Optimal scaling laws in learning hierarchical multi-index models

Leonardo Defilippis LEONARDO.DEFILIPPIS@ENS.PSL.EU
Bruno Loureiro
Departement d'Informatique, École Normale Supérieure, PSL & CNRS, France

Florent Krzakala
Information, Learning and Physics Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

Antoine Maillard
INRIA Paris & Departement d'Informatique, École Normale Supérieure, PSL & CNRS, France

Abstract

In this work, we provide a sharp theory of scaling laws for two-layer neural networks trained on a class of *hierarchical multi-index* targets, in a genuinely representation-limited regime. We derive exact information-theoretic scaling laws for subspace recovery and prediction error, revealing how the hierarchical features of the target are sequentially learned through a cascade of phase transitions. We further show that these optimal rates are achieved by a simple, target-agnostic spectral estimator, which can be interpreted as the small learning-rate limit of gradient descent on the first-layer weights. Once an adapted representation is identified, the readout can be learned statistically optimally, using an efficient procedure. As a consequence, we provide a unified and rigorous explanation of scaling laws, plateau phenomena, and spectral structure in shallow neural networks trained on such hierarchical targets.

1. Introduction

Despite the staggering practical success of neural networks, we still lack a predictive theory answering a deceptively simple question: given a structured learning problem, *how does the network adapts to the task, in what order are the relevant features in the data learned, and how these translate in statistical efficiency?* This question sits at the intersection of three active lines of research. First, the empirical observation of *neural scaling laws* [17, 35, 37] suggest that the performance of large models scale as a power-law in the training resources, yet—with few exceptions—our mathematical understanding of this relationship remains largely confined to linear models or networks in the lazy regime. Second, recent literature of the training dynamics of neural networks increasingly suggest that feature learning is not a smooth process, but are associated to long *plateaus and abrupt transitions* in the risk, with features (or *concepts* in this context) appearing sequentially rather than all at once [54, 55, 61]. Third, empirical analyses of trained networks have uncovered robust regularities in how the learned representations manifest in the trained network weights, such as in their spectral structure, but without a first-principles explanation of *why* particular features are preferred or *when* they should emerge [46, 58, 60].

This paper provides an end-to-end answer to these questions in a mathematically tractable—yet genuinely feature learning—setting of a two-layer neural network trained on *hierarchical multi-index* data, a class of structured supervised learning tasks where the target function depends on a hierarchical combination of functions of the covariates [10, 27, 52]. Here, hierarchical denote the

© L. Defilippis, B. Loureiro, F. Krzakala & A. Maillard.

fact that the direction are *ordered*, in the sense that their relative importance decay as a power-law with their index number. This leads to a quasi-sparsity of the target representation, with ordered hierarchy of feature strengths as is classical in signal processing [28, 44]. In this task, learning is fundamentally *representation-limited*: the labels $y = g(\boldsymbol{W}_\star \boldsymbol{x})$ depend on a low-dimensional but unknown subspace $\mathrm{span}(\boldsymbol{W}_\star) \subset \mathbb{R}^d$ of the input space $\boldsymbol{x} \in \mathbb{R}^d$, and generalization hinges on discovering this subspace as well as possible from the data. Our goal is to turn this qualitative picture into sharp, quantitative predictions. More precisely, our **main contributions** are:

(i) **Optimal Bayes rates for feature recovery.** We derive the exact information-theoretic limits for recovering the features subspace $\mathrm{span}(\boldsymbol{W}_\star)$ from $n = \Theta(d)$ samples. Further, we precisely characterize the associated optimal mean-squared error rates (a.k.a. *scaling laws*), unveiling the presence of cross-overs and plateaus regions that translate a fundamental trade-off between data-scarce and model-limited regions, depending on the hardness of the underlying hierarchical structure. Interestingly, these rates coincide with the minimax bounds for quasi-sparse recovery achieved by the LASSO [51], and with previously conjectured results for shallow networks with quadratic activation [25]. Our analysis shows that these scalings are in fact universal for a broad class of hierarchical multi-index targets, well beyond the quadratic setting.

(ii) **A matching spectral algorithm and sequential feature emergence.** We introduce a simple, *target-agnostic* spectral estimator that provably achieves the Bayes-optimal rates derived above. The recovery proceeds sequentially: the i-th direction in the hierarchy becomes detectable at a sample complexity $n_i = \Theta(i^{2\gamma} d)$ —where γ is the exponent controlling the hierarchy— leading to a cascade of sharp phase transitions.

(iii) **Learnability of the target function by neural networks.** Once the features/concepts subspace is identified, we show that learning the second-layer (readout) weights incurs no additional statistical bottleneck: the resulting prediction error matches the Bayes-optimal rate for subspace recovery.

Together, these results show that representation learning proceeds through a sequence of sharp phase transitions as the number of samples increases. New directions in the signal subspace emerge one after another, leading to plateaus and abrupt drops in the prediction error. This sequential emergence (to borrow the terminology of [55, 61]) of features provides a precise theoretical underpinning for the empirically observed phenomenon of progressive concept learning in neural networks. Strikingly, the resulting phenomenology closely mirrors that recently uncovered for diagonal and quadratic neural networks using heuristic tools from statistical physics, where progressive feature emergence and distinct scaling regimes were observed [27]. Our contribution is to place this picture on firm theoretical ground and to substantially generalize it to a broader and more realistic class of models and targets.

2. Setting

Consider a supervised regression problem with training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$, which we assumed were drawn i.i.d. from a joint distribution over $\mathbb{R}^{d+1}$. Recall that the goal in regression is to learn a target function $f_\star(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$ from the training data $\mathcal{D}$. In the following, we will be interested in particular structured tasks where the dependence of the target $f_\star$ on the covariates $\boldsymbol{x} \in \mathbb{R}^d$ is given by a hierarchical combination of low-dimensional subtasks. Mathematically, this is formalized by *hierarchical multi-index models*.

Definition 1 (Hierarchical multi-index model) *Let $\boldsymbol{W}_\star = (\boldsymbol{w}_k^\star \in \mathbb{R}^d)_{k\in[m_\star]}$ denote a family of $m_\star$ orthogonal vectors of norm $\|\boldsymbol{w}_k^\star\|^2 = d$. A* hierarchical multi-index *target is defined as*

$$f_\star(\boldsymbol{x}) = \sum_{k=1}^{m_\star} a_k^\star g_k(\langle \boldsymbol{w}_k^\star, \boldsymbol{x}\rangle), \tag{1}$$

where $a_1^\star > a_2^\star > \ldots > a_{m_\star}^\star$ satisfy $\sum_{k=1}^{m_\star}(a_k^\star)^2 = 1$ and $g_k : \mathbb{R} \to \mathbb{R}$. Moreover, we say $f_\star$ is a scale-free *hierarchical multi-index model if $a_k^\star = \Theta(k^{-\gamma})$ for some $\gamma > 0$.*

A few remarks are in order.

- As the name suggests, hierarchical multi-index models can be seen as a sum of $m_\star$ effectively one-dimensional tasks $g_k(z_k)$, with decreasing weight $a_k^\star$.
- More generally, multi-index models are functions of the type $f_\star(\boldsymbol{x}) = g(\boldsymbol{W}_\star \boldsymbol{x})$, where $g : \mathbb{R}^{m_\star} \to \mathbb{R}$ is known as the *link function* and

$$\boldsymbol{z} = \boldsymbol{W}_\star \boldsymbol{x} \in \mathbb{R}^{m_\star} \tag{2}$$

  are known as the *indices*. They have been widely studied as statistical models for regression in the statistics literature [5, 41, 64].
- More recently, multi-index models have gained in popularity in the machine learning theory literature as generative models for data, where they have been used to prove feature learning separation results for neural networks [1, 23, 24].
- Scale-free hierarchical multi-index models were considered recently in [10, 27, 52]. Note that when $\gamma > 1/2$, the target is *quasi-sparse* in the index basis [28, 44]. This makes the model a suitable framework for investigating feature learning in neural networks, which are known to exhibit an implicit bias towards sparse estimators [2, 34, 56].
- In scale-free hierarchical multi-index models with $\gamma < 1/2$, the coefficients must scale as $a_k^\star = \Theta(k^{-\gamma} m_\star^{\gamma-1/2})$ to guarantee the boundedness of the labels.

Definition 2 *The training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ is drawn i.i.d. from a scale-free hierarchical multi-index model $f_\star(\boldsymbol{x})$ as in Def. 1, in particular*

$$y_i = f_\star(\boldsymbol{x}_i) + \sqrt{\Delta}\xi_i \tag{3}$$

where $\boldsymbol{x}_i \sim \mathcal{N}(0, {}^1\!/_d \boldsymbol{I}_d)$, $\Delta > 0$ is the noise variance and $\xi_i \sim \mathcal{N}(0,1)$. Moreover, we consider the high-dimensional limit $n, d \to \infty$ with $\alpha = n/d$ fixed.

Our results depend on mild assumptions on the link function, stated formally in Assumption 12.

In order to quantify the recovery of each individual index, or concept, we define following metric.

Definition 3 (Matrix-MSE) *Let $\boldsymbol{w}_k^\star \in \mathbb{S}^{d-1}(\sqrt{d})$ denote one of the target indices $k \in [d]$. We define the matrix-MSE associated to a predictor $\boldsymbol{w} \in \mathbb{R}^d$ as $\mathrm{mse}_k(\boldsymbol{w}) := \frac{1}{d^2}\mathbb{E}\left[\|\boldsymbol{w}\boldsymbol{w}^\top - \boldsymbol{w}_k^\star \boldsymbol{w}_k^{\star\top}\|_F^2\right]$.*

Additionally, we characterize the *weak recovery* transition for the index k, *i.e.* the minimum number of samples required by an estimator to correlate non-trivially with a specific concept.

Definition 4 ($k-$critical threshold) *Given an estimator $\hat{\boldsymbol{w}}_k \in \mathbb{R}^d$ of the signal direction $\boldsymbol{w}_k^\star$ (i.e. a measurable function of the training data $\boldsymbol{X}, \boldsymbol{y}$), the $k-$critical threshold is defined as the minimum sample complexity such that the estimator exhibits a finite overlap with $\boldsymbol{w}_{\star,k}$, namely* $\inf\left\{\alpha > 0 : d^{-1}|\langle \hat{\boldsymbol{w}}_k, \boldsymbol{w}_{\star,k}\rangle| = \Theta_d(1)\right\}$.

The recovery of $\mathrm{span}(\boldsymbol{W}^\star)$ is assessed by the sum of the squared errors for each individual direction, weighted by the contribution of each concept to the target variance.

