# Optimal scaling laws in learning hierarchical multi-index models

**Leonardo Defilippis**
Département d'Informatique
École Normale Supérieure, PSL & CNRS

**Florent Krzakala**
Information, Learning and Physics Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)

**Bruno Loureiro**
Département d'Informatique
École Normale Supérieure, PSL & CNRS

**Antoine Maillard**
Département d'Informatique
École Normale Supérieure, PSL & CNRS
INRIA Paris

## 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 adapt 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* (Kaplan et al., 2020; Brown et al., 2020; Hoffmann et al., 2022) suggests that the performance of large models scales 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, a line of work on the training dynamics of neural networks increasingly suggests that feature learning is not a smooth process, but is associated to long *plateaus and abrupt transitions* in the risk, with features (or *concepts* in this context) appearing sequentially rather than all at once (Saxe et al., 2014; Wei et al., 2022; Schaeffer et al., 2023). 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 emerge (Martin and Mahoney, 2021; Wang et al., 2023; Thamm et al., 2024).

This paper provides an end-to-end answer to these questions in a mathematically tractable—yet exhibiting genuine 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 (Ren et al., 2025; Ben Arous et al., 2025;

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Defilippis et al., 2026c). Here, hierarchical denotes the fact that the directions are *ordered*, in the sense that their relative importance decays 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 (Mallat, 1999; Donoho, 2006).

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, showing that sequential feature emergence, spectral outliers, plateaus in the prediction error and scaling laws are consequences of a single mechanism, which we characterize from first principles. 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*) for a large class of hierarchical models, 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 (Raskutti et al., 2011), with previously conjectured results for shallow networks with quadratic activation (Defilippis et al., 2026c; Boncoraglio et al., 2026) —which we thus prove rigorously— and with recent analysis of SGD in random feature models (Ferbach et al., 2025). Our analysis shows that these scalings are in fact universal for a broad class of hierarchical multi-index targets, namely within the generative-exponent-two regime (provided they possess a non-zero second-order Hermite coefficient), well beyond the quadratic activation 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. This estimator can be either interpreted as a spectral method applied to the network, or as the small–learning-rate limit of gradient descent on a suitably regularized loss, via the Hessian of a modified objective. As a consequence, gradient-based training is shown to recover the signal subspace optimally, provided it is properly tuned. Moreover, 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 $\gamma$ is the exponent controlling the hierarchy— leading to a cascade of sharp phase transitions. This produces plateaus and abrupt drops in the prediction error, providing a theoretical account of the progressive emergence of features observed empirically in neural networks.

(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. Crucially, this two-stage procedure succeeds because the spectral subspace estimator is fully agnostic to the target function and does not require any prior knowledge of its structure or tuning to its decay profile. We prove that this two-stage mechanism—optimal representation discovery followed by efficient readout estimation—applies in the noisy and regularized setting, and to a wide class of multi-index targets.

(iv) **Impact of noise sensitivity on scaling laws.** We investigate how the noise sensitivity of the target affects the algorithmic recovery problem, altering the sample complexity detection thresholds and the error scaling laws, extending the analysis to the entire class of generative-exponent-two functions, including those with vanishing second-order Hermite coefficient. We find that these new rates are governed by the *Noise Sensitivity Exponent* (NSE) of the model, a simple property of the link functions, first introduced by Defilippis et al. (2026b) in the context of single-index models and committee machines. Moreover, since the NSE is not expected to affect information-theoretic recovery, our result points to possible statistical-to-computational gaps in the rates.

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 Wei et al. 2022; Schaeffer et al. 2023) 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 (Defilippis et al., 2026c). 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.

**Further related works —** Most theoretical work on neural scaling laws focuses on effectively linear models, such as kernel methods and random features, where the generalization behavior can be characterized through spectral properties of fixed representations (Caponnetto and De Vito, 2007; Spigler et al., 2020; Cui et al., 2021; 2023; Defilippis et al., 2024; Atanasov et al., 2024; Bahri et al., 2024; Maloney et al., 2022; Paquette et al., 2024; Kunstner and Bach, 2025; Bordelon et al., 2020). Exceptions have focused either on the joint-training of both layers for linear networks (Bordelon et al., 2024; 2025; Worschech and Rosenow, 2024) or non-linear settings with fixed-features (Wortsman and Loureiro, 2025; Worschech and Rosenow, 2024; Braun et al., 2025). 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 is a recent line of work analyzing two-layer networks with structured first-layer weights and hierarchical multi-index targets (Ren et al., 2025; Ben Arous et al., 2025; Defilippis et al., 2026c). In particular, Ren et al. (2025); Ben Arous et al. (2025) studied one-pass SGD in this setting, with Ben Arous et al. (2025) focusing on the case of a quadratic neural network architecture with an idealized, continuous time dynamics. A major difference with this work is that we analyze instead full-batch ERM with a fixed number of samples, rigorously characterizing the optimal sample scaling laws and showing that they can be achieved computationally by a gradient-descent like algorithm, thus providing a fundamental benchmark for sample complexity rates in these problems.

Complementary, Defilippis et al. (2026c) heuristically derived rates for ERM in quadratic neural networks trained on quadratic targets, building on earlier work by Erba et al. (2025) (see also Maillard et al. 2024; Xu et al. 2025; Martin et al. 2026). Our results show that these rates are universal within a large class of target functions (any generative exponent two functions in the language of Damian et al. 2024, with non-zero second Hermite coefficient), closing a gap between the purely quadratic setting and general hierarchical multi-index targets. The key underlying these universal scaling laws is the combination of the *quasi-sparsity* of the target representation, encoded by a heavy-tailed spectrum that induces an ordered hierarchy of feature strengths (Mallat, 1999; Donoho, 2006), and an implicit *rank-sparsity* bias, leading to LASSO-like behavior (Raskutti et al., 2011).

Our proof technique is of independent interest and builds on the toolbox of approximate message passing (AMP) and its associated state evolution (Bayati and Montanari, 2011; Donoho et al., 2013; Javanmard and Montanari, 2013; Berthier et al., 2020; Zou and Yang, 2022; Feng et al., 2022; Gerbelot and Berthier, 2023; Dudeja et al., 2023; Erba et al., 2025). Our results leverage the connection between AMP and Bayes-optimal estimation for single- (Barbier et al., 2019) and multi-index (Aubin et al., 2018; Troiani et al., 2025) models. Similarly, we build up on the literature on spectral methods derived from AMP for single- (Mondelli and Montanari, 2019; Lu and Li, 2020; Maillard et al., 2022; Damian et al., 2024) and multi-index (Kovačević et al., 2025; Defilippis et al., 2026a) functions.

## 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 2.1** (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 $\|\boldsymbol{w}_k^\star\|^2 = d$, namely $\boldsymbol{W}^\star = \sqrt{d}\boldsymbol{Q}^\star$, with $\boldsymbol{Q}^\star$ uniformly sampled from the Stiefel manifold $\mathrm{St}(m_\star, d)$. 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 $a_k^\star$. Here, the hierarchy refers only to the relative importance of the features, and not to a compositional structure of the target, as in Cagnetta et al. (2024); Dandi et al. (2025); Tabanelli et al. (2026).
- 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 (Li, 1991; Yuan, 2011; Babichev and Bach, 2018).
- 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 (Damian et al., 2022; Abbe et al., 2023; Dandi et al., 2024).
- Scale-free hierarchical multi-index models were considered recently in Ren et al. (2025); Ben Arous et al. (2025). Note that when $\gamma > 1/2$, the target is *quasi-sparse* in the index basis (Mallat, 1999; Donoho, 2006). 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 (Gunasekar et al., 2017; Soudry et al., 2018; Andriushchenko et al., 2023).
- In scale-free hierarchical multi-index models with $\gamma < 1/2$, the coefficients must scale as $a_k^\star \asymp m_\star^{\gamma - 1/2} k^{-\gamma}$ to guarantee the condition $\sum_k (a_k^\star)^2 = 1$.

We consider a supervised learning task with data generated from a hierarchical multi-index model.

**Definition 2.2.** *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. 2.1. More precisely*
$$y_i = f_\star(\boldsymbol{x}_i) + \sqrt{\Delta}\xi_i, \tag{3}$$
*where* $\boldsymbol{x}_i \sim \mathcal{N}(0, {}^1\!/_d \boldsymbol{I}_d)$, $\Delta > 0$ *is the noise variance and* $\xi_i \sim \mathcal{N}(0,1)$. *Furthermore, we consider the high-dimensional limit* $n, d \to \infty$ *with* $\alpha = n/d$ *and* $m_\star$ *fixed. Only after taking this limit do we consider the regime* $\alpha, m_\star \gg 1$ *in which the rates are expressed. In this sense,* $m_\star = \Theta_d(1)$ *means that* $m_\star$ *does not grow with* $d$.

Moreover, our results depend on the following additional assumption on the link functions $g_k$.

