Deep Learning as Neural Low-Degree Filtering

A Spectral Theory of Hierarchical Feature Learning

Yatin Dandi with M. Vilucchio, L. Arnaboldi, H. Tabanelli, F. Krzakala

IdePHICS & SPOC Laboratories, EPFL

ProbAI / Scaling Laws Workshop · University of Warwick

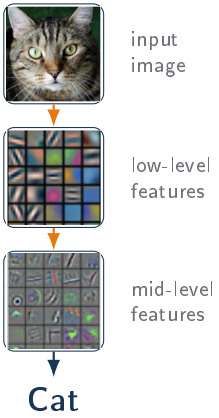
github.com/IdePHICS/Neural-LoFi-Theory

Deep networks build representations, but how?

Empirically, depth *progressively constructs* structured features:

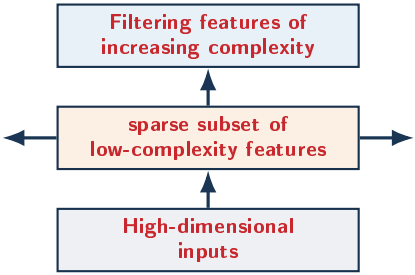
- **Vision:** edges $\rightarrow$ textures $\rightarrow$ parts $\rightarrow$ objects [Zeiler–Fergus '14]
- **Transfer:** early layers general, late layers task-specific [Yosinski et al. '14]
- **Platonic representations:** different models/architectures converge to *aligned* representations [Huh et al. '24]

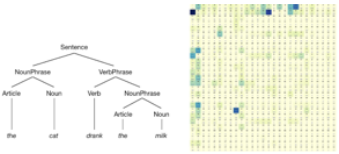
Yet we lack a *simple predictive mechanism* for *how* this happens.



Intuition: feature learning as iterative *filtering*

vision hierarchy





language and sparse structure

*Can we formulate this mathematically for **generic** data distributions?*

Introducing Neural LoFi

Neural Low-Degree Filtering (Neural LoFi) is an interpretable spectral surrogate for GD training.

It is easier to analyze than GD, while still capturing hierarchical feature learning and working well on real data.

hard-to-analyze GD dynamics

$\Rightarrow$

interpretable spectral surrogate

Why introduce a surrogate?

Direct gradient descent is the object we care about, but its feature-learning dynamics are **coupled, nonlinear, and layer-dependent**.

	Feature learning?	Deep GD tractable?
Lazy / NTK [Jacot '18]	✗ no	✓
Mean-field [Mei–Montanari '18]	✓ yes	✗ shallow / hard
Direct GD for deep feature learning	✓ yes	✗ generally hard
Neural LoFi surrogate	✓ **multi-level**	✓ **explicit layer-wise**

*Can we keep the useful GD feature-discovery signal, but make it **simpler, interpretable, and efficient**?*

Roadmap

1. **Surrogate**: derive Neural LoFi from early layer-wise GD.
2. **Predictions**: which features, when they emerge, and why depth helps.

***Key idea**: turn hierarchical learning into an iterative spectral method.*

Motivating the surrogate

Motivation 1: data-agnostic, non-asymptotic

A standard fully-connected network *(no skip connections)*:

$$z_0 = x, \qquad z_\ell = \sigma(W_\ell z_{\ell-1}),$$
$$\widehat{f}(x) = \langle a_L, z_L \rangle .$$

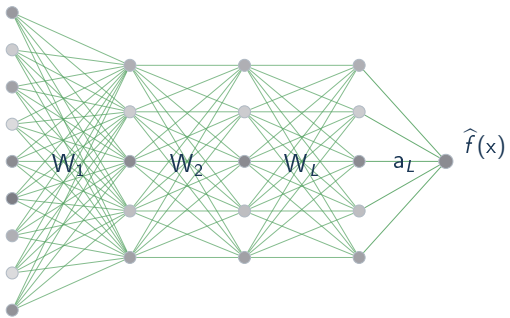
Layer-wise vanishing initialization

$$\|a_L\| \ll \|W_{L-1}\| \ll \cdots \ll \|W_1\| \ll 1.$$

Train layers in sequence (squared loss), earlier layers frozen.

*What does one layer do at **early time**?*

$$\mathcal{L} = \frac{1}{2n} \sum_{\mu=1}^{n} \left(y_\mu - \widehat{f}(x_\mu)\right)^2$$



What one neuron sees at layer ℓ

Neuron $w_{\ell,i}$, residual $r = y - \widehat{f}$. The gradient back-propagated from later layers is

$$-\nabla_{w_{\ell,i}} \mathcal{L} \;\approx\; \widehat{\mathbb{E}}_n\big[\, r\, \bar{a}_{\ell,i}(x,t)\, \sigma'\big(\langle w_{\ell,i}, z_{\ell-1}\rangle\big)\, z_{\ell-1}\big].$$

Layer-wise vanishing initialization

The effective readout is approximately fixed:

$$\bar{a}_{\ell,i}(x,t) = \bar{a}_{\ell,i} + o(1).$$

Downstream layers therefore act as a **fixed effective readout** while W_ℓ moves.

later layers $\approx$ linear

From one-neuron GD to LoFi operators

With the effective readout frozen, the gradient reduces to

$$-\nabla_{w_{\ell,i}} \mathcal{L} \;\approx\; \bar{a}_{\ell,i}\, \widehat{\mathbb{E}}_n\big[\, r\, \sigma'(\langle w_{\ell,i}, z_{\ell-1}\rangle)\, z_{\ell-1}\big].$$

Expand $\sigma'(u) = c_0 + c_1 u + O(u^2)$ near small pre-activations. Since $\widehat{f} \approx 0$ under vanishing init, $r = y - \widehat{f} \approx y$:

$$\partial_t w_{\ell,i} \;\propto\; c_{0,i}\, \underbrace{\widehat{\mathbb{E}}_n[\, y\, z_{\ell-1}]}_{\widehat{u}^{\ell}} \;+\; c_{1,i}\, \underbrace{\widehat{\mathbb{E}}_n[\, y\, z_{\ell-1} z_{\ell-1}^{\top}]}_{\widehat{C}^{(\ell)}}\, w_{\ell,i} \;+\; O(\|w_{\ell,i}\|^2).$$

- drift along the label-correlated direction $\widehat{u}^{\ell}$ (degree-1)
- linear map by the **label-weighted moment operator** $\widehat{C}^{(\ell)}$, the first specialization term

Layer-wise GD = power iteration on a weighted covariance

Proposition (Layer-wise GD, informal)

Under layer-wise vanishing init, for small enough time each neuron evolves as

$$w_{\ell,i}(t) \;\approx\; \exp\big(c_{1,i}\,\widehat{C}^{(\ell)}\,t\big)\,w_{\ell,i}(0), \qquad \widehat{C}^{(\ell)} = \widehat{\mathbb{E}}_n[\,y\,z_{\ell-1}z_{\ell-1}^\top].$$