Definition 5 (Weighted MSE) *Let $\boldsymbol{W} \in \mathbb{R}^{m_\star \times d}$. We define the weighted mean-squared error as*

$$\mathrm{MSE}_\gamma(\boldsymbol{W}) := \sum_{k\in K} (a_k^\star)^2 \mathrm{mse}_k(\boldsymbol{w}_k). \tag{4}$$

For the second goal — showing that neural networks can efficiently learn $f_\star$ — we will focus on two-layer neural networks:

$$f(\boldsymbol{x}; \boldsymbol{\Theta}) := \boldsymbol{a}^\top \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \tag{5}$$

with weights $\boldsymbol{\Theta} := \left(\boldsymbol{a} \in \mathbb{R}^p,\ \boldsymbol{W} \in \mathbb{R}^{p\times d}\right)$ and a generic activation function $\sigma : \mathbb{R} \to \mathbb{R}$. As we shall see, two-layer neural networks can agnostically learn with respect to the model, i.e. without knowledge of the individual tasks g_k nor the details in Def. 1. We quantify the generalization capacity of this model through its excess risk

$$R(\boldsymbol{\Theta}) = \mathbb{E}\left[\left(f_\star(\boldsymbol{x}) - f(\boldsymbol{x}; \boldsymbol{\Theta})\right)^2\right]. \tag{6}$$

3. Main results

3.1. Bayes-optimal rates for feature recovery

As a first result, we characterize the information-theoretic limits for the recovery of the subspace $\mathrm{span}(\boldsymbol{W}_\star)$ in terms of the decay of the target coefficients $\boldsymbol{a}^\star$, the sample complexity α and the subspace dimension $m_\star$.

Definition 6 (Optimal MSE) *Let $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ be drawn as in def. 2. Then, the optimal (weighted) mean-square error is achieved by the posterior average*

$$\mathrm{MMSE}_\gamma := \sum_{k=1}^{m_\star} \frac{(a_k^\star)^2}{d^2} \mathbb{E}\left[\|\mathbb{E}[\boldsymbol{w}_k \boldsymbol{w}_k^\top \mid \mathcal{D}] - \boldsymbol{w}_k^\star \boldsymbol{w}_k^{\star\top}\|_F^2\right]$$

.

By definition, MMSE_γ is a lower-bound for the smallest achievable weighted mean-squared error $\mathrm{MSE}_\gamma(\hat{\boldsymbol{W}})$ by any estimator $\hat{\boldsymbol{W}}$ that is a function of the dataset $\mathcal{D}$. Our first main result quantifies precisely the rates of MMSE_γ as $\alpha \gg 1$, which by construction define the optimal scaling laws for subspace reconstruction in the class of scale-free hierarchical multi-index models. A proof of this result is discussed in Appendix 5.

Theorem 7 (Optimal scaling-laws) *In the setting of Definitions 1, 2, under Assumption 12, for $\alpha, m_\star \gg 1$, the Bayes-optimal mean-squared error satisfies*

$$\mathrm{MMSE}_\gamma = \Theta_{\alpha, m_\star} \begin{cases} \min(\alpha^{-1+\frac{1}{2\gamma}}, \frac{m_\star}{\alpha}) & \gamma > 1/2, \\ \min(1, \frac{m_\star}{\alpha}), & \gamma < 1/2 \end{cases} \tag{7}$$

Moreover, the k-critical threshold of the Bayes estimator satisfies $\alpha_k^{\mathrm{Bayes}} = \Theta_k(k^{2\gamma} m_\star^{\max((1-2\gamma),0)})$

3.2. Optimal Agnostic subspace recovery

While Theorem 7 provides a fundamental benchmark, an equally important question is whether the optimal reconstruction rates can be *efficiently* achieved by an algorithm which is agnostic of the underlying data distribution. In this section, we give an affirmative answer to this question by constructing an explicit spectral method achieving these rates.

Definition 8 (Spectral estimator) *Given a pre-processing function $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ on the labels, consider the symmetric random matrix*

$$\boldsymbol{T} = \sum_{i=1}^{n} \mathcal{T}(y_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top, \tag{8}$$

with spectrum $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. Define $r := \min\{i \in \mathbb{N} \,|\, \lambda_{i+1} - \lambda_{i+2} < C/\sqrt{d}\}$, for some arbitrary constant C. We define the spectral estimator $\hat{\boldsymbol{W}}^{\rm sp} = (\hat{\boldsymbol{w}}_1, ..., \hat{\boldsymbol{w}}_r)^\top \in \mathbb{R}^{r\times d}$ where $\hat{\boldsymbol{w}}_k$ is the $k \in [r]$ eigenvector of $\boldsymbol{T}$ corresponding to λ_k, normalized such that $\|\hat{\boldsymbol{w}}_i\|^2 = d$.

We refer to Appendix 5.2.1 for an interpretation of the spectral estimator in terms of a gradient descent algorithm. Crucially, we consider $\mathcal{T}$ to be data-agnostic, satisfying only the following mild condition.

Assumption 9 *Given $Y \sim \mathcal{N}\left(\sum_{k=1}^{m_\star} a_k^\star g_k(z_k), \Delta\right)$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{m_\star})$, assume that $\mathbb{P}(\mathcal{T}(Y) = 0) < 1$ and $\mathcal{T}$ is bounded and there exists $\tau > 0$ such that $\tau = \inf\{c : \mathbb{P}(\mathcal{T}(Y) < c) = 1\}$.*

Our second main theoretical result shows that the data-agnostic spectral method defined above indeed achieves the optimal reconstruction rates of Theorem 7.

Theorem 10 *In the setting of Definitions 1, 2, under Assumption 12, denote $\hat{\boldsymbol{W}} = (\hat{\boldsymbol{W}}^{\rm sp}, \boldsymbol{0}_{(m_\star - r)\times d}) \in \mathbb{R}^{m_\star \times d}$, where $\hat{\boldsymbol{W}}^{\rm sp} \in \mathbb{R}^{r\times d}$, $r \leq m_\star$, is the spectral estimator defined in 8, with $\mathcal{T}$ satisfying Assumption 9. Then, for $\alpha, m_\star \gg 1$, denoting the spectral estimator's k-critical threshold as $\alpha_k^{\rm sp}$, we have that $\mathrm{MSE}_\gamma(\hat{\boldsymbol{W}}) = \Theta_{\alpha,m_\star}(\mathrm{MMSE}_\gamma)$, and $\alpha_k^{\rm sp} = \Theta_k(\alpha_k^{\rm Bayes})$.*

Figure 2 shows the decay of the weighted MSE for three scale-free hierarchical model variants as a function of the sample complexity α, based on finite-size experiments, using the spectral method introduced in Definition 8. Figure 3 illustrates the sequential emergence of new concepts as sample complexity grows, visualized as spikes detaching from the eigenvalue bulk of the matrix $\boldsymbol{T}$ in eq. (8).

3.3. Learning multi-index with neural networks

Finally, we turn our attention to the problem of learning hierarchical multi-index targets function with a two-layer neural network and show that the excess risk associated with a suitable training procedure described in Algorithm 1 in Appendix 6, achieves the optimal rates in Theorem 7.

Theorem 11 *In the setting of Definitions 1, 2, under Assumption 12, with $g_k^\star$, $k \in [m_\star]$, polynomials of finite degree. There exist n_0 such that, for $n > n_0$, $p = \omega(n^{1/2})$, $\lambda = \Theta(n^{-1/2})$, the excess risk of eq. (6) for a two-layer neural network eq. (5), with σ bounded and continuous, trained according to Algorithm 1 satisfies*

$$R_{\rm NN} = \Theta(\mathrm{MMSE}_\gamma) + O(n^{-1/2}). \tag{9}$$

References

[1] Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2552–2623. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

[2] Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

[3] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[4] Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

[5] Dmitry Babichev and Francis Bach. Slice inverse regression with score functions. *Electronic Journal of Statistics*, 12(1):1507 – 1543, 2018. doi: 10.1214/18-EJS1428. URL https://doi.org/10.1214/18-EJS1428.

[6] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

[7] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Phase transitions, optimal errors and optimality of message-passing in generalized linear models. *CoRR*, abs/1708.03395, 2017. URL http://arxiv.org/abs/1708.03395.

[8] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. doi: 10.1073/pnas.1802705116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1802705116.

[9] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2): 764–785, 2011.

[10] Gérard Ben Arous, Murat A Erdogdu, N Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. *arXiv preprint arXiv:2508.03688*, 2025.

[11] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020.

[12] Tony Bonnaire, Giulio Biroli, and Chiara Cammarota. The role of the time-dependent hessian in high-dimensional optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):083401, 2025.

[13] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

[14] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *Proceedings of the 41st International Conference on Machine Learning*, 2024. arXiv preprint arXiv:2402.01092.

[15] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8): 084002, 2025.

[16] Guillaume Braun, Bruno Loureiro, Ha Quang Minh, and Masaaki Imaizumi. Fast escape, slow convergence: Learning dynamics of phase retrieval under power-law data. *arXiv preprint arXiv:2511.18661*, 2025.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[18] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[19] Andreas Christmann and Ingo Steinwart. Support vector machines. *Systems and Virtualization Management. Standards and New Technologies*, 2008.

[20] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

[21] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Error scaling laws for kernel classification under source and capacity conditions. *Machine Learning: Science and Technology*, 4(3):035033, 2023.

[22] Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

[23] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

[24] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024.

[25] Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. *Advances in Neural Information Processing Systems*, 37:104630–104693, 2024.

[26] Leonardo Defilippis, Yatin Dandi, Pierre Mergny, Florent Krzakala, and Bruno Loureiro. Optimal spectral transitions in high-dimensional multi-index models, 2025. URL https://arxiv.org/abs/2502.02545.

[27] Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. *arXiv preprint arXiv:2509.24882*, 2025.

[28] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006.

[29] David L Donoho, Matan Gavish, and Andrea Montanari. The phase transition of matrix recovery from Gaussian measurements matches the minimax mse of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21):8405–8410, 2013.

[30] Rishabh Dudeja, Yue M. Lu, and Subhabrata Sen. Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.

[31] Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of ERM in overparameterized quadratic networks. *NeurIPS 2025*, 2025. arXiv preprint arXiv:2505.17958.

[32] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 15 (4):335–536, 2022. ISSN 1935-8237. doi: 10.1561/2200000092. URL http://dx.doi.org/10.1561/2200000092.

[33] Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.

[34] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

[35] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in neural information processing systems*, 35:30016–30030, 2022.

[36] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[37] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[38] Filip Kovačević, Yihan Zhang, and Marco Mondelli. Spectral estimators for multi-index models: Precise asymptotics and optimal weak recovery. *arXiv preprint arXiv:2502.01583*, 2025.

[39] Filip Kovačević, Zhang Yihan, and Marco Mondelli. Spectral estimators for multi-index models: Precise asymptotics and optimal weak recovery. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3354–3404. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/kovacevic25a.html.