**Assumption 2.3.** *For each index* $k \in [m_\star]$, $g_k$ *is an even function. Moreover, for* $z \sim \mathcal{N}(0,1)$, *for some constants* $b, C, D > 0$, *independent of* $k, n, d, \alpha, m_\star$,
$$\begin{cases} |g_k(z)| \le C(1 + |z|^b), \\ D < \mathbb{E}_z[g_k^2(z)] < C, \quad D < |\mathbb{E}_z[g_k(z)(z^2 - 1)]| < C. \end{cases} \tag{4}$$

**Remark 2.4.** *The parity assumption on* $g_k$ *ensures that learning* $\mathrm{span}(\boldsymbol{W}_\star)$ *is* non-trivial *(Troiani et al., 2025). It rules out linear correlations that would yield weak recovery at arbitrarily small sample complexity, a regime where the emergence phenomenon cannot occur. 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*
$$\alpha_c := \left( \sup_{\boldsymbol{M} \in \mathbb{S}^+_{m_\star},\, \|\boldsymbol{M}\|_{\mathrm{F}} = 1} \|\mathbb{E}_y \boldsymbol{G}(y) \boldsymbol{M} \boldsymbol{G}(y)\|_{\mathrm{F}} \right)^{-1}, \tag{5}$$
*with* $\boldsymbol{G}(y) := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_{m_\star} \mid Y = y]$ *(Barbier et al., 2019; Mondelli and Montanari, 2019; Lu and Li, 2020; Damian et al., 2024; Troiani et al., 2025).*

*The bound of eq. (4) on* $\mathbb{E}_z[g_k^2(z)]$ *controls the label variance, while the lower bound on* $\mathbb{E}_z[g_k(z)(z^2 - 1)]$ *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* $a_k^\star$.

Hierarchical multi-index models with generative exponent equal to 2 include the vast majority of cases of interest, while the class of models with exponent larger than 2 mostly includes fine-tuned examples (Barbier et al., 2019; Damian et al., 2024). Specifically, Assumption 2.3 considers a subset of the targets with a generative exponent equal to two – namely, all link functions with a non-zero second-order coefficient in their Hermite expansion. According to the recent framework of Defilippis et al. (2026b), this corresponds to a *noise sensitivity exponent* equal to one. Within this class, we prove the universality of the rates obtained for quadratic models in Defilippis et al. (2026c). In Section 3.4 we relax this assumption to encompass all hierarchical models with a generative exponent of two, allowing us to explore how generic noise sensitivity exponents affect the error rates.

Our key goals are twofold: (i) to characterize what are the best achievable rates (a.k.a. scaling laws) in recovering concepts and features for this class of hierarchical tasks; (ii) to show that out-of-the box models such as neural networks can learn efficiently the features, and the entire functions as well, with those optimal scaling rates.

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

**Definition 2.5** (Matrix-MSE)**.** *Let $\boldsymbol{w}_k^\star \in \mathbb{S}^{d-1}(\sqrt{d})$ denote one of the target indices $k \in [m_\star]$. We define the matrix-MSE associated to a predictor $\boldsymbol{w} \in \mathbb{R}^d$ as:*

$$\mathrm{mse}_k(\boldsymbol{w}) := \frac{1}{d^2}\mathbb{E}\left[\|\boldsymbol{w}\boldsymbol{w}^\top - \boldsymbol{w}_k^\star \boldsymbol{w}_k^{\star\top}\|_{\mathrm{F}}^2\right]. \tag{6}$$

*In our results, we consider this metric in the limit $d \to \infty$, which we assume to exist.*

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 2.6** ($k-$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 $\boldsymbol{w}_k^\star$, namely*

$$\alpha_k := \inf\left\{\alpha > 0 : \mathbb{E}\left[\frac{|\langle \hat{\boldsymbol{w}}_k, \boldsymbol{w}_{\star,k}\rangle|}{\|\hat{\boldsymbol{w}}_k\|\|\boldsymbol{w}_{\star,k}\|}\right] = \Theta_d(1)\right\}. \tag{7}$$

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 2.7** (Weighted MSE)**.** *Let $\boldsymbol{W} \in \mathbb{R}^{m_\star \times d}$, with rows $\boldsymbol{w}_1, \dots, \boldsymbol{w}_{m_\star}$. We define the weighted mean-squared error as*

$$\mathrm{MSE}_\gamma(\boldsymbol{W}) := \min_{\pi \in \mathfrak{S}_{m_\star}} \sum_{k=1}^{m_\star} (a_k^\star)^2 \mathrm{mse}_k(\boldsymbol{w}_{\pi(k)}), \tag{8}$$

*where $\mathfrak{S}_{m_\star}$ is the set of permutations of $[m_\star]$.*

The minimum over permutations accounts for the fact that multi-index models are identifiable only up to a relabeling of the indices: an estimator is not required to return the directions in the order of the hierarchy. For the second goal — showing that neural networks can efficiently learn $f_\star$ — we will focus on two-layer neural networks:

$$f(\boldsymbol{x}; \boldsymbol{\Theta}) := \frac{1}{\sqrt{p}}\boldsymbol{a}^\top \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \tag{9}$$

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. 2.1. We quantify the generalization capacity of this model through its excess risk

$$R(\boldsymbol{\Theta}) := \mathbb{E}[(f_\star(\boldsymbol{x}) - f(\boldsymbol{x}; \boldsymbol{\Theta}))^2]. \tag{10}$$

## 3 Main results

We now discuss our main contributions. First, we derive the information-theoretical optimal reconstruction error for the class of scale-free hierarchical multi-index models introduced in Def. 2.2

under Asm. 2.3, as quantified by the weighted reconstruction error in Eq.(8). This result will serve as a fundamental benchmark on the difficulty of subspace recovery for a given $\gamma > 0$. Second, we construct a spectral estimator for subspace recovery, and prove that this algorithm achieves the aforementioned optimal reconstruction error. Then, we consider two-layer neural networks, proving that a two-step training procedure also achieves the optimal scaling laws for the excess risk. Finally, relaxing Asm. 2.3, we derive the optimal algorithmic rates for a broader class of targets, highlighting their connection to the function's noise sensitivity exponent.

### 3.1 Bayes-optimal rates for feature recovery

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 $\alpha$ and the subspace dimension $m_\star$.

**Definition 3.1** (Optimal MSE)**.** *Let $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ be drawn as in def. 2.2. Then, the optimal (weighted) mean-square error is achieved by the posterior average:*

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

*In our results, we consider this metric in the limit $d \to \infty$, which we assume to exist.*

By definition, $\mathrm{MMSE}_\gamma$ 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 $\mathrm{MMSE}_\gamma$ 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. We then have the following theorem (proven in Appendix A).

**Theorem 3.2** (Optimal scaling laws)**.** *In the setting of Definitions 2.1, 2.2, under Assumption 2.3, for $\alpha, m_\star \gg 1$, the Bayes-optimal mean-squared error satisfies*

$$\mathrm{MMSE}_\gamma = \Theta_{\alpha, m_\star}\begin{cases} \alpha^{-1+1/2\gamma}, & \gamma > 1/2,\ \alpha \ll m_\star^{2\gamma}, \\ m_\star/\alpha, & \gamma > 1/2,\ m_\star^{2\gamma} \ll \alpha, \\ 1, & \gamma < 1/2,\ \alpha \ll m_\star, \\ m_\star/\alpha, & \gamma < 1/2,\ m_\star \ll \alpha. \end{cases} \tag{11}$$

*Moreover, the $k$-critical threshold of the Bayes estimator satisfies (with $x_+ := \max(0, x)$).*

$$\alpha_k^{\mathrm{Bayes}} = \Theta_k(k^{2\gamma} m_\star^{(1-2\gamma)_+}), \tag{12}$$

### 3.2 Optimal agnostic subspace recovery

While Theorem 3.2 provides a fundamental benchmark, an equally important question is whether the optimal reconstruction rates can be *efficiently* achieved by an algorithm which is agnostic to 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 3.3** (Spectral estimator)**.** *Given a pre-processing function $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ on the labels, consider the symmetric random matrix*

$$\boldsymbol{T} = \sum_{i=1}^{n} \mathcal{T}(y_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top, \tag{13}$$

*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 $\hat{\boldsymbol{W}}^{\mathrm{sp}} = (\hat{\boldsymbol{w}}_1, ..., \hat{\boldsymbol{w}}_r)^\top \in \mathbb{R}^{r \times d}$, where $\hat{\boldsymbol{w}}_k$ is the $k \in [r]$ eigenvector of $\boldsymbol{T}$ corresponding to $\lambda_k$, normalized such that $\|\hat{\boldsymbol{w}}_i\|^2 = d$.*

Such estimators have been studied in the context of multi-index models (Kovačević et al., 2025; Defilippis et al., 2026a) and proven to achieve the computationally optimal weak recovery threshold (Troiani et al., 2025), for a tailored choice of the pre-processing function $\mathcal{T}$. Instead, here we consider $\mathcal{T}$ to be data-agnostic, satisfying only the following mild conditions.