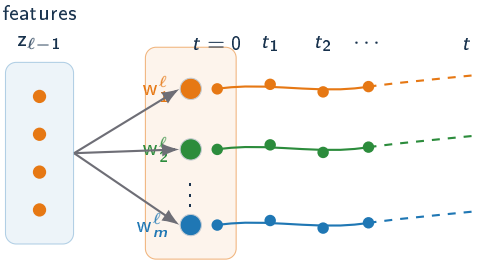
Stacking neurons, with $\widehat{C}^{(\ell)} = \widehat{V}\Lambda\widehat{V}^\top$:

$$W_\ell(t) \;\approx\; \underbrace{W_\ell(0)\,\widehat{V}}_{\text{random mixing}} \;\times\; \underbrace{\exp(c_1\Lambda t)\,\widehat{V}^\top}_{\text{spectral filter} \;\to\; \text{top-}|\lambda|\text{ projection}}.$$

- amplifies directions with **large** $|\lambda_r^{(\ell)}|$ ($c_{1,i}$ random sign $\Rightarrow$ *magnitude*)
- left factor is a fixed random transform of those eigenvectors

Small-init training = ***spectral filtering*** *of the current representation.*

Neurons run parallel power iterations



$$\partial_t w_i \propto \widehat{C}^{(\ell)} w_i,$$
$$\widehat{C}^{(\ell)} = \widehat{\mathbb{E}}_n[y\, z_{\ell-1} z_{\ell-1}^\top].$$

Shared-operator picture

The neurons share the same label-weighted operator, but start from independent random initializations.

Each neuron is an approximate power iteration on $\widehat{C}^{(\ell)}$.

Motivation: isolate mixing from filtering

Stacking the neuron dynamics gives

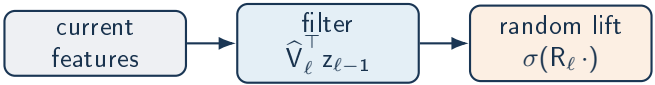
$$W_\ell(t) \approx \underbrace{W_\ell(0)\widehat{V}}_{\text{random mixing from init}} \times \underbrace{\exp(c_1 \Lambda t)\,\widehat{V}^\top}_{\text{spectral filter selects top-}|\lambda|} .$$

Random mixing

Comes from initial weights. It randomly recombines the directions selected by the filter.

Spectral filter

Comes from the label-weighted covariance. This is the task-adaptive part of early GD.



Neural LoFi = filter explicitly, then random lift.

Algorithm: Neural LoFi

Neural LoFi, layer ℓ (backprop-free, one pass)

1. **Statistics:** $\widehat{u}^{\ell} = \frac{1}{n}\sum_{\mu} y_\mu z_{\ell-1}(x_\mu), \quad \widehat{C}^{(\ell)} = \frac{1}{n}\sum_{\mu} y_\mu\, z_{\ell-1}(x_\mu) z_{\ell-1}(x_\mu)^\top.$
2. **Filter:** set $\widehat{v}_0^{\ell} = \widehat{u}^{\ell}/\|\widehat{u}^{\ell}\|$ and let $\widehat{v}_1^{(\ell)}, \dots, \widehat{v}_{k_\ell}^{(\ell)}$ be the top eigenvectors of $\widehat{C}^{(\ell)}$ by $|\lambda|$. Then $\widehat{V}_\ell = [\widehat{v}_0^{\ell}, \widehat{v}_1^{(\ell)}, \dots, \widehat{v}_{k_\ell}^{(\ell)}]$ and $g_\ell = \widehat{V}_\ell^\top z_{\ell-1} \in \mathbb{R}^{k_\ell+1}$.
3. **Lift:** $z_\ell = \frac{1}{\sqrt{p_\ell}}\,\sigma(R_\ell g_\ell)$, $R_\ell \in \mathbb{R}^{p_\ell \times (k_\ell+1)}$ random.

- $\widehat{u}^{\ell}$ is the degree-1 label-feature correlation; $\widehat{C}^{(\ell)}$ gives degree-2 specialization directions.
- For vector labels $y \in \mathbb{R}^c$, the linear filter is SVD of $\widehat{\mathbb{E}}_n[y\, z_{\ell-1}^\top]$; its right singular vectors already give up to c label-aligned directions.

Neural LoFi: full scalar-label algorithm

Algorithm 1 Neural Low-Degree Filtering

Require: Dataset $\{(x_\mu, y_\mu)\}_{\mu=1}^{n}$, depth L, ranks $\{k_\ell\}_{\ell=1}^{L}$, widths $\{p_\ell\}_{\ell=1}^{L}$.

1: Initialize $z_0(x) \leftarrow x$ and $p_0 = d$.

2: **for** $\ell = 1, \dots, L$ **do**

3: $\widehat{u}^{\ell} \leftarrow \frac{1}{n}\sum_{\mu=1}^{n} y_\mu z_{\ell-1}(x_\mu), \quad \widehat{v}_0^{\ell} \leftarrow \widehat{u}^{\ell}/\|\widehat{u}^{\ell}\|.$

▷ estimate linear component

4: Form the label-weighted moment operator

$$\widehat{C}^{(\ell)} \leftarrow \frac{1}{n}\sum_{\mu=1}^{n} y_\mu z_{\ell-1}(x_\mu) z_{\ell-1}(x_\mu)^{\top} \in \mathbb{R}^{p_{\ell-1} \times p_{\ell-1}}.$$

5: $\widehat{V}_\ell = [\widehat{v}_0^{\ell}, \widehat{v}_1^{(\ell)}, \dots, \widehat{v}_{k_\ell}^{(\ell)}]$, with ordered eigenvectors of $\widehat{C}^{(\ell)}$ by decreasing $|\widehat{\lambda}|$.

6: $g_\ell(x) \leftarrow \widehat{V}_\ell^{\top} z_{\ell-1}(x) \in \mathbb{R}^{k_\ell+1}$.

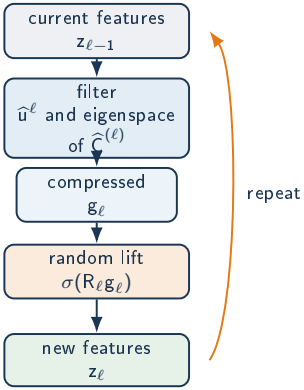
▷ project onto learned features

7: $z_\ell(x) \leftarrow p_\ell^{-1/2}\sigma(R_\ell g_\ell(x)), \quad R_\ell \in \mathbb{R}^{p_\ell \times (k_\ell+1)}$.

▷ nonlinear random lift

8: Fit final linear/logistic readout a on $z_L(x)$.

9: **return** $\widehat{f}(x) = \langle a, z_L(x)\rangle$.



Relation to multi-index spectral estimators

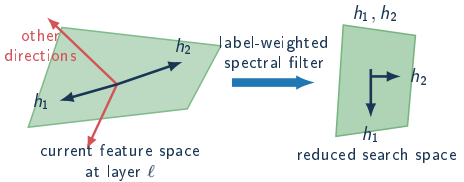
Why second order? Linear $\widehat{\mathbb{E}}_n[yz]$ gives label-aligned directions. Nonlinear features first appear through the matrix statistic $\widehat{\mathbb{E}}_n[yzz^\top]$.