[40] Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under Zipf's law. *arXiv preprint arXiv:2505.19227*, 2025.

[41] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[42] Yue M Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional nonconvex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541, 2020.

[43] Antoine Maillard, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. Construction of optimal spectral methods in phase retrieval. In *Mathematical and Scientific Machine Learning*, pages 693–720. PMLR, 2022.

[44] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[45] Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

[46] Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

[47] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, 19(3):703–773, Jun 2019. ISSN 1615-3383. doi: 10.1007/s10208-018-9395-y. URL https://doi.org/10.1007/s10208-018-9395-y.

[48] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[49] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37:16459–16537, 2024.

[50] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

[51] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10): 6976–6994, 2011.

[52] Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D Lee. Emergence and scaling laws in SGD learning of shallow neural networks. *arXiv preprint arXiv:2504.19983*, 2025.

[53] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/61b1fb3f59e28c67f3925f3c79be81a1-Paper.pdf.

[54] A Saxe, J McClelland, and S Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of the International Conference on Learning Represenatations 2014*. International Conference on Learning Represenatations 2014, 2014.

[55] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

[56] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

[57] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

[58] Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix theory analysis of neural network weight matrices. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

[59] Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborova, Bruno Loureiro, and Florent Krzakala. Fundamental computational limits of weak learnability in high-dimensional multi-index models. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2467–2475. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/troiani25a.html.

[60] Zhichao Wang, Andrew Engel, Anand D Sarwate, Ioana Dumitriu, and Tony Chiang. Spectral evolution and invariance in linear-width neural networks. *Advances in neural information processing systems*, 36:20695–20728, 2023.

[61] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[62] Roman Worschech and Bernd Rosenow. Analyzing neural scaling laws in two-layer networks with power-law data spectra. *arXiv preprint arXiv:2410.09005*, 2024.

[63] Arie Wortsman and Bruno Loureiro. Kernel ridge regression under power-law data: spectrum and generalization. *arXiv preprint arXiv:2510.04780*, 2025.

[64] Ming Yuan. On the identifiability of additive index models. *Statistica Sinica*, pages 1901–1911, 2011.

[65] Bohan Zhang, Zihao Wang, Hengyu Fu, and Jason D. Lee. Neural networks learn generic multi-index models near information-theoretic limit, 2025. URL https://arxiv.org/abs/2511.15120.

[66] Qiuyun Zou and Hongwen Yang. A concise tutorial on approximate message passing. *arXiv preprint arXiv:2201.07487*, 2022.

4. Further related works

Most theoretical work on neural scaling laws focus on effectively linear models, such as kernel methods and random features, where the generalization behavior can be characterized through spectral properties of fixed representations [3, 6, 13, 18, 20, 21, 25, 40, 45, 49, 57]. Exceptions have focused either in the joint-training of both layers for linear networks [14, 15, 62] or non-linear settings with fixed-features [16, 62, 63]. A central difference is that these works introduce scaling through the covariate distribution, while in our work it is induced by the hierarchical nature of the task. Moreover, our work departs from this literature by addressing scaling laws in a genuinely non-linear, feature-learning regime.

Closer to us are a recent line of work analyzing two-layer networks with structured first-layer weights and hierarchical multi-index target [10, 27, 52]. In particular, [10, 52] studied one-pass SGD in this setting, with [10] focusing on the case of a quadratic neural network architecture. The main difference with this work is that we analyze full-batch ERM, characterizing the optimal scaling laws and showing that they can be achieved computationally by a gradient-descent like algorithm, thus providing a fundamental benchmark for the SGD rates in these works.

Complementary, [27] derived rates for ERM in quadratic neural networks trained on quadratic targets. Our results show that the rates from [27] are universal a large class of target functions (any generative exponent two functions in the language of [22]), closing a gap between the purely quadratic setting and general hierarchical multi-index targets. The key underlying these universal scaling laws is the combination *quasi-sparsity* of the target representation, encoded by a heavy-tailed spectrum that induces an ordered hierarchy of feature strengths [28, 44], and an implicit *rank-sparsity* bias, leading to LASSO-like behavior [51].

On a technical level, our work builds on the toolbox of approximate message passing (AMP) and its associated state evolution [9, 11, 29–33, 36, 66]. In particular, our results leverage the connection between AMP and Bayes-optimal estimation for single- [8] and multi-index [4, 59] models. Similarly, we build up on the literature on optimal spectral methods derived from AMP for single- [22, 42, 43, 47] and multi-index [26, 38] functions.

5. Proof sketches of Theorems 7 and 10

In this section, we prove the information-theoretic results stated in Theorem 7. In particular, we derive matching bounds for the MMSE_γ, defined 6, and the k-critical threshold α_k (Definition 4). Our results depend on the following additional assumption of the link functions.

Assumption 12 *For each index $k \in [m_\star]$, g_k is an even function. Moreover, for $z \sim \mathcal{N}(0,1)$, for some constants $C, D > 0$, independent of $k, n, d, \alpha, m_\star$,*

$$D < \mathbb{E}_z[g_k^2(z)] < C, \quad D < |\mathbb{E}_z[g_k''(z)]| < C. \tag{10}$$

Remark 13 *The parity assumption on g_k ensures that learning $\mathrm{span}(\boldsymbol{W}_\star)$ is* non-trivial*, by ruling out linear correlations that would allow recovery at arbitrarily small sample complexity. Indeed, given $\boldsymbol{z}$ as in eq. (2) and y as in eq. 3, when $\mathbb{E}[\boldsymbol{z} \mid y] = 0$ a.s.—which holds here by parity—no efficient algorithm can even weakly recover $\mathrm{span}(\boldsymbol{W}_\star)$ below the critical threshold*

$$\alpha_c := \left(\sup_{\boldsymbol{M} \in \mathbb{S}^+_{m_\star},\, \|\boldsymbol{M}\|_F = 1} \|\mathbb{E}_y \boldsymbol{G}(y) \boldsymbol{M} \boldsymbol{G}(y)\|_F \right)^{-1}, \tag{11}$$

with $\boldsymbol{G}(y) := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_{m_\star} \mid Y = y]$ *[8, 22, 42, 47, 59].*

The condition $\mathbb{E}_z[g_k^2(z)] < C$ *controls the label variance, while the lower bound on* $\mathbb{E}_z[g_k''(z)]$ *ensures detectability in the proportional regime* $n = \Theta(d)$ *(equivalently, a generative exponent equal to* 2*), so that the relative difficulty of each index is governed solely by the decay of* $a_k^\star$. [1]

Hierarchical multi-index models with generative exponent equal to 2 includes the vast majority of cases of interest, while the class of models with exponent larger than 2 *mostly includes fine-tuned examples [7, 22].*

In Appendix 5.1, we analyze an oracle estimator that exploits additional information to achieve a weighted mean-squared error smaller than MMSE_γ. Finally, in Appendix 5.2, we complete the proof by characterizing the scaling laws and k-critical thresholds of the spectral estimator defined in 8, which simultaneously establishes the results in Theorem 10.

5.1. Lower bound

Definition 14 (Oracle Estimator) *Consider a dataset* $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i\in[n]}$*, where the labels are generated by a multi-index model* $y \sim \mathsf{P}(\cdot|\langle \boldsymbol{w}_1^\star, \boldsymbol{x}\rangle, \dots, \langle \boldsymbol{w}_{m_\star}^\star, \boldsymbol{x}\rangle)$*, with weights* $\boldsymbol{W}^\star$ *drawn from a distribution* $\mathsf{P}_{\boldsymbol{W}}$*. We define the Oracle Estimator as the matrix* $\boldsymbol{W}^{\mathrm{oracle}} \in \mathbb{R}^{m\times d}$ *with rows*

$$\boldsymbol{w}_k^{\mathrm{oracle}} = \arg\min_{\boldsymbol{w}} \mathbb{E}\left[\|\boldsymbol{w}_k\boldsymbol{w}_k^\top - \boldsymbol{w}_k^\star\boldsymbol{w}_k^{\star\top}\|_F^2 \,|\, \mathcal{D}, \{\boldsymbol{w}_h\}_{h\neq k}\right], k \in [m_\star]. \tag{12}$$

Analogously to the Bayes-optimal estimator case, the weighed mean-squared error $\mathrm{MSE}_\gamma(\boldsymbol{W}^{\mathrm{oracle}})$ is lower bounded by

$$\mathrm{MMSE}_\gamma^{\mathrm{oracle}} = \frac{1}{d^2}\sum_{k=1}^{m_\star}(a_k^\star)^2\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{w}_k\boldsymbol{w}_k^\top|\mathcal{D}, \{\boldsymbol{w}_h^\star\}_{h\neq k}] - \boldsymbol{w}_k^\star\boldsymbol{w}_k^{\star\top}\|_F^2\right]. \tag{13}$$

As a consequence of the law of total variance

$$d^2\,\mathrm{mmse}_k := \mathbb{E}\left[\mathrm{Cov}\left(\boldsymbol{w}_k\boldsymbol{w}_k^\top \,|\, \mathcal{D}\right)\right] \tag{14}$$

$$= \mathbb{E}\left[\mathrm{Cov}\left(\boldsymbol{w}_k\boldsymbol{w}_k^\top \,|\, \mathcal{D}, \{\boldsymbol{w}_h\}_{h\neq k}\right)\right] + \underbrace{\mathbb{E}\left[\mathrm{Cov}_{\{\boldsymbol{w}_h\}_{h\neq k}}\left(\mathbb{E}\left[\boldsymbol{w}_k\boldsymbol{w}_k^\top \,|\, \mathcal{D}, \{\boldsymbol{w}_h\}_{h\neq k}\right]\right)\right]}_{\geq 0} \tag{15}$$

$$\geq d^2\,\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{w}_k\boldsymbol{w}_k^\top|\mathcal{D}, \{\boldsymbol{w}_h^\star\}_{h\neq k}] - \boldsymbol{w}_k^\star\boldsymbol{w}_k^{\star\top}\|_F^2\right]. \tag{16}$$

Therefore,

$$\mathrm{MMSE}_\gamma^{\mathrm{oracle}} \leq \mathrm{MMSE}_\gamma. \tag{17}$$

We now focus on the specific setting of our interest defined in Section 2. For each index $k \in [m_\star]$, conditioning on $\{\boldsymbol{w}_h\}_{h\neq k}$, the problem of the optimal estimation of $\boldsymbol{w}_k^\star$ becomes statistically

1. Our analysis can be easily adapted to the less restrictive assumption that there exists an integer $\beta \geq 1$ such that $D < |\mathbb{E}_Z[g_k^\beta(Z)(Z^2-1)]| < C$. This would result in a simple modification of the scaling laws and not affect the overall message of the present manuscript.