**Assumption 3.4.** *Given $Y \sim \mathcal{N}\left(\sum_{k=1}^{m_\star} a_k^\star g_k(z_k), \Delta\right)$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{m_\star})$, we assume that $\mathcal{T}$ is bounded, strictly increasing and $\inf_y \mathcal{T}(y) < 0 < \sup_y \mathcal{T}(y) =: \tau$.*

Under Assumption 3.4, Kovačević et al. (2025) show that, in the proportional high-dimensional limit $n, d \to \infty$ with $n = \Theta(d)$, the empirical spectrum of $\boldsymbol{T}$ in eq. (13) converges to a bounded continuous density — referred to as the *bulk* — plus $r \leq m_\star$ isolated eigenvalues — referred to as the *spikes* — with $r$ depending on the sample complexity $\alpha = {}^n\!/_d$. This phenomenon is reminiscent of the classical Baik–Ben Arous–Péché (BBP) transition in random matrix theory (Baik et al., 2005).

The spikes are separated from the bulk by a finite gap and their corresponding eigenvectors exhibits a non-trivial correlation with $\mathrm{span}(\boldsymbol{W}_\star)$. Therefore, the criterion described in Definition 3.3 (reminiscent of the heuristic "elbow" method in principal components analysis, see Fig. 3) for selecting the number of eigenvalues agnostically retains only the informative components, discarding the noise contributions associated with the bulk. Note also that the choice of cutoff $C/\sqrt{d}$ is arbitrary, as in general we want to separate the bulk with spacing $o_d(1)$ and the spikes with spacing $\Theta_d(1)$.

**Remark 3.5** (GD interpretation of 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 (Bonnaire et al., 2025). The early dynamics of GD on the first layer of (9), for weights initialized on a sphere with small radius, will thus follow the proposed spectral method. This construction was recently used in Zhang et al. (2025) 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 gradient-based algorithm, reads as a power iteration*

$$\boldsymbol{w}_k^{t+1} = \boldsymbol{w}_k^t - \nabla_{\boldsymbol{w}_k} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \boldsymbol{a}^\top \sigma(\boldsymbol{W}^t \boldsymbol{x}_i)/\sqrt{p}\right) \approx (\boldsymbol{I}_d - a_k \sigma''(0) \boldsymbol{T}/\sqrt{p}) \boldsymbol{w}_k^t, \tag{14}$$

*where $\boldsymbol{T}$ has the structure defined in eq. (13) with pre-processing $\mathcal{T} = \ell'(\,\cdot\,, 0)$, and $\ell'$ denotes the derivative of $\ell$ with respect to its second argument.*

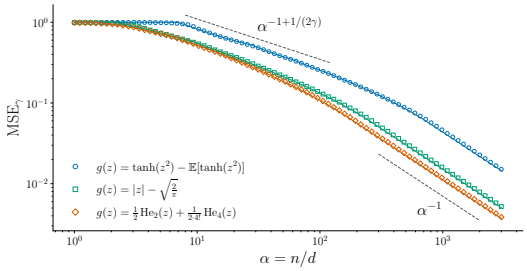



Figure 1: Weighted mean square error $\mathrm{MSE}_\gamma$ – see Def. 2.7 – of the spectral estimator of Def. 3.3 with preprocessing function $\mathcal{T}(y) = y/(1+|y|)$. The target is given by the hierarchical multi-index model 2.1, with $g_k(z) = g(z)\ \forall k$, stated in the legend, and $a_k^\star \propto k^{-\gamma}$, $\gamma = 1$. Solid lines: asymptotic prediction for $d \to \infty$, from Theorem A.2 and the overlaps derived in App. A.2. Dots represent averages over 30 instances, for covariates with dimension $d = 2000$, feature space dimension $m_\star = 10$ and label noise variance $\Delta = 0.5$.

Our second main theoretical result shows that the data-agnostic spectral method defined above indeed achieves the optimal reconstruction rates of Theorem 3.2.

**Theorem 3.6.** *In the setting of Definitions 2.1, 2.2, under Assumption 2.3, denote $\hat{\boldsymbol{W}} = (\hat{\boldsymbol{W}}^{\mathrm{sp}}, \boldsymbol{0}_{(m_\star - r)\times d}) \in \mathbb{R}^{m_\star \times d}$, where $\hat{\boldsymbol{W}}^{\mathrm{sp}} \in \mathbb{R}^{r\times d}$, $r \leq m_\star$, is the spectral estimator defined in 3.3, with $\mathcal{T}$ satisfying Assumption 3.4. Then, for $g_k$ strictly increasing on $(0,\infty)$, and $\alpha, m_\star \gg 1$,*

$$\mathrm{MSE}_\gamma(\hat{\boldsymbol{W}}) = \Theta_{\alpha, m_\star}(\mathrm{MMSE}_\gamma), \tag{15}$$

*and the spectral estimator's $k$-critical threshold $\alpha_k^{\mathrm{sp}}$ satisfies $\alpha_k^{\mathrm{sp}} = \Theta_k(\alpha_k^{\mathrm{Bayes}})$.*

**Remark 3.7.** *Note that, as the spectral method 3.3 retrieves $r \leq m_\star$ directions, we construct the full estimator by filling the remaining columns with zeros to evaluate $\mathrm{MSE}_\gamma$ in Def. 2.7. This zero-padding is harmless: the added zeros do not impact the derived scaling laws or the estimator's agnostic nature. Moreover, the spectral thresholds depend on the link functions $g_k$, so that the spikes do not necessarily emerge in the order of the hierarchy, and the $k^{\mathrm{th}}$ eigenvector of $\boldsymbol{T}$ can be aligned with $\boldsymbol{w}_h^\star$, $h \neq k$. This is accounted for by the minimum over permutations in Definition 2.7, and $\alpha_k^{\mathrm{sp}}$ denotes the threshold at which an eigenvector aligned with $\boldsymbol{w}_k^\star$ emerges.*

**Remark 3.8.** *The additional technical assumption of $g_k$ strictly increasing on $(0,\infty)$ ensures that the informative outliers emerge from the right edge of the bulk. If we retain the unrestricted class of admissible link functions, the theorem continues to hold after a minor extension of the proof that accounts for outliers emerging from both edges of the bulk. This extension does not change any of the stated thresholds, rates, or scaling exponents.*

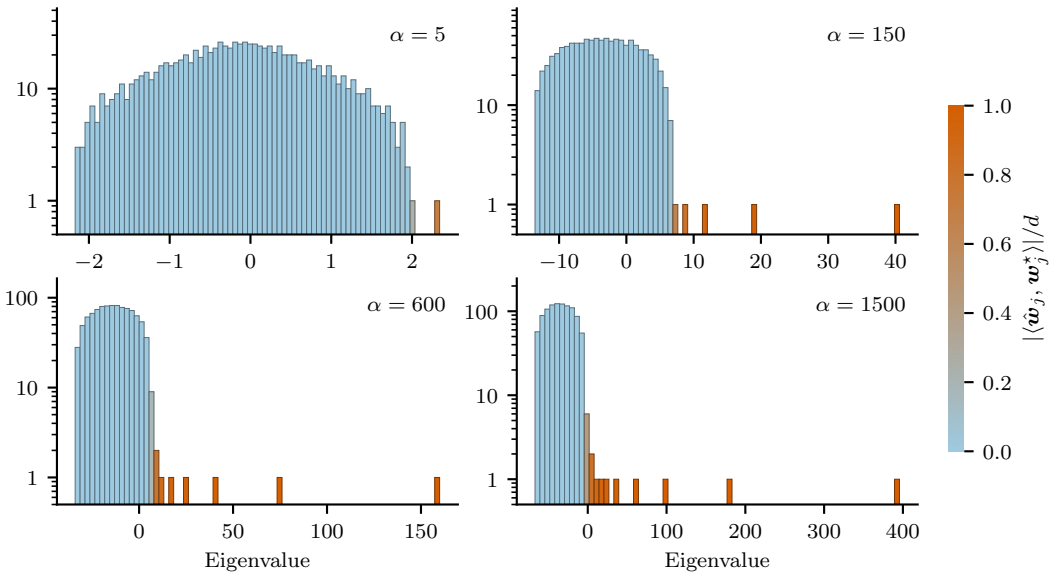



Figure 2: Empirical spectrum distribution of the matrix $\boldsymbol{T}$ defined in eq. (13), with preprocessing function $\mathcal{T}(y) = y/(1+|y|)$, for a single realization at increasing sample complexity $\alpha$ (indicated in each panel), highlighting the sequential emergence of concepts (spikes) from the bulk. Each bin is colored by the overlap $|\langle \hat{\boldsymbol{w}}_j, \boldsymbol{w}_j^\star \rangle|/d$ between the eigenvector $\hat{\boldsymbol{w}}_j$ of the $j^{\text{th}}$ largest eigenvalue, normalized as in Def. 3.3, and the target direction $\boldsymbol{w}_j^\star$ (averaged over the eigenvalues in the bin, and set to zero for $j > m_\star$). The target is given by a hierarchical multi-index model 2.1, with $g_k(z) = \frac{1}{2}\mathrm{He}_2(z) + \frac{1}{2\cdot 4!}\mathrm{He}_4(z)\ \forall k$, $a_k^\star \propto k^{-\gamma}$ with $\gamma = 1.3$ and $\sum_k (a_k^\star)^2 = 1$, and $\Delta = 0.5$. The covariates dimension is $d = 1000$, the feature space dimension is $m_\star = 20$.