Classical multi-index estimators.

$$\frac{1}{n}\sum_{\mu=1}^{n} g(y_\mu)\, x_\mu x_\mu^\top$$

including sliced inverse regression and related spectral methods.

Neural LoFi adds iteration. Filter in the current representation, lift, then filter again.



Same spectral estimator, repeated after each learned lift.

Multi-index estimators: Lu and Li '17; Mondelli and Montanari '18; Maillard et al. '20; Troiani et al. '24

Incorporating inductive bias

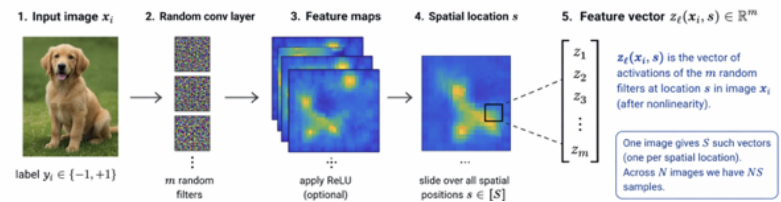
For convolutional layers, replace one feature vector by local feature vectors $z_{\ell-1}(x_\mu, s)$ at spatial locations $s \in [S]$:

$$\widehat{C}_{\mathrm{conv}}^{(\ell)} = \frac{1}{nS} \sum_{\mu=1}^{n} \sum_{s=1}^{S} y_\mu \, z_{\ell-1}(x_\mu, s) z_{\ell-1}(x_\mu, s)^\top.$$

- $z_{\ell-1}(x, s)$ comes from random convolution + nonlinearity.
- One image gives S local samples; n images give nS samples.

Preserved bias

The estimator preserves translation invariance and locality.



Motivation 2: higher order decay (more general, data-dependent)

The Taylor expansion on the gradient slide can also be viewed in an *orthonormal polynomial basis*. Testing the exact population gradient direction against a reference v,

$$\langle \mathsf{G}_\ell(\mathrm{w}_0), \mathrm{v} \rangle = \big\langle\, a_{\ell,\mathrm{v}},\ \sigma'(\langle \mathrm{w}_0, Z \rangle) \,\big\rangle_{L^2(P_\ell)}, \qquad a_{\ell,\mathrm{v}}(Z) = \mathbb{E}[y\, \langle Z, \mathrm{v} \rangle \mid Z].$$

Removing lower degrees gives

$$\langle \mathsf{G}_\ell(\mathrm{w}_0), \mathrm{v} \rangle = \underbrace{c_0\, \langle \mathrm{u}_\ell, \mathrm{v} \rangle}_{\text{constant / degree 1}} + \underbrace{c_1\, \langle \mathsf{C}_\ell \mathrm{w}_0, \mathrm{v} \rangle}_{\text{linear in } \mathrm{w}_0 \text{ / degree 2}} + \sum_{r \geq 2} c_r\, R_{\ell,r}(\mathrm{w}_0, \mathrm{v}).$$

The constant and linear terms are exactly the LoFi operators $\mathrm{u}_\ell, \mathsf{C}_\ell$; higher-order terms are the only missing piece.

Higher order decay ansatz

Ansatz

Let $Z = \mathsf{z}_{\ell-1}(X)$ have m_ℓ **effective degrees of freedom**. For a fixed normalized degree-k polynomial p_k (orthogonal to lower degrees) and random $\mathsf{w}_0 \sim \mathrm{Unif}(\mathbb{S}^{p_{\ell-1}-1})$,

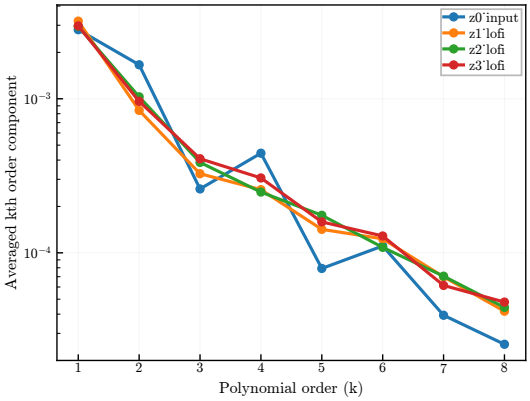
$$\left| \mathbb{E}_Z[\, p_k(Z)\, q_k(\langle \mathsf{w}_0, Z\rangle)\,] \right|^2_{\mathrm{typ}} \asymp_k \; m_\ell^{-k}.$$

Dimension counting: degree-k polynomials (lower degrees removed) span $\binom{m_\ell+k-1}{k} \asymp_k m_\ell^k$ directions; a random unit vector has squared overlap $\sim m_\ell^{-k}$ with any fixed one.

$$\Rightarrow \; |R_{\ell,r}(\mathsf{w}_0, \mathsf{v})|_{\mathrm{typ}} \lesssim_r m_\ell^{-r}, \;\; \langle \mathsf{G}_\ell(\mathsf{w}_0), \mathsf{v}\rangle = c_0 \langle \mathsf{u}_\ell, \mathsf{v}\rangle + c_1 \langle \mathsf{C}_\ell \mathsf{w}_0, \mathsf{v}\rangle + O(m_\ell^{-2}).$$

Higher-order interactions exist on single samples but are ***suppressed in the averaged gradient****.*

Higher order decay on real data (CIFAR-10)

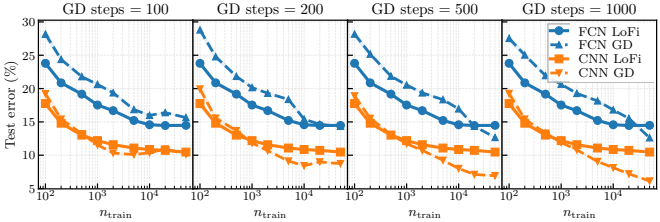


Median squared correlation of y with a normalized residual degree-k *random ridge* component (finite-sample floor subtracted), per layer.

- clear **decay with** k for input *and* every LoFi layer
- consistent with a degree-k space of size $\sim m_\ell^k$

Empirical justification for stopping at second order, without layer-wise scale separation.

Neural LoFi tracks early gradient descent



Binary CIFAR-10 (animals vs. vehicles), FCN & CNN.

- LoFi **matches or beats** GD in the low-data regime and at early training times
- a *one-pass spectral* surrogate, no backprop

A tractable stand-in for the feature-discovery phase of GD.

OK, we have a **simpler surrogate** to GD in the Neural LoFi regime!

Now what can we **learn** from it?