equivalent to the Bayes-optimal estimation given a dataset $\mathcal{D}_k := \{(\boldsymbol{x}_\nu, \overline{y}_\nu)\}_{\nu\in[n]}$, where the labels are generated by the single-index model

$$\overline{y}_i = a_k^\star g_k(\langle \boldsymbol{w}_k^\star, \boldsymbol{x}_i\rangle) + \sqrt{\Delta}\xi_i, \tag{18}$$

where ξ_i is the same label noise as in the original dataset. We can now characterize the oracle estimator using the results from the literature on *single-index models*, in particular [8]. In Appendix 7 we show that the information-theoretic weak recovery threshold for single-index models is a monotonic decreasing function of the signal-to-noise ratio. Combined with Assumption 12, this implies that the sequence of k-critical thresholds α_k is bounded by strictly increasing functions.[2] In particular, the first result in Corollary 19 implies that the sequence is bounded by

$$\alpha_k^{\text{oracle}} = \Theta\left((a_k^\star)^{-2}\right) = \Theta(k^{2\gamma} m_\star^{(1-2\gamma)_+}), \tag{19}$$

with $(x)_+ = \max(0, x)$. Further, denoting by k_α the largest index k such that $\alpha > \alpha_k$, i.e. $k_\alpha = \max(m_\star, \Theta(\alpha^{1/(2\gamma)}))$ for $\gamma > 1/2$ or $k_\alpha = \max(m_\star, \Theta(m_\star^{-1+1/(2\gamma)}\alpha^{1/(2\gamma)}))$ for $\gamma < 1/2$, and exploiting the second result in Corollary 19,

$$\text{MMSE}_\gamma^{\text{oracle}} = \left(\sum_{k=1}^{k_\alpha} + \sum_{k=k_\alpha+1}^{m_\star}\right)(a_k^\star)^2 \mathbb{E}\left[\|\mathbb{E}[\boldsymbol{w}_k\boldsymbol{w}_k^\top|\mathcal{D}, \{\boldsymbol{w}_h^\star\}_{h\neq k}] - \boldsymbol{w}_k^\star\boldsymbol{w}_k^{\star\top}\|_F^2\right] \tag{20}$$

$$\geq C\left(\sum_{k=1}^{k_\alpha}\frac{(a_k^\star)^2}{(a_k^\star)^2\,\alpha} + \sum_{k=k_\alpha+1}^{m_\star}(a_k^\star)^2\right) \tag{21}$$

$$= \begin{cases} \Theta\left(\alpha^{-1+1/(2\gamma)}\right), & \gamma > 1/2,\ \alpha \ll m_\star^{2\gamma} \\ \Theta\left(m_\star/\alpha\right), & \gamma > 1/2,\ \alpha \gg m_\star^{2\gamma} \\ \Theta\left(1\right), & \gamma < 1/2,\ \alpha \ll m_\star \\ \Theta\left(m_\star/\alpha\right), & \gamma < 1/2,\ \alpha \gg m_\star \end{cases} \tag{22}$$

Note that we have used

$$\sum_{k=k_\alpha+1}^{m_\star} k^{-2\gamma} = \begin{cases} \Theta(k_\alpha^{1-2\gamma}), & \gamma > 1/2, \\ \Theta(m_\star^{1-2\gamma}), & \gamma < 1/2. \end{cases} \tag{23}$$

In eq. (21), the first term is a lower bound to the (weighted) mean-squared error of weakly recovered features, while the second corresponds to the underfitting contribution of the unlearned ones.

5.2. Upper Bound: Spectral Method

By definition, any estimator that is a function of the dataset only has a weighted mean-squared error larger than MMSE_γ. In this section we consider the spectral method defined in 8. Note that, as the spectral method 8 retrieves $r \leq m_\star$ directions, we construct the full estimator by filling the remaining columns with zeros to evaluate MSE_γ in Def. 5. This zero-padding is harmless: the added zeros do not impact the derived scaling laws or the estimator's agnostic nature. The following Theorem, proven in [39], is valid for generic Gaussian multi-index models with $m_\star$ indices. We refer to the original work for further details. In the rest of the section we denote by $\mathbb{E}$ the expected value with respect to $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}_{m_\star}, \boldsymbol{I}_{m_\star})$ and $y \sim \mathcal{N}\left(\sum_{k=1}^{m_\star}(a_k^\star)^2 g_k(z_k),\ \Delta\right)$.

2. If $g_k = g$, $\forall k$, the sequence α_k^{oracle} itself is strictly increasing.

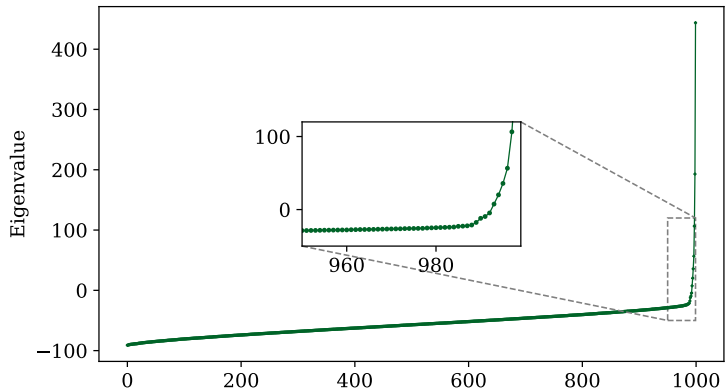

Figure 1: Empirical spectrum of $\boldsymbol{T}$ eq. (8) with preprocessing $\mathcal{T}(y) = y/(1+|y|)$, for a hierarchical multi-index model with $g_k(z) = \frac{1}{2}\mathrm{He}_2(z) + \frac{1}{2\cdot 4!}\mathrm{He}_4(z)$ and $\gamma = 1.3$. The covariates dimension is $d = 1000$, while the feature space dimension is $m_\star = 20$. The figure illustrates the change in scale of the eigenvalue gaps, transitioning from the informative spikes ($\Theta_d(1)$) to the uninformative bulk ($o_d(1)$). This behavior forms the basis of the selection method described in Def. 8.

Theorem 15 (Theorem 4.1 in [39]) *Let $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ be a preprocessing function subject to Assumption 9 and $\boldsymbol{T}$ defined as in eq. 8. Let $t_1 \geq t_2 \geq \ldots \geq t_r \geq \tau$, for some $r \in [m]$, be all the solutions to*

$$\det\left(\alpha\mathbb{E}\left[\frac{(\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_m)\mathcal{T}(y)}{t - \mathcal{T}(y)}\right] - \boldsymbol{I}_p\right) = 0 \tag{24}$$

such that

$$t_k \geq \bar{t}_\alpha := \operatorname*{arg\,min}_{t\geq\tau} \ \zeta_\alpha(t), \quad \forall k \in [j], \tag{25}$$

where

$$\zeta_\alpha(t) := t\left(1 + \alpha\mathbb{E}\left[\frac{\mathcal{T}(y)}{t - \mathcal{T}(y)}\right]\right). \tag{26}$$

Then, denote $\lambda_1^{\boldsymbol{T}}, \ldots, \lambda_{m_\star}^{\boldsymbol{T}}$ the largest $m_\star$ eigenvalues of $\boldsymbol{T}$. For the top r eigenvalues it holds that

$$\lambda_1^{\boldsymbol{T}}, \ldots, \lambda_r^{\boldsymbol{T}} \xrightarrow{\text{a.s.}} \zeta_\alpha(t_1), \ldots, \zeta_\alpha(t_r), \tag{27}$$

and for the remaining $m_\star - r$ eigenvectors it holds that

$$\lambda_{r+1}^{\boldsymbol{T}}, \ldots, \lambda_{m_\star}^{\boldsymbol{T}} \xrightarrow{\text{a.s.}} \zeta_\alpha(\bar{t}_\alpha). \tag{28}$$

As a first result, we derive a bound for α_k.

Consider the matrix in eq. (24)

$$\boldsymbol{G}(y) := \mathbb{E}\left[\frac{(\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_m)\mathcal{T}(y)}{t - \mathcal{T}(y)}\right] = \mathbb{E}_y\left[\frac{\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_m|y]\mathcal{T}(y)}{t - \mathcal{T}(y)}\right]. \tag{29}$$

It is straightforward to show that it is a diagonal matrix, due to the parity of each function g_k. Indeed, given $k \neq h$,

$$G_{kh}(y) = \mathbb{E}[z_k z_h|y] \propto \int e^{-\|\boldsymbol{z}\|^2/2} \exp\left(-\left(y - \sum_{k=1}^{m} a_k g_k(z_k)\right)^2 /(2\Delta)\right) z_h z_k \, \mathrm{d}\boldsymbol{z} = 0. \quad (30)$$

For simplicity, we denote $G_k(y) := G_{kk}(y)$. Therefore, the solutions of eq. (24) coincides with the solutions of

$$\alpha^{-1} = \mathbb{E}_y \left[\frac{\mathcal{T}(y)}{t - \mathcal{T}(y)} G_k(y)\right], \quad k \in [m]. \quad (31)$$

Further, note that, by definition $\bar{t}_\alpha$ is the solution of

$$\alpha^{-1} = \mathbb{E}_y \left[\left(\frac{\mathcal{T}(y)}{t - \mathcal{T}(y)}\right)^2\right]. \quad (32)$$

A spectral transition occurs if, for sample complexity α and index $k \in [m]$, the value $\bar{t}_\alpha$ is also solution of eq. (31). Indeed by Theorem 4.2 in [39] – which characterizes the overlap of the principal r eigenvalues, such sample complexity corresponds to the k-critical threshold α_k. This implies, for $\alpha = \alpha_k^{\mathrm{sp}}$

$$\begin{cases} (\alpha_k^{\mathrm{sp}})^{-1} = \mathbb{E}_y \left[\frac{\mathcal{T}(y)}{t - \mathcal{T}(y)} G_k(y)\right] \\ (\alpha_k^{\mathrm{sp}})^{-1} = \mathbb{E}_y \left[\left(\frac{\mathcal{T}(y)}{t - \mathcal{T}(y)}\right)^2\right]. \end{cases} \quad (33)$$