In Appendix A.2, we show that the rates in Theorem 3.6 allow for an interpretation reminiscent of the "universal error decomposition" in Defilippis et al. (2026c). Let $k_\alpha \asymp \min(\alpha^{1/(2\gamma)}, m_\star)$ denote the number of weakly learnable features at sample complexity $\alpha$ for $\gamma > 1/2$ – corresponding to the number of spikes of $\boldsymbol{T}$ eq. (13). Then, the spectral mean-squared error decomposes into two main contributions: (i) the approximation error of the learnable concepts, scaling as $k_\alpha \cdot \Theta(\alpha^{-1})$; (ii) the underfitting due to the unlearnable concepts, scaling as $\Theta\left(\sum_{k=k_\alpha+1}^{m_\star} (a_k^\star)^2\right)$.

Figure 1 shows the decay of the weighted MSE for three scale-free hierarchical model variants as a function of the sample complexity $\alpha$, based on finite-size experiments, using the spectral method introduced in Definition 3.3. Notably, around $\alpha \asymp m_\star^{2\gamma} \approx 100$, we observe a crossover between the rates $\alpha^{-1+1/(2\gamma)}$ and $m_\star/\alpha$, as predicted by Theorem 3.6. Figure 2 illustrates the sequential emergence of new concepts as the sample complexity grows, visualized as spikes detaching from the eigenvalue bulk of the matrix $\boldsymbol{T}$ in eq. (13).

### 3.3 Learning hierarchical multi-index models with neural networks

Finally, we turn our attention to the problem of learning hierarchical multi-index target functions with a two-layer neural network. In particular, we are interested in characterizing the scaling laws associated with the excess risk in Equation (10) (rather than the reconstruction error in Equation (8)). Surprisingly, our main result in this section will show that the excess risk associated with a suitable training procedure described in Alg. 1, achieves the optimal rates in Theorem 3.2 (proof in App.B):

**Theorem 3.9.** *Under the assumptions of Theorem 3.6, let each $g_k$, $k \in [m_\star]$, be weakly differentiable with $\mathbb{E}_z[g_k'(z)^2] \le C'$, for a constant $C'$ independent of $k$, and let $\sigma$ be bounded, continuous and nonconstant. For every $\alpha$, there exist a width $p(\alpha)$ and a regularization $\lambda(\alpha) > 0$ such that, for $p \ge p(\alpha)$ and $0 < \lambda \le \lambda(\alpha)$, the excess risk $R_{\text{NN}}$ eq. (10), where the expectation is over all sources of randomness, of the two-layer neural network eq. (9) trained according to Algorithm 1 satisfies*

$$R_{\text{NN}} = \Theta(\text{MMSE}_\gamma). \tag{16}$$

**Algorithm 1** Spectral Initialization and Ridge Training

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

*1. Data Splitting*
Partition the dataset $\mathcal{D}$ into two disjoint sets $\mathcal{D}_1$ and $\mathcal{D}_2$ such that $|\mathcal{D}_1| = |\mathcal{D}_2| = \frac{n}{2}$.

*2. Feature Learning (Spectral Initialization on $\mathcal{D}_1$)*
Compute the spectral estimator $\hat{\boldsymbol{W}}^{\text{sp}} \in \mathbb{R}^{r\times d}$ of Def. 3.3 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: $\boldsymbol{W} = \boldsymbol{Z}\hat{\boldsymbol{W}}^{\text{sp}}$.

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

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

**Remark 3.10.** *This final result demonstrates that neural networks can efficiently learn the target function, achieving the exact same rates as those of the optimal feature subspace recovery. The proof effectively reduces the problem to random features regression on the $r$-dimensional projected inputs, whose dimension stays finite as $n, d \to \infty$. 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. (10). 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 3.2. We leave such a precise analysis of the optimal risk as a future direction.*

### 3.4 Effects of a generic noise sensitivity exponent

The results presented in Sections 3.1, 3.2, 3.3 rely on the condition $\mathbb{E}[g_k(z)(z^2-1)] \neq 0$ – or, equivalently, a Noise Sensitivity Exponent (NSE) equal to one (Defilippis et al., 2026b), as implied by Assumption 2.3. This regime ensures easier features detectability and leads to matching bounds between information-theoretic and algorithmic rates. For generic link functions $g : \mathbb{R} \to \mathbb{R}$, the NSE is defined as follows

**Definition 3.11** (Noise Sensitivity Exponent (Defilippis et al., 2026b))**.** *Let $g : \mathbb{R} \to \mathbb{R}$ such that $|g(z)| \leq C(1 + |z|^k)$ for some $C, k > 0$. The noise-sensitivity exponent is defined as*

$$\beta_\star(g) \coloneqq \min\{\beta \in \mathbb{N} : \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[g^\beta(z)(z^2-1)\right] \neq 0\}. \tag{17}$$

While typical functions have NSE $= 1$, NSE $= 2$ captures the majority of functions with zero second Hermite coefficient, including, for instance, higher-order Hermite polynomials $g(z) = He_{2k}(z)$ with $k > 1$. We refer to Defilippis et al. (2026b) for more examples.

Introduced in the context of single-index models and committee machines, the NSE can be shown to govern the optimal computational weak-recovery threshold of features in large noise regimes. In this section, we provide an example showing how a generic NSE can affect the optimal algorithmic weak-recovery.

**Assumption 3.12.** *For all $k \in [m_\star]$, $g_k = g$ is an even function. Moreover, for $z \sim \mathcal{N}(0,1)$, for some constants $b, C, D > 0$, independent of $k, n, d, \alpha, m_\star$, it satisfies $|g(z)| \leq C(1 + |z|^b)$, $D < \mathbb{E}_z[g^2(z)] < C$, and the noise sensitivity exponent of $g$ is given by $\beta_\star \in \mathbb{N}$.*

In agreement with Aubin et al. (2018); Troiani et al. (2025); Defilippis et al. (2026b), we assess the performance of the Bayes-AMP estimator $\hat{\boldsymbol{W}}_{\mathrm{AMP}}$, as the optimal estimator among first-order iterative methods (Celentano et al., 2020), *i.e.* the minimizer of the mean-squared error eq. 8.

**Theorem 3.13.** *In the setting of Definitions 2.1, 2.2, under Assumption 3.12 and* $1 \ll \alpha \ll m_\star^{2\gamma\beta_\star}$, $\gamma > 1/2$, *the Bayes-AMP estimator* $\hat{\mathbf{W}}^{\mathrm{AMP}}$ *(Aubin et al., 2018; Troiani et al., 2025) achieves*

$$\mathrm{MSE}_\gamma\left(\hat{\mathbf{W}}^{\mathrm{AMP}}\right) = \Theta_{\alpha,m_\star}\left(\alpha^{-(1-1/2\gamma)/\beta_\star}\right). \tag{18}$$

*Moreover, its k-critical threshold* $\alpha_k^{\mathrm{AMP}}$ *satisfies* $\alpha_k^{\mathrm{AMP}} = \Theta_k\left(k^{2\gamma\beta_\star}\right)$.

The proof of this result is discussed in Appendix C, in which we also show that the spectral estimator of Definition 3.3 achieves the algorithmic bounds in Theorem 3.13.

There is an interesting consequence to this result if one follows the conclusion of the statistical mechanics approach to statistical-to-computational gaps (Bandeira et al., 2018; Aubin et al., 2018; Tsirkas et al., 2026): while the NSE changes the computational rate for feature emergence, one expects the information-theoretic rates to remain insensitive to the NSE. Indeed the replica-symmetric picture (Aubin et al., 2018; Troiani et al., 2025) suggests that information-theoretic recovery is governed by the global maximizer of a functional called the "replica-symmetric potential", whereas efficient recovery is controlled by the local instability of a trivial critical point of this potential (Zdeborová and Krzakala, 2016; Franz and Parisi, 1998). As such, a larger NSE not only modifies the computational rates, but also seems to open a genuine statistical-to-computational gap.

## 4 Conclusion and limitations

We established a sharp theory of scaling laws for hierarchical multi-index models in a genuinely representation-limited regime. The main message is that optimal representation learning is sequential: as the sample size grows, the relevant directions become detectable one by one, each through a new outlier in the spectrum, producing plateaus and abrupt drops in prediction error. We derive the corresponding Bayes-optimal rates for subspace recovery and prediction, show that they are achieved by a simple target-agnostic spectral estimator, and prove that once this representation is recovered, the readout can be learned without an additional statistical bottleneck. These results unify neural scaling laws, progressive feature emergence, and the structured spectra observed in trained weights within a single tractable framework, while showing that the power laws previously identified in quadratic settings extend to a (much) broader class of hierarchical multi-index targets. We also provide exceptions to this universality class.