The three questions this talk answers

*Which features are discovered during training, **in which order**, and at **what sample complexity**? And **why** is depth helpful?*

WHICH?

relevance–complexity trade-off in feature space

WHEN?

emergence criterion set by the *effective dimension*

WHY DEPTH?

low-degree compositionality

Recall: the second-order LoFi criterion

At layer ℓ, the quadratic filter diagonalizes

$$\widehat{C}_\ell \equiv \widehat{C}^{(\ell)} = \widehat{\mathbb{E}}_n\left[y\, z_{\ell-1}(x) z_{\ell-1}(x)^\top\right].$$

For $\varphi_v(x) = \langle v, z_{\ell-1}(x)\rangle$,

$$R_\ell(v) := v^\top \widehat{C}_\ell v = \widehat{\mathbb{E}}_n\left[y\, \varphi_v(x)^2\right].$$

Variational view of the eigenvectors. Let $S_{j-1}^{(\ell)} = \mathrm{span}\{\widehat{v}_1^{(\ell)}, \dots, \widehat{v}_{j-1}^{(\ell)}\}$, with $S_0^{(\ell)} = \{0\}$. Ordered by $|\lambda|$,

$$\widehat{v}_j^{(\ell)} \in \underset{\substack{\|v\|_2=1 \\ v \perp S_{j-1}^{(\ell)}}}{\arg\max} |R_\ell(v)| = \underset{\substack{\|v\|_2=1 \\ v \perp S_{j-1}^{(\ell)}}}{\arg\max} \left|\widehat{\mathbb{E}}_n\left[y\, \varphi_v(x)^2\right]\right|.$$

To interpret what is learned, translate this coordinate criterion into a criterion over features in the current feature space.

What is learned: a relevance–complexity principle

Representer theorem: induced features live in an RKHS

The eigenvectors $\widehat{v}_j^{(\ell)}$ live in random-feature coordinates. We interpret them through the induced functions

$$\varphi_v(x) = \langle v, z_{\ell-1}(x) \rangle , \qquad K_{\ell-1}(x, x') = \langle z_{\ell-1}(x), z_{\ell-1}(x') \rangle .$$

Representer theorem on the training sample

With $\mathcal{H}_{\ell-1}$ the RKHS of $K_{\ell-1}$ and $K_{\mu\nu} = K_{\ell-1}(x_\mu, x_\nu)$,

$$\varphi(\cdot) = \sum_{\mu=1}^{n} \alpha_\mu K_{\ell-1}(x_\mu, \cdot), \qquad \boldsymbol{\varphi} = K\alpha, \qquad \|\varphi\|^2_{\mathcal{H}_{\ell-1}} = \alpha^\top K \alpha = \boldsymbol{\varphi}^\top K^\dagger \boldsymbol{\varphi}.$$

A learned vector is best viewed through the feature it induces on the data.

From Euclidean norm to a function-space problem

The top LoFi direction is selected in weight space by

$$\widehat{\mathrm{v}}_1^{(\ell)} \in \operatorname*{arg\,max}_{\|\mathrm{v}\|_2=1} \left| \frac{1}{n} \sum_{\mu=1}^{n} y_\mu \, \langle \mathrm{v}, \mathrm{z}_{\ell-1}(\mathrm{x}_\mu) \rangle^2 \right|.$$

Geometry transfer

For the minimum-norm representative of $\varphi_{\mathrm{v}}(\mathrm{x}) = \langle \mathrm{v}, \mathrm{z}_{\ell-1}(\mathrm{x}) \rangle$,

$$\|\varphi_{\mathrm{v}}\|_{\mathcal{H}_{\ell-1}} = \|\mathrm{v}\|_2.$$

Equivalent feature problem

$$\varphi_1^{(\ell)} \in \operatorname*{arg\,max}_{\|\varphi\|_{\mathcal{H}_{\ell-1}}=1} \left| \widehat{\mathbb{E}}_n[y \, \varphi(\mathrm{x})^2] \right|.$$

The Euclidean unit sphere for v *becomes an RKHS unit sphere for the feature.*

RKHS norm is current-layer complexity

Let $\psi_1, \psi_2, \dots$ be eigenfunctions of the current kernel $K_{\ell-1}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$. If $\varphi = \sum_j a_j \psi_j$, then

$$\|\varphi\|^2_{\mathcal{H}_{\ell-1}} = \sum_j \frac{a_j^2}{\lambda_j}.$$

Cheap directions

Large λ_j directions are easy to express in the current representation.

Expensive directions

Small λ_j directions require large RKHS norm and are complex in this kernel.

Neural LoFi seeks low-degree label correlation under a low-complexity budget in the current kernel.

Theorem (Variational characterization & infinite-width limit)

The linear feature maximizes $|\widehat{\mathbb{E}}_n[y\,\psi]|$; the 2nd-order features successively maximize

$$\varphi_k = \operatorname*{arg\,max}_{\substack{\|\varphi\|_{\mathcal{H}_{\ell-1}}=1 \\ \varphi \perp \varphi_1,\dots,\varphi_{k-1}}} \left| \widehat{\mathbb{E}}_n[\, y\, \varphi(\mathrm{x})^2\,] \right|,$$

and as $p_{\ell-1} \to \infty$ they converge (L^2) to eigenfunctions of the limiting kernel.

relevance: large $\widehat{\mathbb{E}}_n[y\,\varphi^2]$
(low-degree correlation with the label)

complexity: small $\|\varphi\|_{\mathcal{H}_{\ell-1}}$
(simple in the current geometry)

*Neural LoFi seeks features that are **simple** in the previous layer's geometry but **predictive** through low-degree correlation with the task.*

Depth = a sequence of task-adaptive kernels

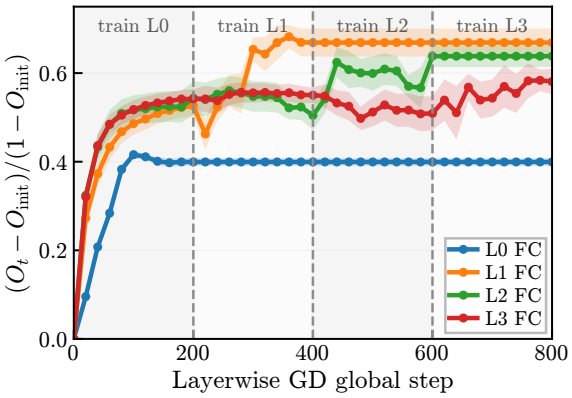
Once $\varphi_1^{(\ell)}, \dots, \varphi_{k_\ell}^{(\ell)}$ are selected, the random lift induces the next kernel

$$K_\ell(x, x') = \mathbb{E}_r\big[\sigma(r^\top g_\ell(x))\,\sigma(r^\top g_\ell(x'))\big], \qquad g_\ell = (\varphi_1^{(\ell)}, \dots, \varphi_{k_\ell}^{(\ell)}).$$

Each transition is **supervised**: the next kernel is built from the low-complexity features whose squared activations correlate most with y.

Not a single fixed kernel chosen before seeing labels: an ***adaptive multilayer kernel construction****.*

Empirical link: LoFi tracks *layer-wise training*



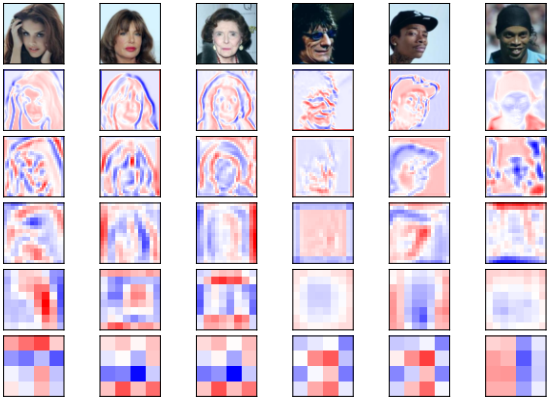
Layer-wise GD on CIFAR-10 (4 hidden layers). During each shaded window *only one* layer is trained. Plotted: normalized overlap gain between the GD layer's top singular features and the **fixed LoFi filtered features**.