In order to characterize the threshold, we consider the following expansions for small SNR $a_k\star \ll 1$. Define Z the marginal distribution of y and consider the Fourier representation $(2\pi)^{-1/2} e^{-x^2/2} = (2\pi)^{-1} \int \mathrm{d}\omega e^{i\omega x - \omega^2/2}$, then

$$\mathsf{Z}(y) := \frac{1}{\sqrt{2\pi\Delta}} \mathbb{E}_{\boldsymbol{z}} \left[\exp\left(-\left(y - \sum_{k=1}^{m_\star} a_k^\star g_k(z_k)\right)^2 /(2\Delta)\right)\right] \quad (34)$$

$$= \frac{1}{2\pi} \int \mathrm{d}\omega e^{i\omega y - \omega^2 \Delta/2} \left(\prod_{h \neq k} \mathbb{E}_z \left[e^{-i\omega a_h^\star g_h(z)}\right]\right) \left[e^{-i\omega a_k^\star g_k(z)}\right] \quad (35)$$

$$= \frac{1}{2\pi} \int \mathrm{d}\omega e^{i\omega y - \omega^2 \Delta/2} \left(\prod_{h \neq k} \mathbb{E}_z \left[e^{-i\omega a_h^\star g_h(z)}\right]\right) \left[1 + \sum_{\beta \geq 1} \frac{(-i\omega a_k^\star)^\beta}{\beta!} \mathbb{E}_z[g_k^\beta(z)]\right] \quad (36)$$

$$= \frac{1}{2\pi} \int \mathrm{d}\omega e^{i\omega y - \omega^2 \Delta/2} \left(\prod_{h \neq k} \mathbb{E}_z \left[e^{-i\omega a_h^\star g_h(z)}\right]\right) + O\left(a_k^\star \mathsf{Z}'(y)\right), \quad (37)$$

where we have used the identity $\int \mathrm{d}\omega e^{i\omega y}\omega f(\omega) = \frac{\partial}{\partial y}\int \mathrm{d}\omega e^{i\omega y} f(\omega)$. Similarly, one can bound the following quantity

$$\left| G_k(y) + a_k^\star \mathbb{E}_z[g_k(z)(z^2-1)]\frac{\mathsf{Z}'(y)}{\mathsf{Z}(y)} \right| \tag{38}$$

$$= \frac{1}{\mathsf{Z}(y)} \left| \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_p)} \left[\mathsf{P}(y|\boldsymbol{z})(z_k^2-1) + a_k^\star \mathbb{E}_z[g_k(z)(z^2-1)]\partial_y \mathsf{P}(y|\boldsymbol{z}) \right] \right| \tag{39}$$

$$= \frac{1}{2\pi \mathsf{Z}(y)} \left| \int_{\mathbb{R}} \mathrm{d}\omega e^{i\omega y - \omega^2\Delta/2} \prod_{h\neq k} \mathbb{E}_z[e^{-i\omega a_h^\star g_h(z)}] \left(\mathbb{E}_z\left[e^{-i\omega a_k^\star g_k(z)}(z^2-1) \right] + i\omega a_k^\star \mathbb{E}_z[g_k(z)(z^2-1)]\mathbb{E}_z[e^{-i\omega a_k^\star g_k(z)}] \right) \right| \tag{40}$$

$$= \frac{1}{2\pi \mathsf{Z}(y)} \left| \int_{\mathbb{R}} \mathrm{d}\omega e^{i\omega y - \omega^2\Delta/2} \prod_{h\neq k} \mathbb{E}_z[e^{-i\omega a_h^\star g_h(z)}] O\left((a_k^\star)^2\omega^2 \right) \right| \tag{41}$$

$$= \frac{1}{2\pi \mathsf{Z}(y)} O\left((a_k^\star)^2 \left| \frac{\partial^2}{\partial y^2} \int_{\mathbb{R}} \mathrm{d}\omega e^{i\omega y - \omega^2\Delta/2} \prod_{k\neq h} \mathbb{E}_z[e^{-i\omega a_k^\star g_k(z)}] \right| \right) \tag{42}$$

$$= O\left((a_k^\star)^2 \left| \frac{\mathsf{Z}''(y)}{\mathsf{Z}(y)} \right| \right). \tag{43}$$

Note that $\mathbb{E}_z[g_k(z)(z^2-1)] \overset{\text{parts}}{=} \mathbb{E}_z[g_k''(z)] \neq 0 \, \forall k \in [m_\star]$ by Assumption 12. We now look for $t > \tau$ and α that are solutions of eqs. (33). In particular, the oracle k-critical threshold in Appendix 5.1 is a lower bound to the spectral one, therefore $\alpha_k^{\text{sp}} = \Omega(\alpha_k^{\text{oracle}})$ and diverges with k. The second equation of (33) – due to the monotonicity of the RHS, implies that also the solution t diverges with k. At leading order in $a_k^\star$ and t

$$\begin{cases} (\alpha_k^{\text{sp}})^{-1} \approx -\int \mathrm{d}y \mathsf{Z}(y) \frac{\mathcal{T}(y)\mathsf{Z}'(y)}{t\mathsf{Z}(y)} a_k^\star \mathbb{E}_z[g_k(z)(z^2-1)] \approx t^{-1} a_k^\star \mathbb{E}_z[g_k(z)(z^2-1)]\mathbb{E}[\mathcal{T}'(y)] \\ (\alpha_k^{\text{sp}})^{-1} \approx t^{-2}\mathbb{E}[\mathcal{T}(y)^2]. \end{cases} \tag{44}$$

The system is thereby solved for

$$t^{-1} = \Theta(a_k^\star), \qquad \alpha_k^{\text{sp}} = \Theta(\alpha_k^{\text{oracle}}). \tag{45}$$

We can now derive the scaling laws for the spectral weighted mean-squared error. Taking $\alpha \gg \alpha_k^{\text{sp}}$, for k large enough, the solution t_k of eq. (31) scales as $t_k = \Theta(\alpha a_k^\star)$. As a consequence of Theorem 4.2 in [39], specialized in our setting – where we can exploit the diagonality of $\boldsymbol{G}(y)$, we have that the overlap between the eigenvector $\boldsymbol{v}_k$ corresponding to the eigenvalue $\lambda_k^{\boldsymbol{T}}$, and the true signal $\boldsymbol{w}_h^\star$, converges, in the high-dimensional limit, to an overlap

$$m_{hk}^2 := \frac{1}{d^2}\langle \boldsymbol{w}_h^\star, \boldsymbol{v}_k\rangle^2 = \delta_{kh}\frac{\zeta_\alpha'(t_k)}{\zeta_\alpha'(t_k) + \frac{\mathrm{d}}{\mathrm{d}t}R_k(t_k)}, \tag{46}$$

where $R_k(t) = \mathbb{E}[z_k^2\mathcal{T}(y)/(t-\mathcal{T}(y))]$. For $\alpha \gg \alpha_k^{\text{sp}}$ we have that $t_k \gg 1$ and eq. (31) is solved by $t_k = \Theta(a_k^\star\alpha)$. Putting all together

$$m_{hk}^2 = \delta_{kh}\left(1 - \Theta\left(\frac{(\alpha_k^\star)^{-2}}{\alpha}\right)\right). \tag{47}$$

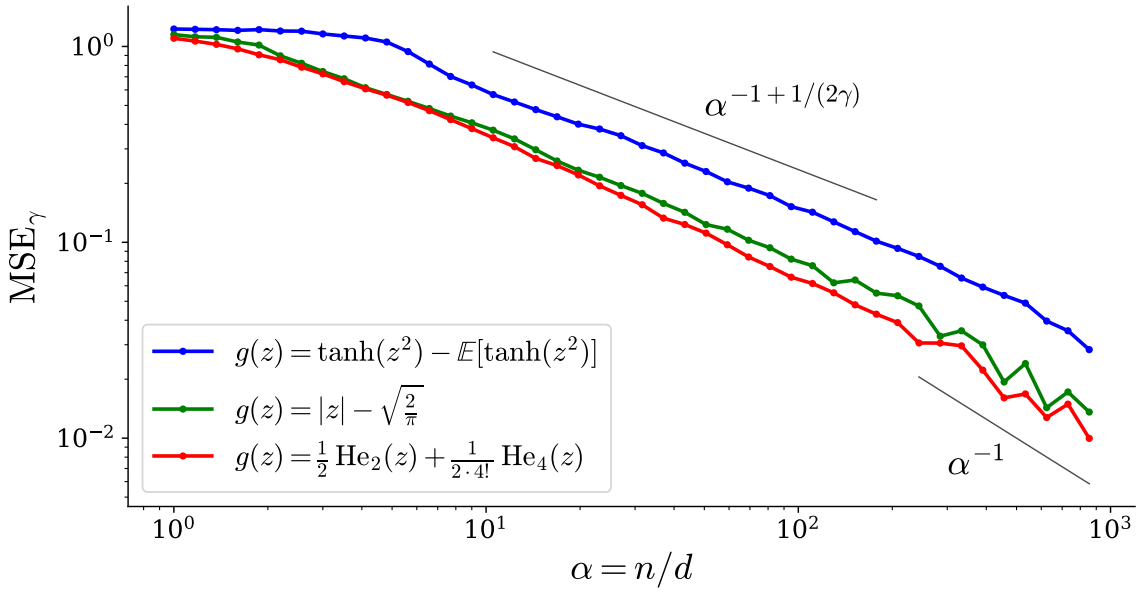

Figure 2: Weighted mean square error MSE_γ – see Def. 5 – of the spectral estimator of Def. 8 with preprocessing function $\mathcal{T}(y) = y/(1+|y|)$, averaged over 70 instances. The target is given by the hierarchical multi-index model 1, with $g_k(z) = g(z)\ \forall k$, stated in the legend, and $a_k^\star \propto k^{-\gamma}$, $\gamma = 1.3$. The covariates dimension is $d = 1000$, the feature space dimension is $m_\star = 10$.

Recall that

$$\mathrm{mse}_k(\boldsymbol{v}_k) = \frac{1}{d^2}\mathbb{E}\left[\|\boldsymbol{v}_k\boldsymbol{v}_k\top\|_{\mathrm{F}}^2 + \|\boldsymbol{w}_k^\star\boldsymbol{w}_k^{\star\top}\|_{\mathrm{F}}^2 - 2\langle\boldsymbol{w}_k^\star, \boldsymbol{v}_k\rangle^2\right] \to 2(1-m_{kk}^2). \tag{48}$$

By repeating computations analogous to the ones in Appendix 5.1, one finds, setting $\hat{\boldsymbol{W}} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_r)^\top$,

$$\mathrm{MMSE}_\gamma(\hat{\boldsymbol{W}} := (\hat{\boldsymbol{W}}^{\mathrm{sp}}, \boldsymbol{0}_{(m_\star - r)\times d})) = \Theta(\mathrm{MMSE}_\gamma^{\mathrm{oracle}}), \tag{49}$$

which proves Theorem 7 and 10.