Our analysis deliberately focuses on an idealized setting: isotropic Gaussian inputs, two-layer architectures, and high-dimensional asymptotics. These assumptions are a limitation in scope, but also the source of the model's strength. By stripping the problem down to its essential ingredients – hierarchical target geometry, representation-limited learning, and spectral recovery of the signal subspace – the framework yields exact, non-perturbative predictions and exposes the mechanisms behind scaling laws and feature emergence. We therefore view this model not as an endpoint, but as a clean foundation on which more realistic data distributions, architectures, and training dynamics, including SGD, can be incorporated systematically.

## Acknowledgments

We would like to thank Yatin Dandi, Vittorio Erba and Lenka Zdeborová for insightful discussions. BL and LD were supported by the French government, managed by the National Research Agency (ANR), under the France 2030 program with the project references "ANR-23-IACL-0008" (PR[AI]RIE-PSAI) and "ANR-25-CE23-5660" (MAPLE), as well as the Choose France - CNRS AI Rising Talents program. FK acknowledges funding from the Swiss National Science Foundation grants OperaGOST (grant number 200021 200390) and DSGIANGO (grant number 225837). This work was supported by the Simons Collaboration on the Physics of Learning and Neural Computation via the Simons Foundation grant (#1257412).

## References


E. Abbe, E. B. Adserà, and T. Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In G. Neu and L. 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.

M. Andriushchenko, A. V. Varre, L. Pillaud-Vivien, and N. Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

A. Atanasov, J. A. Zavatone-Veth, and C. Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

B. Aubin, A. Maillard, F. Krzakala, N. Macris, L. 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.

D. Babichev and F. 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.

Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

J. Baik, G. B. Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643 – 1697, 2005.

A. Bandeira, A. Perry, and A. S. Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. *Portugaliae mathematica*, 75(2):159–186, 2018.

J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. 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.

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

G. Ben Arous, M. A. Erdogdu, N. M. Vural, and D. Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. *arXiv preprint arXiv:2508.03688*, 2025.

R. Berthier, A. Montanari, and P.-M. Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020.

F. Boncoraglio, V. Erba, E. Troiani, Y. Xu, F. Krzakala, and L. Zdeborová. Single-head attention in high dimensions: A theory of generalization, weights spectra, and scaling laws. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

T. Bonnaire, G. Biroli, and C. Cammarota. The role of the time-dependent Hessian in high-dimensional optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2025 (8):083401, 2025.

B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

B. Bordelon, A. Atanasov, and C. Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4345–4382, 2024.

B. Bordelon, A. Atanasov, and C. Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084002, 2025.

G. Braun, B. Loureiro, H. Q. Minh, and M. Imaizumi. Fast escape, slow convergence: Learning dynamics of phase retrieval under power-law data. *arXiv preprint arXiv:2511.18661*, 2025.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

F. Cagnetta, L. Petrini, U. M. Tomasini, A. Favero, and M. Wyart. How deep neural networks learn compositional data: The Random Hierarchy Model. *Physical Review X*, 14:031001, 2024. doi: 10.1103/PhysRevX.14.031001.

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

M. Celentano, A. Montanari, and Y. Wu. The estimation error of general first order methods. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1078–1141. PMLR, 09–12 Jul 2020.

H. Cui, B. Loureiro, F. Krzakala, and L. 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.

H. Cui, B. Loureiro, F. Krzakala, and L. Zdeborová. Error scaling laws for kernel classification under source and capacity conditions. *Machine Learning: Science and Technology*, 4(3):035033, 2023.

A. Damian, J. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

A. Damian, L. Pillaud-Vivien, J. D. Lee, and J. Bruna. Computational-statistical gaps in Gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

Y. Dandi, F. Krzakala, B. Loureiro, L. Pesce, and L. Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024.

Y. Dandi, L. Pesce, L. Zdeborová, and F. Krzakala. The computational advantage of depth in learning high-dimensional hierarchical targets. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

L. Defilippis, B. Loureiro, and T. Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. *Advances in Neural Information Processing Systems*, 37: 104630–104693, 2024.

L. Defilippis, Y. Dandi, P. Mergny, F. Krzakala, and B. Loureiro. Optimal spectral transitions in high-dimensional multi-index models. *Journal of Statistical Mechanics: Theory and Experiment*, 2026:084010, 2026a. doi: 10.1088/1742-5468/ae937b.

L. Defilippis, F. Krzakala, B. Loureiro, and A. Maillard. A noise sensitivity exponent controls large statistical-to-computational gaps in single- and multi-index models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 23562–23590, 2026b.

L. Defilippis, Y. Xu, J. Girardin, E. Troiani, V. Erba, L. Zdeborová, B. Loureiro, and F. Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. In *International Conference on Learning Representations*, 2026c.

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

D. L. Donoho, M. Gavish, and A. 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.

R. Dudeja, Y. M. Lu, and S. Sen. Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.

V. Erba, E. Troiani, L. Zdeborová, and F. Krzakala. The nuclear route: Sharp asymptotics of ERM in overparameterized quadratic networks. In *Advances in Neural Information Processing Systems*, volume 38, pages 88862–88901, 2025.

O. Y. Feng, R. Venkataramanan, C. Rush, and R. 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.

D. Ferbach, K. Everett, G. Gidel, E. Paquette, and C. Paquette. Dimension-adapted momentum outscales SGD. *arXiv preprint arXiv:2505.16098*, 2025.

S. Franz and G. Parisi. Effective potential in glassy systems: theory and simulations. *Physica A: Statistical Mechanics and its Applications*, 261(3-4):317–339, 1998.

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

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in neural information processing systems*, 35:30016–30030, 2022.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991.

A. Javanmard and A. 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.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

F. Kovačević, Y. Zhang, and M. Mondelli. Spectral estimators for multi-index models: Precise asymptotics and optimal weak recovery. In N. Haghtalab and A. 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.

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

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Y. M. Lu and G. Li. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541, 2020.

A. Maillard, F. Krzakala, Y. M. Lu, and L. Zdeborová. Construction of optimal spectral methods in phase retrieval. In *Mathematical and Scientific Machine Learning*, pages 693–720. PMLR, 2022.

A. Maillard, E. Troiani, S. Martin, F. Krzakala, and L. Zdeborová. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. *Advances in Neural Information Processing Systems*, 37:82085–82132, 2024.

S. Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

A. Maloney, D. A. Roberts, and J. Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

C. H. Martin and M. 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.

S. Martin, G. Biroli, and F. Bach. High-dimensional analysis of gradient flow for extensive-width quadratic neural networks. *arXiv preprint arXiv:2601.10483*, 2026.

M. Mondelli and A. 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.

A. Montanari and Y. Wu. Statistically optimal first-order algorithms: a proof via orthogonalization. *Information and Inference: A Journal of the IMA*, 13(4):iaae027, 2024. doi: 10.1093/imaiai/iaae027.

R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

E. Paquette, C. Paquette, L. Xiao, and J. Pennington. 4+3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37:16459–16537, 2024.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

Y. Ren, E. Nichani, D. Wu, and J. D. Lee. Emergence and scaling laws in SGD learning of shallow neural networks. *arXiv preprint arXiv:2504.19983*, 2025.

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 Representations 2014*. International Conference on Learning Representations 2014, 2014.

R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

S. Spigler, M. Geiger, and M. Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020 (12):124001, 2020.

H. Tabanelli, Y. Dandi, L. Pesce, and F. Krzakala. Deep learning of compositional targets with hierarchical spectral methods. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 117916–117938, 2026.

M. Thamm, M. Staats, and B. Rosenow. Random matrix theory analysis of neural network weight matrices. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

E. Troiani, Y. Dandi, L. Defilippis, L. Zdeborova, B. Loureiro, and F. Krzakala. Fundamental computational limits of weak learnability in high-dimensional multi-index models. In Y. Li, S. Mandt, S. Agrawal, and E. 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.

K. Tsirkas, L. Wang, and I. Zadik. The monotonicity of the Franz-Parisi potential is equivalent with low-degree MMSE lower bounds. *arXiv preprint arXiv:2603.20070*, 2026.

Z. Wang, A. Engel, A. D. Sarwate, I. Dumitriu, and T. Chiang. Spectral evolution and invariance in linear-width neural networks. *Advances in neural information processing systems*, 36:20695–20728, 2023.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

R. Worschech and B. Rosenow. Analyzing neural scaling laws in two-layer networks with power-law data spectra. *arXiv preprint arXiv:2410.09005*, 2024.

A. Wortsman and B. Loureiro. Kernel ridge regression under power-law data: spectrum and generalization. *arXiv preprint arXiv:2510.04780*, 2025.