- a clean **staircase**: the active layer becomes aligned with the LoFi features for that layer
- already-trained layers stay put

Layer-wise GD moves toward the directions selected by LoFi.

Generality: CelebA (gender attribute)

first-layer filters



activations of a mid-ranked feature, deepening by row

6 conv + 1 FC Neural LoFi on CelebA.

- structured filters & selective activations
- pure spectral filtering

When do features emerge? The effective-dimension criterion

Learning is not smooth: features emerge one at a time

Training shows long plateaus then **abrupt jumps**; new directions emerge sequentially [Wei et al. '22; Arora–Goyal '23]. Neural LoFi gives a mechanism. The k-th feature at layer ℓ maximizes

$$\widehat{\rho}_\ell^{(k)} = \sup_{\substack{\|\varphi\|_{\mathcal{H}_{\ell-1}}=1 \\ \varphi \perp \varphi_{<k}}} \left| \widehat{\mathbb{E}}_n[\, y\, \varphi(\mathrm{x})^2 \,] \right|.$$

Empirical correlation $\widehat{c}_{\ell,n}(\varphi)$ fluctuates around its population value $c_\ell(\varphi)$. Over the candidate class $\mathcal{S}_k^\ell$ define

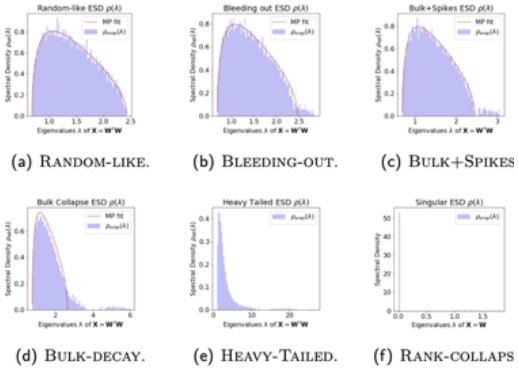
$$\rho_\ell^{(k)} = \sup_{\mathcal{S}_k^\ell} |c_\ell(\varphi)| \quad \text{(signal)}, \qquad \tau_\ell^k(n) = \sup_{\mathcal{S}_k^\ell} \left| \widehat{c}_{\ell,n}(\varphi) - c_\ell(\varphi) \right| \quad \text{(noise floor)}.$$

Emergence criterion = a BBP-type transition

A feature becomes learnable when $\rho_\ell^{(k)} \gg \tau_\ell^k(n)$.

Below threshold: the leading empirical direction is *noise*.
Above threshold: a task-relevant direction **separates from the bulk**.

- exactly the BBP / Baik–Ben Arous–Péché spike transition [Baik et al. '05]
- same phenomenology as spiked matrix models / spectral inference



(a) RANDOM-LIKE. (b) BLEEDING-OUT. (c) BULK+SPIKES

(d) BULK-DECAY. (e) HEAVY-TAILED. (f) RANK-COLLAPS

empirical spectra: bulk, outliers, and collapse regimes

Residual effective dimension: definition

Fix a layer ℓ and a candidate degree block $\mathcal{S}_k^\ell$. Let $T_{\ell,k}$ be the population kernel/covariance operator on this block, and let $\Pi_{<k}$ project onto directions already extracted by earlier filters.

$$T_{\ell,k}^{\perp} = (I - \Pi_{<k})\, T_{\ell,k}\, (I - \Pi_{<k}), \qquad T_{\ell,k}^{\perp} e_j = \lambda_j^{\perp} e_j.$$

At resolution $r > 0$, define

$$\boxed{D_{\ell,k}^{\mathrm{eff},\perp}(r) = \mathrm{Tr}\Big[T_{\ell,k}^{\perp}\big(T_{\ell,k}^{\perp} + r^2 I\big)^{-1}\Big] = \sum_{j \geq 1} \frac{\lambda_j^{\perp}}{\lambda_j^{\perp} + r^2}}.$$

- soft count of *residual* directions with eigenvalue above r^2
- after each feature is removed, the count and the noise floor change
- estimable from the empirical kernel spectrum after projection

Emergence is controlled by the number of directions still competing with the next feature.

Proof sketch: residual effective dimension

Uniform fluctuation. For residual candidates $\varphi \in \mathcal{S}_k^\ell$, define

$$f_\varphi(X, Y) = Y\,\varphi(X)^2 - \mathbb{E}[Y\,\varphi(X)^2].$$

Localized empirical noise is

$$\tau_{\ell,r}^k(n) = \sup_{\varphi \in \mathcal{S}_k^\ell(r)} \left|(P_n - P) f_\varphi\right|.$$

Symmetrization gives

$$\mathbb{E}\,\tau_{\ell,r}^k(n) \lesssim \mathfrak{R}_n(\mathcal{S}_k^\ell(r)).$$

Residual spectrum. Expand in the eigenbasis of $T_{\ell,k}^\perp$, after projecting out previously found features. Only directions with $\lambda_j^\perp \gtrsim r^2$ matter:

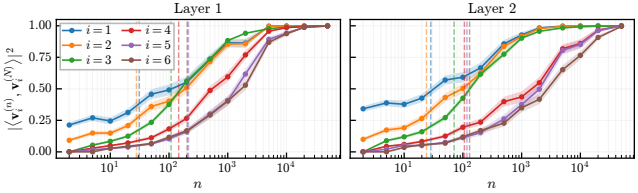
$$D_{\ell,k}^{\mathrm{eff},\perp}(r) := \sum_j \frac{\lambda_j^\perp}{\lambda_j^\perp + r^2} \approx \#\{j : \lambda_j^\perp \gtrsim r^2\}.$$

Thus, up to logarithms,

$$\mathfrak{R}_n(\mathcal{S}_k^\ell(r)) \lesssim r\sqrt{\frac{D_{\ell,k}^{\mathrm{eff},\perp}(r)}{n}}.$$

Residual effective dimension is the local complexity of the remaining search space; localized matrix concentration gives the same scale.

Predicting *when* each concept emerges on real data

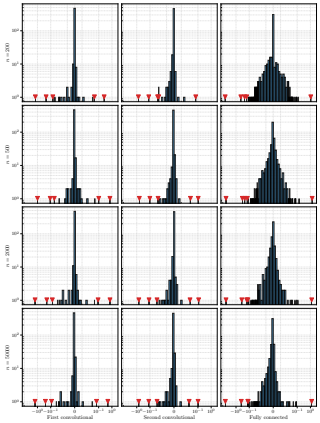


Squared overlap $|\left\langle \mathrm{v}_i^{(n)}, \mathrm{v}_i^{(N)} \right\rangle|^2$ of each eigenvector with a large-sample reference, vs. n (layers 1 & 2).