5.2.1. GD INTERPRETATION OF THE SPECTRAL METHOD

Note that this spectral method also admits an interpretation in terms of a gradient descent algorithm. Indeed, considering a slightly modified loss than the pure square one, this matrix is nothing but the Hessian of the loss [12]. The early dynamics of GD on the first layer of (5), for weights initialized on a sphere with small radius, will thus follow the proposed spectral method. This construction was recently used in [65] to show how two-layer neural networks identify features within the first few GD steps in the first layer weights. More precisely, they show that the early dynamics of a suitable

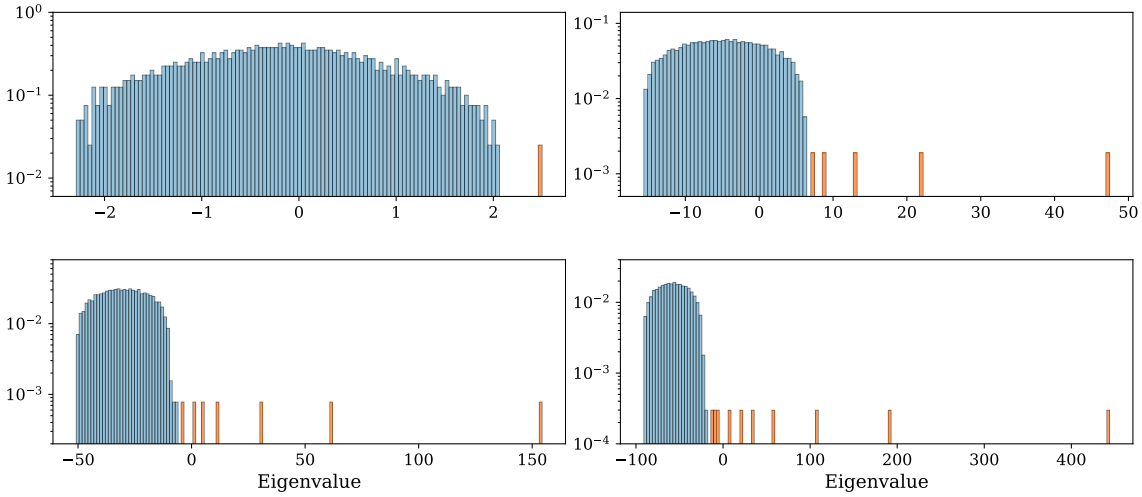

Figure 3: Empirical spectrum density of the matrix $\boldsymbol{T}$ defined in eq. (8), with preprocessing function $\mathcal{T}(y) = y/(1+|y|)$, at different sample complexities, highlighting the sequential emergence of concepts as the sample size increases. The target is given by a hierarchical multi-index model 1, with $g_k(z) = \frac{1}{2}\mathrm{He}_2(z) + \frac{1}{2\cdot 4!}\mathrm{He}_4(z)\ \forall k$ and $a_k \propto k^{-\gamma}$, $\gamma = 1.3$. The covariates dimension is $d = 1000$, the feature space dimension is $m_\star = 20$. (**top left**) $\alpha = 5$, (**top right**) $\alpha = 164$, (**bottom left**) $\alpha = 611$, (**bottom right**) $\alpha = 1638$.

gradient-based algorithm, reads as a power iteration

$$\boldsymbol{w}_k^{t+1} = \boldsymbol{w}_k^t - \nabla_{\boldsymbol{w}_k} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \boldsymbol{a}^\top \sigma(\boldsymbol{W}^t \boldsymbol{x}_i)\right) \tag{50}$$

$$\approx (\boldsymbol{I}_d - a_k \sigma''(0) \boldsymbol{T}) \boldsymbol{w}_k^t, \tag{51}$$

where $\boldsymbol{T}$ has the structure defined in eq. (8) with pre-processing $\mathcal{T} = \ell'(\,\cdot\,, 0)$, and ℓ' denoting the derivative of ℓ with respect to its second argument.

6. Proof sketch of Theorem 11

In this section we prove Theorem 11. This final result demonstrate that neural networks can efficiently learn the target function, achieving the exact same rates as those of the optimal feature subspace recovery. We note that our results do not imply that these rates correspond to the decay of the Bayes risk, defined as $\mathbb{E}\left[(f_\star(\boldsymbol{x}) - \mathbb{E}[f_\star(\boldsymbol{x})|\boldsymbol{x}, \mathcal{D}])^2\right]$, and which constitutes a lower bound for the risk $R(\boldsymbol{\Theta})$ in Eq. (6). Nevertheless, we notice that under the hypothesis of a *fully specialized* Bayes-optimal estimator—i.e., assuming $d^{-1}\hat{\mathbf{W}}^{\text{Bayes}}(\mathbf{W}^\star)^\top$ converges to a diagonal matrix in the high-dimensional limit—the risk for a network with matching architecture and Bayes weights, $f(\mathbf{x}) = \sum_{k=1}^{m_\star} a_k^\star g_k(\hat{\boldsymbol{w}}_k^{\text{Bayes}})$, achieves the rates established in Theorem 7. We leave a more precise analysis of the optimal risk as a future direction.

Algorithm 1 Spectral Initialization and Ridge Training

Input: Dataset $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y}) \in \mathbb{R}^{2n \times d} \times \mathbb{R}^{2n}$, hidden layer width m, regularization λ.
Output: Network parameters $\boldsymbol{W} \in \mathbb{R}^{p \times d}$, $\boldsymbol{a} \in \mathbb{R}^p$.

1. Data Splitting
Partition the dataset $\mathcal{D}$ into two disjoint sets $\mathcal{D}_1$ and $\mathcal{D}_2$ such that $|\mathcal{D}_1| = |\mathcal{D}_2| = n$.

2. Feature Learning (Spectral Initialization on $\mathcal{D}_1$)
Compute the spectral estimator $\hat{\boldsymbol{W}}^{\text{sp}} \in \mathbb{R}^{r \times d}$ Def. 8 using $\mathcal{D}_1$.
Sample a random matrix $\boldsymbol{Z} \in \mathbb{R}^{p \times r}$ with i.i.d. entries drawn from $\mathcal{N}(0, r^{-1})$.
Set first-layer weights: $\boldsymbol{W} = \boldsymbol{Z}\hat{\boldsymbol{W}}$.

3. Readout Training (Ridge Regression on $\mathcal{D}_2$)
Compute the feature matrix $\Psi \in \mathbb{R}^{n \times p}$ on $\mathcal{D}_2$, where $\Psi_{ij} = \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}_i^{(2)} \rangle + b_j)$, $b_j \sim \mathcal{N}(0, 1)$.
Solve for the second-layer weights:
$\boldsymbol{a} = (\Psi^\top \Psi + n\lambda \boldsymbol{I}_p)^{-1} \Psi^\top \boldsymbol{y}^{(2)}$

return $\boldsymbol{W}, \boldsymbol{a}$

Denote $\hat{\boldsymbol{W}}^{\text{sp}} \in \mathbb{R}^{r \times d}$ the spectral estimator defined in 8. In Appendix 5.2 we have shown that each column k correlates with the signal component $\boldsymbol{w}_k^\star$ only, with squared overlap $m_k = 1 - \Theta((\alpha_k^\star)^{-2}/\alpha)$. Denote $\boldsymbol{v}_k$ as the kth column of the spectral estimator and $\eta_k^2 := (\alpha_k^\star)^2/\alpha \ll 1$. Up to negligible corrections $O(\eta_k^2)$

$$\boldsymbol{w}_k^\star = \boldsymbol{v}_k + \eta_k \boldsymbol{\xi}_k, \tag{52}$$

with $\boldsymbol{\xi}_k$ a unit vector orthogonal to all $\{\boldsymbol{v}_h\}_{h\in[r]}$. Given a covariate $\boldsymbol{x} \in \mathbb{R}^d$, define the projected inputs $\boldsymbol{s} \in \mathbb{R}^r$ such that $s_k = \langle \boldsymbol{v}_k, \boldsymbol{x}\rangle$, $k \in [r]$. The first layer pre-activations are given by

$$\boldsymbol{W}\boldsymbol{x} = \boldsymbol{Z}\hat{\boldsymbol{W}}^{\mathrm{sp}}\boldsymbol{x} = \boldsymbol{Z}\boldsymbol{s}. \tag{53}$$

Similarly, up to negligible corrections $O(\eta_k^2)$, the target function

$$f_\star(\boldsymbol{x}) = \sum_{k=1}^{r} a_k^\star g_k(s_k + \eta_k\langle \boldsymbol{\xi}_k, \boldsymbol{x}\rangle) + \sum_{k=r+1}^{m_\star} a_k^\star g_k(\langle \boldsymbol{w}_k^\star, \boldsymbol{x}\rangle) \tag{54}$$

$$= \sum_{k=1}^{r} a_k^\star g_k(s_k) + \sum_{k=1}^{r} a_k^\star g_k'(s_k)\eta_k\langle \boldsymbol{\xi}_k, \boldsymbol{x}\rangle + \sum_{k=r+1}^{m_\star} a_k^\star g_k(\langle \boldsymbol{w}_k^\star, \boldsymbol{x}\rangle). \tag{55}$$

Due to the orthogonality between $\boldsymbol{\xi}_k$ and $\boldsymbol{v}_h$, for any $h, k \in [r]$, the variables $\langle \boldsymbol{x}, \boldsymbol{\xi}_k\rangle$ and s_h are independent centered Gaussian variables. With respect to the projected input space, the effective target function is

$$f_\star^{\mathrm{eff}}(\boldsymbol{s}) := \mathbb{E}[f_\star(\boldsymbol{x})|\boldsymbol{s}] = \sum_{i=1}^{r} a_k^\star g_k(s_k). \tag{56}$$

Then, the readout training in Algorithm 1 corresponds to random feature ridge regression [50] with weights $\boldsymbol{Z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_p)^\top$ on the projected covariates $\boldsymbol{s}_i = \hat{\boldsymbol{W}}^{\mathrm{sp}}\boldsymbol{x}_i^{(2)}$, $i \in [n]$, where $\{\boldsymbol{x}_i^{(2)}\}_{i\in[n]}$ are the covariates in the dataset $\mathcal{D}_2$:[3]

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_j\sigma(\langle \boldsymbol{z}_j, \boldsymbol{s}_i\rangle)\right)^2 + \frac{\lambda}{2}\|\boldsymbol{a}\|^2 \tag{57}$$

Since σ is bounded and continuous, we can apply Theorem 1 in [53][4], choosing $p = \omega(n^{1/2})$, $\lambda = \Theta(n^{-1/2})$, so that we obtain that the risk eq. (6) $R(\hat{\boldsymbol{a}}, \boldsymbol{W})$ satisfies

$$R(\hat{\boldsymbol{a}}, \boldsymbol{W}) - R_{f_{\mathcal{H}}} = O(n^{-1/2}), \tag{58}$$

where, given the *reproducing kernel Hibert space* (RKHS) $\mathcal{H}$ associated to the kernel $K(\boldsymbol{s}, \boldsymbol{s}') = \mathbb{E}_{\boldsymbol{z},b}[\sigma(\boldsymbol{s}^\top\boldsymbol{s} + b)\sigma(\boldsymbol{w}^\top\boldsymbol{s}' + b)]$,

$$R_{f_{\mathcal{H}}} = \min_{f\in\mathcal{H}} \mathbb{E}_{y,\boldsymbol{s}}[(f(\boldsymbol{s}) - y)^2]. \tag{59}$$