Y. Xu, A. Maillard, L. Zdeborová, and F. Krzakala. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications. In *Proceedings of the Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 5757–5823, 2025.

M. Yuan. On the identifiability of additive index models. *Statistica Sinica*, pages 1901–1911, 2011.

L. Zdeborová and F. Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

B. Zhang, Z. Wang, H. Fu, and J. D. Lee. Neural networks learn generic multi-index models near information-theoretic limit, 2025. URL https://arxiv.org/abs/2511.15120.

Q. Zou and H. Yang. A concise tutorial on approximate message passing. *arXiv preprint arXiv:2201.07487*, 2022.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the main theoretical contributions and their scope: Bayes-optimal scaling laws, a matching spectral estimator, a two-stage neural-network learning procedure, and the dependence on the noise-sensitivity exponent. These claims are formalized in Section 3 and proved in the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Conclusion and limitations" section discussing the idealized scope of the analysis, including isotropic Gaussian inputs, two-layer architectures, and high-dimensional asymptotics. It also identifies extensions to more realistic data distributions, architectures, and SGD dynamics as future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are stated in Definitions 2.1, 2.2 and Assumptions 2.3, 3.4, and 3.12. Proofs of Theorems 3.2, 3.6, 3.9, and 3.13 are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the synthetic data-generating model, the spectral estimator, and the preprocessing used for the finite-size experiments in Definition 2.2, Definition 3.3, and Figures 1–4. These details are sufficient to reproduce the experiments supporting the illustrative empirical figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not currently provide open-source code or execution commands for reproducing the finite-size synthetic experiments. The experiments use synthetic data generated from the model specified in the paper, so no external dataset access is required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experimental model, preprocessing function, estimator, data dimension, feature dimension, coefficient decay, and number of averaged instances are specified in Definitions 2.2–3.3 and Figures 1–4. Algorithm 1 also specifies the train/test split and ridge readout procedure used in the theoretical neural-network construction.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Figure 1 reports averages over 30 independent instances, but the paper does not currently include error bars, confidence intervals, or statistical significance tests. The empirical figures are illustrative finite-size validations of the proved asymptotic scaling laws rather than the primary basis for the claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The paper specifies the problem dimensions for the reported finite-size experiments, but it does not report the hardware type, memory, runtime per experiment, or total compute budget. The main results are theoretical, and the experiments are lightweight synthetic illustrations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is theoretical and uses synthetic data generated from explicitly defined mathematical models, with no human subjects, private data, or high-risk deployment. To the best of the authors' knowledge, the research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not include a separate broader impacts discussion. The work is foundational theory for learning in hierarchical multi-index models and does not introduce a deployed system, dataset, or model with a direct path to societal harm, although improved understanding of neural-network learning may have broad indirect impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pretrained models, scraped datasets, generative models, or other assets that pose a high risk for misuse. The empirical component uses synthetic data generated from the mathematical model in the paper.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use external datasets, pretrained models, or code assets as experimental inputs. Prior theoretical and methodological results used in the analysis are cited in the text and bibliography.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, code packages, pretrained models, or other reusable assets. The synthetic data-generating process is fully specified as part of the theoretical model.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research does not involve crowdsourcing, participant studies, or other human-subject data collection. All experiments use synthetic data from the model defined in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects or crowdsourcing, so IRB or equivalent approval is not applicable. No participant risks or disclosures arise from the work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

Justification: The core methods and scientific contributions do not involve LLMs as an important, original, or non-standard component. Any possible use of LLMs for writing, editing, or formatting would not affect the methodology, rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

# A Proof of Theorems 3.2 and 3.6

In this section, we prove the information-theoretic results stated in Theorem 3.2. In particular, we derive matching bounds for the $\mathrm{MMSE}_\gamma$, defined 3.1, and the $k$-critical threshold $\alpha_k$ (Definition 2.6). In Appendix A.1, we analyze an oracle estimator that exploits additional information to achieve a weighted mean-squared error smaller than $\mathrm{MMSE}_\gamma$. Finally, in Appendix A.2, we complete the proof by characterizing the scaling laws and $k$-critical thresholds of the spectral estimator defined in 3.3, which simultaneously establishes the results in Theorem 3.6.

## A.1 Lower bound

**Definition A.1** (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 $\boldsymbol{W}^{\mathrm{oracle}} \in \mathbb{R}^{m_\star \times d}$ with rows*

$$\boldsymbol{w}_k^{\mathrm{oracle}} = \arg\min_{\boldsymbol{w}} \mathbb{E}\left[\|\boldsymbol{w}_k\boldsymbol{w}_k^\top - \boldsymbol{w}_k^\star\boldsymbol{w}_k^{\star\top}\|_{\mathrm{F}}^2 \,|\, \mathcal{D}, \{\boldsymbol{w}_h\}_{h\neq k}\right], k \in [m_\star]. \tag{19}$$

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

$$\begin{aligned}\mathrm{MMSE}_\gamma^{\mathrm{oracle}} = \frac{1}{d^2}\sum_{k=1}^{m_\star}(a_k^\star)^2\,\mathbb{E}\Big[\Big\|\mathbb{E}[\boldsymbol{w}_k\boldsymbol{w}_k^\top|\mathcal{D}, \{\boldsymbol{w}_h^\star\}_{h\neq k}] \\ - \boldsymbol{w}_k^\star\boldsymbol{w}_k^{\star\top}\Big\|_{\mathrm{F}}^2\Big].\end{aligned} \tag{20}$$

As a consequence of the law of total variance

$$d^2\,\mathrm{mmse}_k := \mathbb{E}\left[\mathrm{Cov}\left(\boldsymbol{w}_k\boldsymbol{w}_k^\top \,|\, \mathcal{D}\right)\right] \tag{21}$$

$$= \mathbb{E}\left[\mathrm{Cov}\left(\boldsymbol{w}_k\boldsymbol{w}_k^\top \,|\, \mathcal{D}, \{\boldsymbol{w}_h\}_{h\neq k}\right)\right] + \underbrace{\mathbb{E}\left[\mathrm{Cov}_{\{\boldsymbol{w}_h\}_{h\neq k}}\left(\mathbb{E}\left[\boldsymbol{w}_k\boldsymbol{w}_k^\top \,|\, \mathcal{D}, \{\boldsymbol{w}_h\}_{h\neq k}\right]\right)\right]}_{\geq 0} \tag{22}$$

$$\geq d^2\,\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{w}_k\boldsymbol{w}_k^\top|\mathcal{D}, \{\boldsymbol{w}_h^\star\}_{h\neq k}] - \boldsymbol{w}_k^\star\boldsymbol{w}_k^{\star\top}\|_{\mathrm{F}}^2\right]. \tag{23}$$

Therefore,

$$\mathrm{MMSE}_\gamma^{\mathrm{oracle}} \leq \mathrm{MMSE}_\gamma. \tag{24}$$

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 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

$$\overline{y}_i = a_k^\star g_k(\langle \boldsymbol{w}_k^\star, \boldsymbol{x}_i\rangle) + \sqrt{\Delta}\xi_i, \tag{25}$$

where $\xi_i$ 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 Barbier et al. (2019). In Appendix D 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 2.3, this implies that the sequence of $k$-critical thresholds $\alpha_k$ is bounded by strictly increasing functions.[1] In particular, the first result in Corollary D.4, for a signal-to-noise ratio $\lambda = (a_k^\star)^2/\Delta$, implies that the sequence is bounded by

$$\alpha_k^{\mathrm{oracle}} = \Theta\left((a_k^\star)^{-2}\right) = \Theta(k^{2\gamma} m_\star^{(1-2\gamma)_+}), \tag{26}$$

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

[1] If $g_k = g$, $\forall k$, the sequence $\alpha_k^{\mathrm{oracle}}$ itself is strictly increasing.