- **dashed lines** = thresholds from the effective-dimension criterion $\rho \gg \tau$
- each overlap rises *right at* its predicted n_ℓ^i

A quantitative, data-driven theory of layer-wise concept emergence.

Spectrum of the label-weighted operator



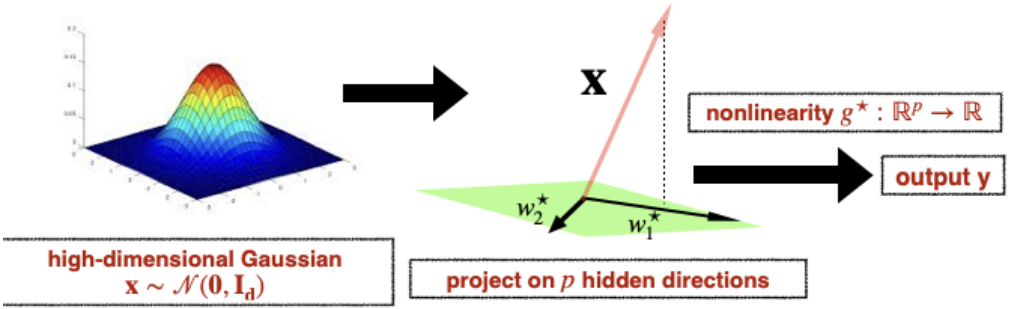
Setting. Binary CIFAR-10 (animals vs. vehicles), convolutional Neural LoFi. Columns are first conv, second conv, and fully connected; rows increase n_{train}; red markers are the five most dominant eigenvalues.

- small n: leading directions sit *inside* the bulk (noise)
- growing n: task-relevant eigenvalues **separate from the bulk**

The BBP/EJ spike picture, layer by layer: the spectral signature of emergence.

Why depth helps: low-degree compositionality

Recap: Gaussian multi-index models



Classical multi-index teacher

$$\mathrm{x} \sim \mathcal{N}(0, I_d), \qquad y = g^\star(U^{\star\top}\mathrm{x}) + \xi, \qquad V^\star = \mathrm{span}\{\mathrm{u}_1^\star, \dots, \mathrm{u}_r^\star\}.$$

Task: recover the hidden subspace $V^\star$ from labels. [Ben Arous et al. '21; Bietti et al. '23; Abbe et al. '23]

From multi-index models to hierarchy

Classical multi-index

$$x \longmapsto U^{\star\top} x \longmapsto y.$$

Recover one hidden subspace $V^\star$.

Hierarchical target

$$x \longmapsto h^{(1)}(x) \longmapsto h^{(2)}(x) \longmapsto y.$$

Recover a sequence of learned representations.

Can we generalize multi-index recovery to ***multiple levels of hierarchy****?*

A solvable model: planted hierarchical target

Draw $x \sim \mathcal{N}(0, I_d)$ and plant $x \in \mathbb{R}^d \to h^{(1)}(x) \in \mathbb{R}^{d_1} \to h^{(2)}(x) \in \mathbb{R} \to y \in \mathbb{R}$.

Data space: $x \sim \mathcal{N}(0, I_d)$.

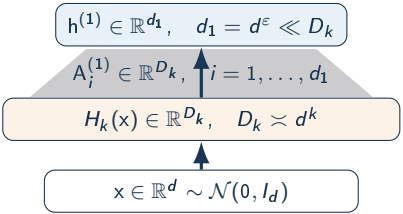
$x \in \mathbb{R}^d \sim \mathcal{N}(0, I_d)$

A solvable model: planted hierarchical target

Draw $\mathrm{x} \sim \mathcal{N}(0, I_d)$ and plant $\mathrm{x} \in \mathbb{R}^d \to \mathrm{h}^{(1)}(\mathrm{x}) \in \mathbb{R}^{d_1} \to h^{(2)}(\mathrm{x}) \in \mathbb{R} \to y \in \mathbb{R}$.

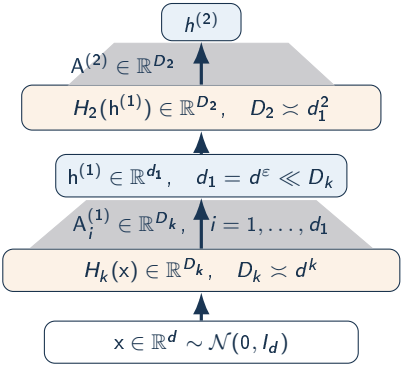
Data space: $\mathrm{x} \sim \mathcal{N}(0, I_d)$.

Sparse degree-k selection: $h_i^{(1)}(\mathrm{x}) = \left\langle \mathrm{A}_i^{(1)}, H_k(\mathrm{x}) \right\rangle$, $i = 1, \dots, d_1 = d^{\varepsilon} \ll d^k$. Entries of $\mathrm{A}_i^{(1)}$ have variance $\asymp d^{-k}$.



A solvable model: planted hierarchical target

Draw $\mathrm{x} \sim \mathcal{N}(0, I_d)$ and plant $\mathrm{x} \in \mathbb{R}^d \to \mathrm{h}^{(1)}(\mathrm{x}) \in \mathbb{R}^{d_1} \to h^{(2)}(\mathrm{x}) \in \mathbb{R} \to y \in \mathbb{R}$.



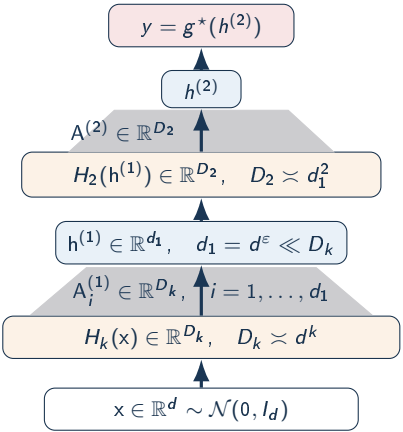
Data space: $\mathrm{x} \sim \mathcal{N}(0, I_d)$.

Sparse degree-k selection: $h_i^{(1)}(\mathrm{x}) = \left\langle \mathrm{A}_i^{(1)}, H_k(\mathrm{x}) \right\rangle$, $i = 1, \dots, d_1 = d^{\varepsilon} \ll d^k$. Entries of $\mathrm{A}_i^{(1)}$ have variance $\asymp d^{-k}$.

Second low-degree step: $h^{(2)}(\mathrm{x}) = \left\langle \mathrm{A}^{(2)}, H_2(\mathrm{h}^{(1)}(\mathrm{x})) \right\rangle$, with entry variance $\asymp d_1^{-2}$.

A solvable model: planted hierarchical target

Draw $x \sim \mathcal{N}(0, I_d)$ and plant $x \in \mathbb{R}^d \to h^{(1)}(x) \in \mathbb{R}^{d_1} \to h^{(2)}(x) \in \mathbb{R} \to y \in \mathbb{R}$.



Data space: $x \sim \mathcal{N}(0, I_d)$.