Such irreducible risk is lower-bounded by

$$R_\star = \mathbb{E}[(f_\star^{\mathrm{eff}}(\boldsymbol{s}) - y)^2] = \Theta\left(\sum_{k=1}^{r}(a_k^\star)^2\eta_k^2 + \sum_{k=r+1}^{m_\star}(a_k^\star)^2\right) = \Theta(\mathrm{MMSE}_\gamma), \tag{60}$$

3. In order to simplify the notation, we denote the labels in $\mathcal{D}_2$ as $y_i := y_i^{(2)}$.
4. Note that the assumption of bounded outputs in Theorem 1 is relaxed in the Appendix of [53], see Assumption 4. Further, the analysis in [53] allows for a more refined error rate in our setting. However, the current result is already subleading with respect to the rates in Theorem 7.

where, as for the spectral method mean-squared error, the two terms correspond to the approximation error of learned features and underfitting of unlearned ones. If $f_\star^{\rm eff} \in \mathcal{H}$, the argument is complete. By assumption, the function σ allows for a decomposition in Hermite polynomials

$$\sigma(z) = \sum_{\beta\geq 0} \frac{\sigma_\beta}{\beta!}\mathrm{He}_\beta(z), \quad \sigma_\beta := \frac{1}{\beta!}\mathbb{E}_z[\mathrm{He}_\beta(z)\sigma(z)]. \tag{61}$$

Further, being bounded, it cannot be a polynomial of finite degree, as there is no β_0 such that $\sigma_\beta = 0$ for all $\beta \geq \beta_0$, excluding the trivial case of constant functions. Considering the Taylor expansion of Hermite polynomials

$$\mathrm{He}_\beta(z+b) = \sum_{\zeta=0}^{\beta} \binom{\beta}{\zeta} b^{\beta-\zeta}\mathrm{He}_\zeta(z), \tag{62}$$

with $\binom{\beta}{\zeta} = \frac{\beta!}{\zeta!(\beta-\zeta)!}$, we find that the kernel K is given by

$$K(\boldsymbol{s},\boldsymbol{s}') = \mathbb{E}_{\boldsymbol{z},b}[\sigma(\boldsymbol{z}^\top\boldsymbol{s}+b)\sigma(\boldsymbol{z}^\top\boldsymbol{s}'+b)] \tag{63}$$

$$= \mathbb{E}_{\boldsymbol{z},b}\sum_{\beta,\beta'\geq 0}\sum_{\zeta=0}^{\beta}\sum_{\zeta'=0}^{\beta'} \frac{\sigma_\beta\sigma_{\beta'}}{\zeta!\zeta'!(\beta-\zeta)!(\beta'-\zeta')!} b^{\beta-\zeta}b^{\beta'-\zeta'}\mathrm{He}_\zeta(\langle\boldsymbol{z},\boldsymbol{s}\rangle)\mathrm{He}_{\zeta'}(\langle\boldsymbol{z},\boldsymbol{s}'\rangle) \tag{64}$$

By Proposition 11.31 in [48],

$$K(\boldsymbol{s},\boldsymbol{s}') = \mathbb{E}_b\sum_{\beta,\beta'\geq 0}\sum_{\zeta=0}^{\min(\beta,\beta')} \frac{\sigma_\beta\sigma_{\beta'}}{\zeta!(\beta-\zeta)!(\beta'-\zeta)!} b^{\beta+\beta'-2\zeta}\frac{\langle\boldsymbol{s},\boldsymbol{s}'\rangle^\zeta}{r^\zeta} \tag{65}$$

$$= \mathbb{E}_b\sum_{\zeta=0}^{\infty}\sum_{\beta,\beta'\geq\zeta} \frac{\sigma_\beta\sigma_{\beta'}}{\zeta!(\beta-\zeta)!(\beta'-\zeta)!} b^{\beta+\beta'-2\zeta}\frac{\langle\boldsymbol{s},\boldsymbol{s}'\rangle^\zeta}{r^\zeta} \tag{66}$$

$$= .\sum_{\zeta=0}^{\infty}\underbrace{\mathbb{E}_b\left(\sum_{\beta\geq\zeta}\frac{\sigma_\beta}{\zeta!(\beta-\zeta)!}b^{\beta-\zeta}\right)^2}\frac{\langle\boldsymbol{s},\boldsymbol{s}'\rangle^\zeta}{r^\zeta} \tag{67}$$

The bracketed term is always strictly positive as $\sum_{\beta\geq\zeta}\frac{\sigma_\beta}{\zeta!(\beta-\zeta)!}b^{\beta-\zeta}$ in b cannot be identically zero. This would require $\sigma_\beta = 0$ for all $\beta > \zeta$, which contradicts the hypothesis of boundedness. Expanding the term $\langle\boldsymbol{s},\boldsymbol{s}'\rangle$, it follows that the kernel admits a decomposition as $K(\boldsymbol{s},\boldsymbol{s}') = \Phi(\boldsymbol{s})^\top\Phi(\boldsymbol{s}')$, where Φ is a feature map with components given by multivariate monomials $\lambda_{\boldsymbol{\beta}}s_1^{\beta_1}\dots s_r^{\beta_r}$, for some $\lambda_{\boldsymbol{\beta}} > 0$ for all degrees $|\boldsymbol{\beta}|$. Theorem 4.21 in [19] readily implies that the RKHS of K includes all finite degree polynomials, completing our argument.

7. Useful results on single-index models

In this section we present various results applied in the derivation of the lower bound to MMSE_γ in Appendix 5.1. In particular, we focus on the following single-index model setting.

Definition 16 *Let $\boldsymbol{w}^\star \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$, $g:\mathbb{R}\to\mathbb{R}$ satisfy Assumption 12. Consider the supervised learning problem of estimating $\boldsymbol{w}_\star$ from n i.i.d. observations $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d\times 1} : i \in [n]\}$ generated as*

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \qquad y_i = \sqrt{\lambda}g(\langle\boldsymbol{x}_i, \boldsymbol{w}^\star\rangle) + \xi_i, \tag{68}$$

where $\lambda \geq 0$ is the signal-to-noise ratio (SNR) and $\xi_i \sim \mathcal{N}(0,1)$ is additive noise.

Given an estimator $\hat{\boldsymbol{w}}$ of $\boldsymbol{w}^\star$ that is a function of the dataset, we evaluate its estimation error using the following matrix-MMSE:

$$\mathrm{mse}(\boldsymbol{w}) := \frac{1}{d^2}\mathbb{E}[\|\boldsymbol{w}\boldsymbol{w}^\top - \boldsymbol{w}^\star\boldsymbol{w}^{\star\top}\|_F^2], \tag{69}$$

where $\mathbb{E}$ computes the expected value with respect to the joint distributions of $\mathcal{D}$ and $\boldsymbol{w}_\star$. A lower bound to this quantity is given by the following optimal matrix-MSE

$$\mathrm{mmse} := \frac{1}{d^2}\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{w}\boldsymbol{w}|\mathcal{D}] - \boldsymbol{w}^\star\boldsymbol{w}^{\star\top}\|_F^2\right] = \underset{\boldsymbol{Q}\in\mathbb{R}^{\mathrm{d}\times d}}{\arg\min}\,\mathbb{E}\left[\|\boldsymbol{Q} - \boldsymbol{w}^\star\boldsymbol{w}^{\star\top}\|_F^2\right]. \tag{70}$$

Intuitively, such optimal estimation error decreases with the SNR.

Lemma 17 *In the setting defined in 16, the optimal matrix-MSE satisfies*

$$\frac{\partial}{\partial\lambda}\mathrm{mmse} < 0. \tag{71}$$

Proof The proof follows from direct computation. We provide here a condensed derivation.

$$\mathrm{mmse} = \mathbb{E}\left[\left\|\boldsymbol{w}\boldsymbol{w}^T\right\|_F^2\right] - \mathbb{E}_{\boldsymbol{y}}\left[\left\|\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T \mid \boldsymbol{y}]\right\|_F^2\right]. \tag{72}$$

Since the first term is independent of λ, it suffices to show that the second term is non-decreasing. Let us denote the posterior mean as $\langle\boldsymbol{w}\boldsymbol{w}^T\rangle_\lambda \equiv \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T \mid \boldsymbol{y}]$. We compute its derivative with respect to λ:

$$\frac{d}{d\lambda}\mathbb{E}_{\boldsymbol{y}}\left[\|\langle\boldsymbol{w}\boldsymbol{w}^T\rangle_\lambda\|_F^2\right] = \int d\boldsymbol{y}\frac{\partial}{\partial\lambda}\left(\mathsf{P}(\boldsymbol{y})\|\langle\boldsymbol{w}\boldsymbol{w}^T\rangle_\lambda\|_F^2\right). \tag{73}$$

Using the explicit form of $\mathsf{P}(\boldsymbol{y}|\boldsymbol{w}) \propto \exp(-\frac{1}{2}\|\boldsymbol{y} - \sqrt{\lambda}g(\boldsymbol{X}\boldsymbol{w})\|^2)$, the derivative $\partial_\lambda \ln \mathsf{P}(\boldsymbol{y})$ introduces terms involving $\boldsymbol{y}^T\langle g(\boldsymbol{X}\boldsymbol{w})\rangle_\lambda$. We handle these terms via the multivariate Stein's Lemma (corresponding to integration by parts), which states that for the Gaussian measure, $\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}^T\mathbf{f}(\boldsymbol{y})] = \mathbb{E}_{\boldsymbol{y}}[\mathrm{Tr}(\nabla\mathbf{f}) + \sqrt{\lambda}\langle h(\boldsymbol{X}\boldsymbol{w})\rangle_\lambda^T\mathbf{f}]$.

Applying this identity results in cancellations of the lower-order terms. The surviving term is proportional to the gradient of the estimator with respect to the observations $\nabla_{\boldsymbol{y}}\langle\boldsymbol{w}\boldsymbol{w}^T\rangle_\lambda = \sqrt{\lambda}\,\mathrm{Cov}(\boldsymbol{w}\boldsymbol{w}^T, g(\boldsymbol{X}\boldsymbol{w}) \mid \boldsymbol{y})$. Therefore,

$$\frac{d}{d\lambda}\mathbb{E}_{\boldsymbol{y}}\left[\|\langle\boldsymbol{w}\boldsymbol{w}^T\rangle_\lambda\|_F^2\right] = \mathbb{E}_{\boldsymbol{y}}\left[\left\|\mathrm{Cov}\left(\boldsymbol{w}\boldsymbol{w}^T, g(\boldsymbol{X}\boldsymbol{w}) \mid \boldsymbol{y}\right)\right\|_F^2\right] \implies \frac{\partial}{\partial\lambda}\mathrm{mmse} < 0. \tag{74}$$

■

We now consider the high-dimensional limit $n, d \to \infty$ with fixed ratio $\alpha = n/d$, referred to as the sample complexity. The following theorem specializes Theorems 1 and 2 of [8] to the setting of interest.