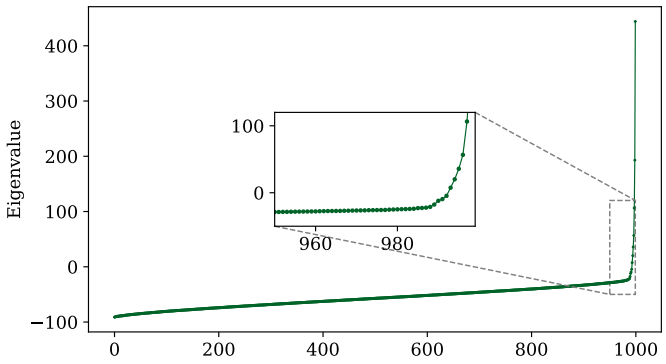



Figure 3: Empirical spectrum of $\boldsymbol{T}$ eq. (13) with preprocessing $\mathcal{T}(y) = y/(1+|y|)$, for a hierarchical multi-index model with $g_k(z) = \frac{1}{2}\mathrm{He}_2(z) + \frac{1}{2\cdot 4!}\mathrm{He}_4(z)$ and $\gamma = 1.3$. The covariates dimension is $d = 1000$, while the feature 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. 3.3.

and exploiting the second result in Corollary D.4,

$$\mathrm{MMSE}_\gamma^{\mathrm{oracle}} = \left(\sum_{k=1}^{k_\alpha} + \sum_{k=k_\alpha+1}^{m_\star}\right)(a_k^\star)^2 \mathbb{E}\left[\|\mathbb{E}[\boldsymbol{w}_k \boldsymbol{w}_k^\top | \mathcal{D}, \{\boldsymbol{w}_h^\star\}_{h\neq k}] - \boldsymbol{w}_k^\star \boldsymbol{w}_k^{\star\top}\|_{\mathrm{F}}^2\right] \tag{27}$$

$$\geq C\left(\sum_{k=1}^{k_\alpha} \frac{(a_k^\star)^2}{(a_k^\star)^2\,\alpha} + \sum_{k=k_\alpha+1}^{m_\star}(a_k^\star)^2\right) \tag{28}$$

$$= \begin{cases} \Theta\left(\alpha^{-1+1/(2\gamma)}\right), & \gamma > 1/2,\ \alpha \ll m_\star^{2\gamma} \\ \Theta\left(m_\star/\alpha\right), & \gamma > 1/2,\ \alpha \gg m_\star^{2\gamma} \\ \Theta\left(1\right), & \gamma < 1/2,\ \alpha \ll m_\star \\ \Theta\left(m_\star/\alpha\right), & \gamma < 1/2,\ \alpha \gg m_\star \end{cases} \tag{29}$$

Note that we have used

$$\sum_{k=k_\alpha+1}^{m_\star} k^{-2\gamma} = \begin{cases} \Theta(k_\alpha^{1-2\gamma}), & \gamma > 1/2, \\ \Theta(m_\star^{1-2\gamma}), & \gamma < 1/2. \end{cases} \tag{30}$$

In eq. (28), 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.

### A.2 Upper bound – (Theorem 3.6)

By definition, any estimator that is a function of the dataset only has a weighted mean-squared error larger than $\mathrm{MMSE}_\gamma$. In this section we consider the spectral method defined in 3.3. The following Theorem, proven in Kovačević et al. (2025), 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 $Y \sim \mathcal{N}\left(\sum_{k=1}^{m_\star} a_k^\star g_k(z_k),\, \Delta\right)$.

**Theorem A.2** (Theorem 4.1 in Kovačević et al. 2025)**.** *Let $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ be a preprocessing function subject to Assumption 3.4 and $\boldsymbol{T}$ defined as in eq. 13. Let $t_1 \geq t_2 \geq \ldots \geq t_r \geq \tau$, for some $r \in [m_\star]$, be all the solutions to*

$$\det\left(\alpha \mathbb{E}\left[\frac{(\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_{m_\star})\mathcal{T}(Y)}{t - \mathcal{T}(Y)}\right] - \boldsymbol{I}_{m_\star}\right) = 0 \tag{31}$$

*such that*

$$t_k \geq \bar{t}_\alpha := \operatorname*{arg\,min}_{t\geq\tau} \zeta_\alpha(t), \quad \forall k \in [r], \tag{32}$$

*where*

$$\zeta_\alpha(t) := t\left(1 + \alpha \mathbb{E}\left[\frac{\mathcal{T}(Y)}{t - \mathcal{T}(Y)}\right]\right). \tag{33}$$

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

$$\lambda_1^{\boldsymbol{T}}, \dots, \lambda_r^{\boldsymbol{T}} \xrightarrow{\text{a.s.}} \zeta_\alpha(t_1), \dots, \zeta_\alpha(t_r), \tag{34}$$

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

$$\lambda_{r+1}^{\boldsymbol{T}}, \dots, \lambda_{m_\star}^{\boldsymbol{T}} \xrightarrow{\text{a.s.}} \zeta_\alpha(\bar{t}_\alpha). \tag{35}$$

In particular, the top $r$ eigenvalues of $\boldsymbol{T}$ are separated from each other and from the remaining ones by gaps of order one, while the gaps between the following $m_\star - r$ eigenvalues vanish as $d \to \infty$. Therefore, for $d$ large enough, the eigenvectors selected by the spectral estimator of Definition 3.3 are those associated with the $r$ outliers of Theorem A.2.

As a first result, we derive a bound for $\alpha_k$.

Consider the matrix in eq. (31)

$$\mathbb{E}\left[\frac{(\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_{m_\star})\mathcal{T}(Y)}{t - \mathcal{T}(Y)}\right] = \mathbb{E}_Y\left[\boldsymbol{G}(Y)\frac{\mathcal{T}(Y)}{t - \mathcal{T}(Y)}\right], \tag{36}$$

with $\boldsymbol{G}(y) = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_{m_\star} | Y = y]$. It is straightforward to show that it is a diagonal matrix, due to the parity of each function $g_k$. Indeed, given $k \neq h$,

$$G_{kh}(y) = \mathbb{E}[z_k z_h | y] \propto \int e^{-\|\boldsymbol{z}\|^2/2} \exp\left(-\left(y - \sum_{k'=1}^{m_\star} a_{k'} g_{k'}(z_{k'})\right)^2 / (2\Delta)\right) z_h z_k \, \mathrm{d}\boldsymbol{z} = 0. \tag{37}$$

For simplicity, we denote $G_k(y) := G_{kk}(y)$. Therefore, the solutions of eq. (31) coincides with the solutions of

$$\alpha^{-1} = \mathbb{E}\left[\frac{\mathcal{T}(Y)}{t - \mathcal{T}(Y)} G_k(Y)\right], \quad k \in [m_\star]. \tag{38}$$

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

$$\alpha^{-1} = \mathbb{E}\left[\left(\frac{\mathcal{T}(Y)}{t - \mathcal{T}(Y)}\right)^2\right]. \tag{39}$$

By Theorem A.2 and Theorem 4.2 in Kovačević et al. (2025), the spectral estimator has an eigenvector with a finite overlap with $\boldsymbol{w}_k^\star$ if and only if eq. (38) has a solution $t_k > \bar{t}_\alpha$. For $k$ large enough, the right-hand side of eq. (38) is positive and decreasing in $t$ for $t \geq \bar{t}_\alpha$. This follows from the expansion of the first line of eq. (60) below, which holds for any $t$ large enough, applied both to it and to its derivative $-\mathbb{E}[\mathcal{T}(Y)G_k(Y)/(t - \mathcal{T}(Y))^2]$, so that such a solution exists if and only if

$$u_k(\alpha) := \alpha \, \mathbb{E}\left[\frac{\mathcal{T}(Y)}{\bar{t}_\alpha - \mathcal{T}(Y)} G_k(Y)\right] > 1. \tag{40}$$

The $k$-critical threshold of the spectral estimator, Definition 2.6, is therefore $\alpha_k^{\text{sp}} = \inf\{\alpha > 0 : u_k(\alpha) > 1\}$. By continuity of $u_k$, $\alpha_k^{\text{sp}}$ solves $u_k(\alpha) = 1$, namely, by definition of $\bar{t}_\alpha$, the system

$$\begin{cases} (\alpha_k^{\text{sp}})^{-1} = \mathbb{E}_Y\left[\frac{\mathcal{T}(Y)}{\bar{t} - \mathcal{T}(Y)} G_k(Y)\right] \\ (\alpha_k^{\text{sp}})^{-1} = \mathbb{E}_Y\left[\left(\frac{\mathcal{T}(Y)}{\bar{t} - \mathcal{T}(Y)}\right)^2\right]. \end{cases} \tag{41}$$

In order to characterize the threshold, we consider the following expansions for small SNR $a_k^\star \ll 1$. Define $\mathsf{Z}$ the marginal distribution of $Y$

$$\mathsf{Z}(y) := \frac{1}{\sqrt{2\pi\Delta}} \mathbb{E}_{\boldsymbol{z}}\left[\exp\left(-\left(y - \sum_{k=1}^{m_\star} a_k^\star g_k(z_k)\right)^2 / (2\Delta)\right)\right]. \tag{42}$$

$$\tag{43}$$

Note that $\mathsf{Z} \in \mathcal{C}^\infty$ and, applying the definition of the probabilist's Hermite polynomials, its $j^{\text{th}}$ derivative reads

$$\frac{\mathrm{d}^j}{\mathrm{d}y^j}\mathsf{Z}(y) = \frac{(-1)^j}{\sqrt{2\pi\Delta^{j+1}}}\mathbb{E}_{\boldsymbol{z}}\left[e^{-\left(y-\sum_{k=1}^{m_\star} a_k^\star g_k(z_k)\right)^2/(2\Delta)}\mathrm{He}_k\left(\frac{y-\sum_{k=1}^{m_\star} a_k^\star g_k(z_k)}{\sqrt{\Delta}}\right)\right]. \tag{44}$$