Sparse degree-k selection: $h_i^{(1)}(x) = \left\langle A_i^{(1)}, H_k(x) \right\rangle$, $i = 1, \dots, d_1 = d^\varepsilon \ll d^k$. Entries of $A_i^{(1)}$ have variance $\asymp d^{-k}$.

Second low-degree step: $h^{(2)}(x) = \left\langle A^{(2)}, H_2(h^{(1)}(x)) \right\rangle$, with entry variance $\asymp d_1^{-2}$.

Label: $y = g^\star(h^{(2)}(x))$. Globally high-degree in x, but low-degree after the right representation is found.

Toy proof I: shallow learning sees the full composition

$$x \xrightarrow{H_k,\ A^{(1)}} h^{(1)} \xrightarrow{H_2,\ A^{(2)}} h^{(2)} \xrightarrow{g^\star} y.$$

One-shot kernel / RF view

It sees only the composite map $x \mapsto y$:

$$\deg_x(y) \sim 2kp, \qquad n_{\text{shallow}} \gtrsim d^{2kp}.$$

Layerwise LoFi view

Ask for the first representation before fitting the whole target:

recover $h^{(1)}$ first.

Then the next problem is low-degree in d_1 latent variables.

The proof separates the hard composite target into a sequence of low-degree recovery problems.

Toy proof II: layer 1 recovers the hidden features

Let $U = H_k(\mathrm{x}) \in \mathbb{R}^{D_k}$ with $D_k \asymp d^k$, and $h_i^{(1)} = \left\langle \mathrm{A}_i^{(1)}, U \right\rangle$ for $i \le d_1 = d^{\varepsilon}$.

At small random lift,

$$\sigma(\langle \mathrm{w}, \mathrm{x} \rangle) = \sum_{j \ge 0} q_j(\mathrm{w})\, \mathrm{He}_j(\mathrm{x}).$$

Lower-degree correlations vanish by construction; the first nonzero supervised term appears at degree k.

The degree-k LoFi statistic is

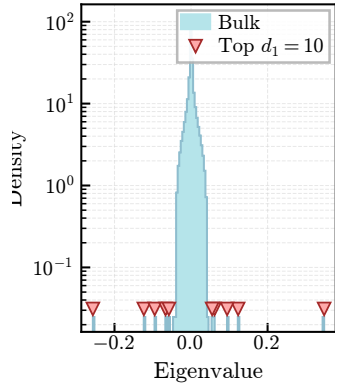
$$\widehat{\mathrm{C}}_{1,k} = \frac{1}{n} \sum_{\mu=1}^{n} y_\mu\, H_k(\mathrm{x}_\mu) H_k(\mathrm{x}_\mu)^\top.$$

Its leading eigenspace targets

$$\mathcal{A}_1 = \mathrm{span}\{\mathrm{A}_i^{(1)} : i \le d_1\}.$$

With $n \gtrsim D_k d_1 = d^{k+\varepsilon}$ samples, LoFi recovers the first hidden representation.

Toy proof III: population spike and empirical noise



Wick / Stein identities give, on the degree-k block,

$$C_{1,k} = \mathbb{E}[y\, UU^\top] \simeq A^{(1)} B A^{(1)\top}, \qquad \mathrm{rank}(B) \le d_1.$$

Empirical noise has bulk scale

$$\tau(n) \asymp \sqrt{\frac{D_k}{n}}, \qquad \rho_{\min} \asymp d_1^{-1/2}.$$

Hence

$$\rho_{\min} \gg \tau(n) \iff n \gg D_k\, d_1 \asymp d^{k+\varepsilon}.$$

The useful directions are BBP-style outliers: once they leave the bulk, the hidden features are recoverable.

Toy proof IV: why this proof is tractable

The point is not to solve full GD dynamics. Neural LoFi turns the argument into **static spectral statistics**:

- scaling of label-feature correlations
- Gaussian / CLT limits for Wiener chaos
- matrix concentration for the empirical operator

This proof template also applies beyond the Gaussian toy setting:

- hierarchical spectral methods [Tabanelli, Dandi, Pesce, Krzakala '26]
- Random Hierarchy Model [Cagnetta et al. '24]
- extensions to structured feature learning [Ren, Dandi, Krzakala, Lee '26]

The surrogate is simple enough to prove hierarchy, but still captures the feature-discovery signal missing from fixed kernels.

Why this model separates depth from fixed kernels

The target is high-degree in the input, but **low-degree one representation at a time**:

$$x \xrightarrow{H_k,\ A^{(1)}} h^{(1)} \xrightarrow{H_2,\ A^{(2)}} h^{(2)} \xrightarrow{g^\star} y.$$

Fixed kernel / RF
sees only $x \mapsto y$; for $k = 2$ needs $n = \Omega(d^4)$ just to beat random guessing.

Two-layer FL
can recover $h^{(1)}$, but still faces a problem over the d_1 latent variables.

Depth / LoFi
first recovers $h^{(1)}$, then solves a quadratic problem in dimension d_1.

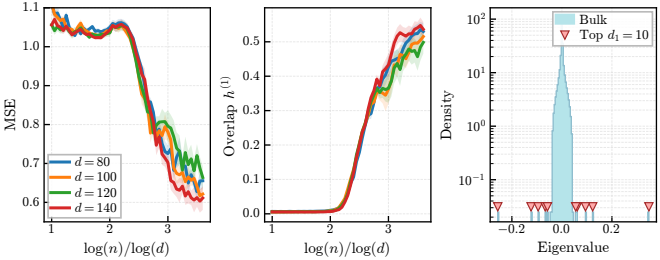
Depth resolves it: once $h^{(1)}$ is recovered, $h^{(2)}$ is a quadratic problem in dimension d_1.

Prediction: emergence at the effective-dimension scale

The first representation $h^{(1)}$ becomes recoverable at

$$n \gg D^{\mathrm{eff}} \cdot d_1 = D_q\, d_1 = O\big(d^{\,q+\varepsilon}\big) \qquad \text{(quadratic: } n \gg d^{2+\varepsilon}\text{)}.$$

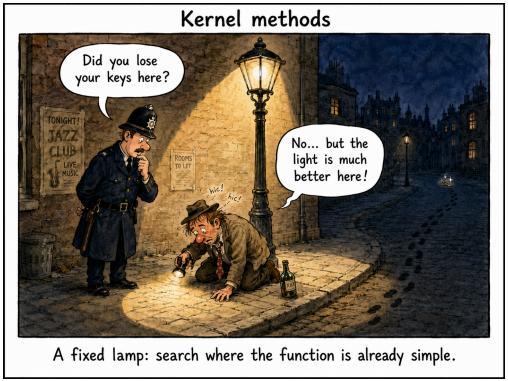
A concrete instance of the general criterion $\rho \gg \tau$, with $D^{\mathrm{eff}} = D_q = O(d^q)$ the size of the degree-q Hermite block.