Theorem 18 ([8]) *Consider the setting defined in 16. Then, in the limit $n, d \to \infty$, with fixed ratio $n/d = \alpha$,*

$$\frac{1}{d^2}\mathbb{E}\|\boldsymbol{w}^\star\boldsymbol{w}^{\star\top} - \mathbb{E}[\boldsymbol{w}\boldsymbol{w}\top \mid \mathcal{D}]\|_F^2 \to 1 - m^2, \tag{75}$$

and, given $\boldsymbol{w} \sim \mathsf{P}(\cdot|\mathcal{D})$,

$$\frac{1}{d}|\boldsymbol{w}^\top \boldsymbol{w}^\star| \xrightarrow{\mathbb{P}} m, \tag{76}$$

where $m = m(\alpha)$ is the maximizer of

$$\sup_{m\in[0,1]} f_{\mathrm{RS}}(m), \quad f_{\mathrm{RS}}(m) := \{m + \log(1-m) + 2\alpha\Psi_{\mathrm{out}}(m)\}, \tag{77}$$

$$\Psi_{\mathrm{out}}(m) := \mathbb{E}_{W,V,Y} \log \mathbb{E}_{w\sim\mathcal{N}(0,1)} \left[\mathsf{P}(Y|\sqrt{m}V + \sqrt{1-m}w)\right], \tag{78}$$

with $V, W \sim \mathcal{N}(0,1)$, $Y \sim \mathsf{P}(\cdot|\sqrt{m}V + \sqrt{1-m}W)$.

As a direct consequence of the theorem, the information-theoretic weak recovery threshold α^{IT} is the smallest sample complexity α such that the maximizer of the free entropy eq. (77) is $m \neq 0$. Equivalently, α^{IT} is the smallest α such that $\mathrm{mmse} < 1$. Lemma 17 readily implies that $\alpha^{\mathrm{IT}} = \alpha^{\mathrm{IT}}(\lambda)$ is decreasing with the SNR.
Finally, we characterize the IT weak recovery threshold in the limit of small SNR. By expanding eq. (78) around $\lambda = 0$, we derive the following Corollary.

Corollary 19 (IT weak recovery threshold in the large noise regime) *Consider the setting of Theorem 18. Then, in the limit $\lambda \to 0$, the information theoretic weak-recovery threshold satisfies*

$$\alpha^{\mathrm{IT}} = \Theta(\lambda^{-1}) \tag{79}$$

and, for $\alpha \to \infty$, and fixed λ the optimal matrix-MSE scales as

$$\mathrm{mmse} = O\left(\frac{1}{\lambda\alpha}\right). \tag{80}$$

Proof In order to simplify the notation, we introduce the shorthand $\phi_\beta(V; m) = \mathbb{E}_{\boldsymbol{w}}[g^\beta(\sqrt{m}V + \sqrt{1-m}w)]$.

$$\mathbb{E}_w\left[\mathsf{P}(Y = y|\sqrt{m}V + \sqrt{1-m}w)\right] = \frac{1}{\sqrt{2\pi}}\mathbb{E}_w\left[\exp\left(-\frac{(y - \sqrt{\lambda}(\sqrt{m}V + \sqrt{1-m}w))^2}{2}\right)\right] \tag{81}$$

$$= \frac{e^{-y^2/2}}{\sqrt{2\pi}}\left(1 + \sum_{\beta\geq 1}\frac{\lambda^{\beta/2}}{\beta!}\phi_\beta(V;m)\mathrm{He}_\beta(y)\right) \tag{82}$$

and therefore, expanding $\mathsf{P}(y|\sqrt{m}V+\sqrt{1-m}W)$ in a similar fashion and leveraging the expansion $\log(1+x) = -\sum_{k\geq 0}(-1)^k x^k/k$, up to constant terms with respect to m,

$$\Psi_{\text{out}}(m) = \mathbb{E}_{V,W,Y}\left[-\frac{y^2}{2} + \sum_{\beta\geq 1}\frac{\lambda^{\beta/2}}{\beta!}\phi_\beta(V;m)\text{He}_\beta(y) - \frac{1}{2}\left(\sum_{\beta\geq 1}\frac{\lambda^{\beta/2}}{\beta!}\phi_\beta(V;m)\text{He}_\beta(y)\right)^2\right] + O(\lambda^{3/2}) \tag{83}$$

$$= \mathbb{E}_V\int\frac{e^{-y^2/2}}{\sqrt{2\pi}}\left(-\frac{y^2}{2} + \sum_{\beta\geq 1}\frac{\lambda^{\beta/2}}{\beta!}\phi_\beta(V;m)\text{He}_\beta(y) - \frac{\lambda}{2}\phi_1^2(V;m)y^2\right)\mathrm{d}y + \tag{84}$$

$$+ \mathbb{E}_V\int\frac{e^{-y^2/2}}{\sqrt{2\pi}}\left(\left(-\frac{y^2}{2} + \sum_{\beta\geq 1}\frac{\lambda^{\beta/2}}{\beta!}\phi_\beta(V;m)\text{He}_\beta(y)\right)\sum_{\beta\geq 1}\frac{\lambda^{\beta/2}}{\beta!}\phi_\beta(V;m)\text{He}_\beta(y)\right)\mathrm{d}y + O(\lambda^{3/2})$$

$$= -\frac{1}{2} + \frac{\lambda}{2}\mathbb{E}_V[\phi_1^2(V;m)] - \frac{\lambda}{2}\mathbb{E}_V[\phi_2(V;m)] + O(\lambda^{3/2}) \tag{85}$$

$$= \frac{1}{2}\left(-1 - \lambda\mathbb{E}_{z\sim\mathcal{N}(0,1)}[g^2(z)] + \lambda\mathbb{E}_{(z_1,z_2)\sim\mathcal{N}(\mathbf{0}_2,\boldsymbol{C})}[g(z_1)g(z_2)]\right) + O(\lambda^{3/2}) \tag{86}$$

with

$$\boldsymbol{C} = \begin{pmatrix} 1 & m \\ m & 1 \end{pmatrix}. \tag{87}$$

In the above we used

$$\mathbb{E}_V[\phi_2(V;m)] = \mathbb{E}_{W,V}[g^2(\sqrt{m}V + \sqrt{1-m}W)] = \mathbb{E}_z[g^2(z)], \tag{88}$$

$$\mathbb{E}_V[\phi_1^2(V;m)] = \mathbb{E}_V[\mathbb{E}_W[g(\sqrt{m}V + \sqrt{1-m}W)]\mathbb{E}_{W'}[g(\sqrt{m}V + \sqrt{1-m}W')]] \tag{89}$$

$$= \mathbb{E}_{(z_1,z_2)\sim\mathcal{N}(\mathbf{0}_2,\boldsymbol{C})}[g(z_1)g(z_2)]. \tag{90}$$

Assumption 12 ensures that g can be decomposed in the Hermite basis as

$$g(z) = \sum_{k\geq 0}\frac{c_k}{k!}\text{He}_k(z), \qquad c_k := \frac{1}{k!}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[g(z)\text{He}_k(z)\right]. \tag{91}$$

Leveraging Proposition 11.31 in [48],

$$\mathbb{E}_{(z_1,z_2)\sim\mathcal{N}(\mathbf{0}_2,\boldsymbol{C})}[g(z_1)g(z_2)] = \sum_{k\geq 0}\frac{c_k^2}{k!}m^k \in [0, \mathbb{E}_z[g^2(z)]], \tag{92}$$

which is a non-decreasing function of $m \in [0,1]$. Note that, since g has generative exponent 2, and $\mathbb{E}_z[g(z)] = 0$, we have that $c_0 = c_1 = 0$ necessarily. Moreover, by Assumption 12, $c_2 \neq 0$ and bounded. Therefore, we are interested in maximizing the quantity[5]

$$f_{\text{RS}}(m) := m + \log(1-m) + \alpha\lambda\sum_{k\geq 2}\frac{c_k^2}{k!}m^k + O(\alpha\lambda^{3/2}) \tag{93}$$

$$= \frac{1}{2}\left(\alpha\lambda c_2^2 - 1\right)m^2 + \sum_{k>2}\left(\frac{\alpha\lambda}{k!}c_k^2 - \frac{1}{k}\right)m^k + O(\alpha\lambda^{3/2}), \tag{94}$$

5. Note that we are neglecting constant terms with respect to m.

where we have expanded $\log(1-m)$. For $\alpha > \lambda^{-1}c_2^{-2}$, $m=0$ is a minimum of the free entropy, therefore $\alpha^{\mathrm{IT}} \le \lambda^{-1}c_2^{-2}$. Denote

$$D := \inf_{m\in(0,1]} \frac{-m-\log(1-m)}{\sum_k \frac{c_k^2}{k!}m^k}, \tag{95}$$

which is strictly positive and well-defined. Note that

$$\lim_{m\to 0^+} \frac{-m-\log(1-m)}{\sum_k \frac{c_k^2}{k!}m^k} = c_2^{-2} \implies \frac{1}{2\mathbb{E}_z[g^2(z)]} \le D \le c_2^{-2} = \frac{1}{2\mathbb{E}_z[g''(z)]}. \tag{96}$$

Then, for all $\alpha < D\lambda^{-1}$, $f(m\neq 0) < f(0) = 0$, *i.e.* $m=0$ is the global maximizer and

$$D \le \alpha^{\mathrm{IT}}\lambda \le c_2^{-2} \implies \alpha^{\mathrm{IT}} = \Theta(\lambda^{-1}), \tag{97}$$

For the second result, we first consider the setting $\lambda \to 0$ and $\alpha \to \infty$, with $\alpha \gg \lambda$. There exists a non-zero maximizer m, which satisfies[6]

$$\frac{\mathrm{d}}{\mathrm{d}m} f_{\mathrm{RS}}(m) \overset{!}{=} 0 \implies \frac{1}{1-m} = \alpha\lambda \sum_{k\ge 2} \frac{c_k^2}{k!} m^{k-2} \implies \sum_{k\ge 2} \frac{c_k^2}{k!}\left(m^{k-2} - m^{k-1}\right) = \frac{1}{\alpha\lambda}, \tag{98}$$

The equation is solved by $m = 1 - \Theta\left(\frac{1}{\alpha\lambda}\right)$. Theorem 18 implies that, in this regime, $\mathrm{mmse}(\lambda) = \Theta((\alpha\lambda)^{-1})$. Together Lemma 17, the result for arbitrary λ is proved. ■

6. In the following we retain the leading terms in the free entropy expansion.