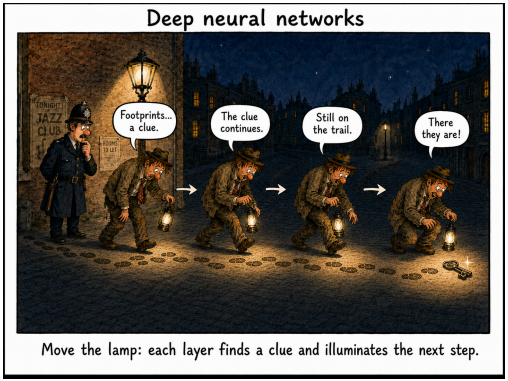


At $n \propto d^{2.5}$ (k=2, ε=$\frac{1}{2}$), *simultaneously*:

- MSE **drops**
- overlap with $h^{(1)}$ **jumps**
- leading eigenvalues **leave the bulk**

Lamp-and-keys view of adaptive search





Filtering is not just selection: it ***adapts the search space*** *for the next low-degree signal.*

Depth changes the geometry in which we search

Each layer *lifts* the selected features, so the next layer searches for signal in a **new representation**. A target can be high-degree in x yet low-degree in z_ℓ.

$$x \longrightarrow z_1 \longrightarrow z_2 \longrightarrow \cdots \longrightarrow z_L,$$

Low-degree compositionality

Depth helps when, *at each stage*, the next useful feature is visible through a **low-degree statistic** of the current representation: i.e. the population correlation clears the noise floor.

*A high-degree target can be learned as a **sequence of low-degree problems**: each filtering step moves the lamp for the next search.*

A data-adaptive compositional information exponent

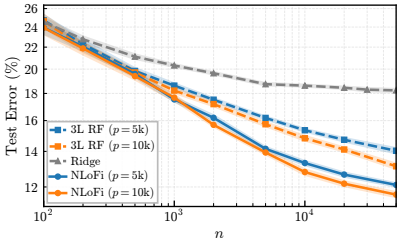
Planted-hierarchy view [Dandi et al. '24]: smallest degree q with

$$\left\| \mathbb{E}[(h_\ell^\star(x))^{\otimes q} f^\star(x)] \right\|_F = \Theta(1).$$

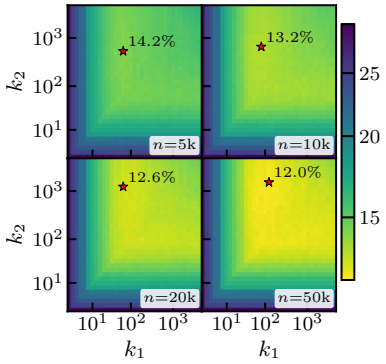
- low compositional exponent $\Leftrightarrow$ strong low-degree dependence on the hidden representation
- Neural LoFi does **not** assume access to $h_\ell^\star$; it *searches the learned feature space* via $\widehat{\mathbb{E}}_n[y\varphi^2]$
- acts as a **data-adaptive compositional information test**

A supervised coarse-graining: keep a few task-relevant directions, discard the rest, in the spirit of renormalization across scales.

FCN on CIFAR-10: filtering beats fixed random features



test error vs. n_{train}



error over retained features (k_1, k_2)

- label-weighted operator **extracts directions unavailable to fixed RF** $\Rightarrow$ gain over ridge / RF
- feature selection is non-trivial: best k is *not* the largest, the optimal k *grows with* n

Takeaways

What we now understand about layer-wise training

The mechanism

- layer-wise GD $\to$ spectral filter of the label-weighted operator $\widehat{\mathrm{C}}^{(\ell)} = \widehat{\mathbb{E}}_n[y\, \mathrm{z}\mathrm{z}^\top]$
- filter + random lift, repeated with depth

What is learned

- relevance–complexity principle in feature/kernel space
- an adaptive multilayer kernel $K_0 \to \cdots \to K_L$

When

- emergence = BBP transition; noise floor set by the **effective dimension**
- predicts per-concept thresholds on real data

Why depth

- low-degree compositionality: a hard problem becomes a *sequence of low-degree ones*

*Layer-wise training is now **quantitatively understood**: which features, in which order, and why depth helps.*

Open directions

- **Beyond second order:** higher-degree statistics $\rightarrow$ tensor spectral problems; generative-/leap-exponent learnability.
- **Multi-pass LoFi:** alternate filtering with target/residual transformations and backward feature correction $\Rightarrow$ closer to trained networks.
- **Long-time dynamics** of the task-adaptive kernels K_ℓ.
- **Structured architectures:** locality, equivariance, and *attention*.
- **From surrogate to primitive:** initialization, feature pretraining, pruning, diagnostics, and **scaling-law prediction**.

`github.com/IdePHICS/Neural-LoFi-Theory`

Thank you!

Neural LoFi: layer-wise training as low-degree spectral filtering

which features · in which order · why depth

Backup: selected references

Regimes & feature learning
Jacot, Gabriel, Hongler '18 (NTK)
Mei, Montanari, Nguyen '18 (mean-field)
Ben Arous, Gheissari, Jagannath '21 (information exp.)
Ghorbani et al. '20; Bietti et al. '22/'23; Abbe et al. '23 (multi-index)
Abbe, Boix-Adsera, Misiakiewicz '23 (leap)

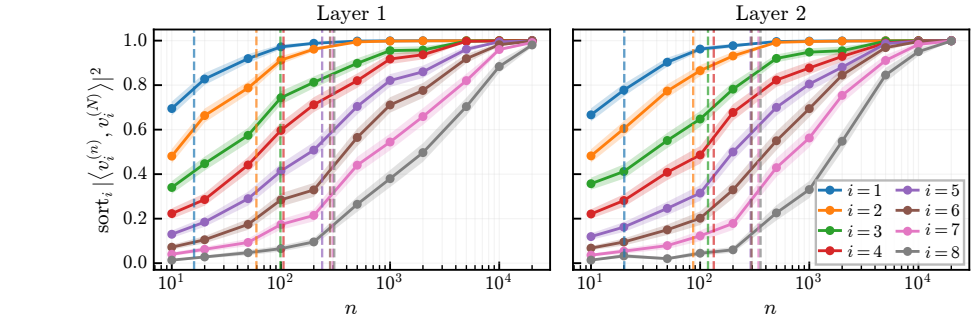
Depth & compositionality
Dandi et al. '24 (compositional information exp.)
Tabanelli, Dandi, Pesce, Krzakala '26 (hierarchical spectral methods)
Cagnetta et al. '24 (Random Hierarchy Model); Ren, Dandi, Krzakala, Lee '26
Wang et al. '23; Nichani et al. '23

Spectral / Hessian learning
Baik, Ben Arous, Péché '05 (BBP)
Bartlett, Bousquet, Mendelson '05 (local Rademacher)
Caponnetto, De Vito '07 (effective dimension)

Mechanistic feature learning
Beaglehole et al. '23 (mechanism of feature learning)

Emergence
Wei et al. '22; Arora, Goyal '23; Schaeffer et al. '23

Backup: feature emergence also holds for CNNs



Per-eigenvector overlap with a large-sample reference vs. n for a *convolutional* Neural LoFi (layers 1 & 2, features $i = 1, \ldots, 8$); dashed lines mark the predicted emergence thresholds.

The effective-dimension criterion also predicts layer-wise emergence for ***convolutional*** *architectures.*