We now aim to approximate integrals of the type $\mathbb{E}_Y[h(Y)G_k(Y)]$, where $h : \mathbb{R} \to \mathbb{R}$ is bounded, that are involved in eq. (31). In particular, we derive a bound for

$$\mathcal{E}_k := \mathbb{E}[h(Y)G_k(Y)] + a_k^\star \mathbb{E}_z[g_k(z)(z^2-1)]\int \mathrm{d}y h(y)\mathsf{Z}'(y). \tag{45}$$

In the following, we leverage the Taylor's expansion of the complex exponential, ensuring that, for all $\omega \in \mathbb{R}$, $k \in \mathbb{N}$, there exists a constant $c \in [0,\omega]$ such that[2]

$$R_k(\omega) := e^{i\omega} - \sum_{j=0}^{k}\frac{(i\omega)^j}{j!}, \qquad R_k(\omega) = \frac{(i\omega)^{k+1}}{k!}\int_0^1 e^{i\omega t}(1-t)^k \mathrm{d}t \tag{46}$$

In addition, we use the Fourier representation of the Gaussian density $(2\pi)^{-1/2}e^{-x^2/2} = (2\pi)^{-1}\int \mathrm{d}\omega e^{i\omega x-\omega^2/2}$ and the identity

$$\frac{1}{2\pi}\int \mathrm{d}\omega e^{i\omega x-\omega^2}(i\omega)^j = \frac{1}{2\pi}\frac{\mathrm{d}^j}{\mathrm{d}x^j}\int \mathrm{d}\omega e^{i\omega x-\omega^2} = (-1)^j\frac{e^{-x^2/2}}{\sqrt{2\pi}}\mathrm{He}_j(x). \tag{47}$$

Define the shorthand $\rho_k^{jj'} := \mathbb{E}_z[g_k^j(z)\mathrm{He}_{j'}(z)]$, for $j,j' \in \{0,1,2\}$, which are non-zero and finite by Assumption 2.3. In order to simplify the following equations, let $\mathbb{E}_{\boldsymbol{z}_{\setminus k}}$ denote the expected value with respect to $\{z_h\}_{h\neq k}$ and $S_k(\boldsymbol{z}_{\setminus k}) := \sum_{h\neq k} a_h^\star g_h(z_h)$. Then, we have

$$\begin{aligned}
\mathcal{E}_k &= \int \mathrm{d}y h(y)\mathbb{E}_{\boldsymbol{z}}\left[\mathsf{P}(y|\boldsymbol{z})(z_k^2-1) + a_k^\star\rho_k^{12}\partial_y\mathsf{P}(y|\boldsymbol{z})\right] && (48)\\
&\overset{(47)}{=}\mathbb{E}_{\boldsymbol{z}_{\setminus k}}\frac{1}{2\pi}\int \mathrm{d}y h(y)\int \mathrm{d}\omega e^{i\omega(y-S_k(\boldsymbol{z}_{\setminus k}))-\omega^2\Delta/2}\mathbb{E}_z\left[e^{-i\omega a_k^\star g_k(z)}(z^2-1)\right] && (49)\\
&\quad + \frac{a_k^\star\rho_k^{12}}{2\pi}\mathbb{E}_{\boldsymbol{z}_{\setminus k}}\int \mathrm{d}y h(y)\int \mathrm{d}\omega(i\omega)e^{i\omega(y-S_k(\boldsymbol{z}_{\setminus k}))-\omega^2\Delta/2}\mathbb{E}_z[e^{-i\omega a_k^\star g_k(z)}] && (50)\\
&= \frac{1}{2\pi}\mathbb{E}_{\boldsymbol{z}_{\setminus k}}\int \mathrm{d}y h(y)\int \mathrm{d}\omega e^{i\omega(y-S_k(\boldsymbol{z}_{\setminus k}))-\omega^2\Delta/2}\\
&\quad \times \underbrace{\mathbb{E}_z\left[(1-i\omega a_k^\star g_k(z) + R_1(\omega a_k^\star g_k(z)))(z^2-1)\right]}_{0-i\omega a_k^\star\rho_k^{12}+\mathbb{E}_z[R_1(\omega a_k^\star g_k(z)))(z^2-1)]} && (51)\\
&\quad + \frac{a_k^\star\rho_k^{12}}{2\pi}\mathbb{E}_{\boldsymbol{z}_{\setminus k}}\int \mathrm{d}y h(y)\int \mathrm{d}\omega(i\omega)e^{i\omega(y-S_k(\boldsymbol{z}_{\setminus k}))-\omega^2\Delta/2}\mathbb{E}_z[1+R_0(\omega a_k^\star g_k(z)))]. && (52)
\end{aligned}$$

[2]The following expression is obtained as the integral form of the remainder in Taylor's theorem applied to the real functions $\sin(\omega)$ and $\cos(\omega)$ in $e^{i\omega} = \cos(\omega) + i\sin(\omega)$.

After the cancellation of the terms in $a_k^\star$, substituting the integral form of the remainder in eq. (46)

$$\begin{aligned}
\mathcal{E}_k =& \frac{(a_k^\star)^2}{2\pi}\mathbb{E}_{\boldsymbol{z}} g_k(z_k)(z_k^2-1)\int_0^1 \mathrm{d}t(1-t) \\
&\times \int \mathrm{d}y h(y)\int \mathrm{d}\omega (i\omega)^2 e^{i\omega(y-S_k(\boldsymbol{z}_{\setminus k}))-\omega^2\Delta/2} e^{it\omega a_k^\star g_k(z_k)} \qquad (53)\\
&+ \frac{(a_k^\star)^2\rho_k^{12}}{2\pi}\mathbb{E}_{\boldsymbol{z}} g_k(z_k)\int_0^1 \mathrm{d}t \\
&\times \int \mathrm{d}y h(y)\int \mathrm{d}\omega (i\omega)^2 e^{i\omega(y-S_k(\boldsymbol{z}_{\setminus k}))-\omega^2\Delta/2} e^{it\omega a_k^\star g_k(z_k)} \qquad (54)\\
\overset{(47)}{=}& \frac{(a_k^\star)^2}{\Delta\sqrt{2\pi\Delta}}\mathbb{E}_{\boldsymbol{z}} g_k^2(z_k)(z_k^2-1)\int_0^1 \mathrm{d}t(1-t) \\
&\times \int \mathrm{d}y h(y) e^{-(y-S_k(\boldsymbol{z}_{\setminus k})+ta_k^\star g_k(z_k))^2/(2\Delta)}\mathrm{He}_2\left(\frac{y-S_k(\boldsymbol{z}_{\setminus k})+ta_k^\star g_k(z_k)}{\sqrt{\Delta}}\right) \qquad (55)\\
&+ \frac{(a_k^\star)^2\rho_k^{12}}{\Delta\sqrt{2\pi\Delta}}\mathbb{E}_{\boldsymbol{z}} g_k(z_k)\int_0^1 \mathrm{d}t \\
&\times \int \mathrm{d}y h(y) e^{-(y-S_k(\boldsymbol{z}_{\setminus k})+ta_k^\star g_k(z_k))^2/(2\Delta)}\mathrm{He}_2\left(\frac{y-S_k(\boldsymbol{z}_{\setminus k})+ta_k^\star g_k(z_k)}{\sqrt{\Delta}}\right) \qquad (56)\\
=& \frac{(a_k^\star)^2}{\Delta\sqrt{2\pi}}\mathbb{E}_{\boldsymbol{z}} g_k^2(z_k)(z_k^2-1)\int_0^1 \mathrm{d}t(1-t) \\
&\times \int \mathrm{d}y h\left(\sqrt{\Delta}y+S_k(\boldsymbol{z}_{\setminus k})-ta_k^\star g_k(z_k)\right) e^{-y^2/2}\mathrm{He}_2(y) \qquad (57)\\
&+ \frac{(a_k^\star)^2\rho_k^{12}}{\Delta\sqrt{2\pi}}\mathbb{E}_{\boldsymbol{z}} g_k(z_k)\int_0^1 \mathrm{d}t \\
&\times \int \mathrm{d}y h\left(\sqrt{\Delta}y+S_k(\boldsymbol{z}_{\setminus k})-ta_k^\star g_k(z_k)\right) e^{-y^2/2}\mathrm{He}_2(y)\,. \qquad (58)
\end{aligned}$$

Recalling that $h$ is a bounded function, we obtain

$$|\mathcal{E}_k| \le \sup_{y\in\mathbb{R}}|h(y)|\frac{(a_k^\star)^2}{\Delta\sqrt{2\pi}}\int \mathrm{d}y e^{-y^2/2}|\mathrm{He}_2(y)|\left(\frac{\rho_k^{22}}{2}+\rho_k^{12}\rho_k^{10}\right). \qquad (59)$$

Note that, by Assumption 2.3, the constants $\rho_k^{jj'}$ do not scale with $k$, therefore $|\mathcal{E}_k| = O((a_k^\star)^2)$. We now look for $t > \tau$ and $\alpha$ that are solutions of eqs. (41). In particular, the oracle $k$-critical threshold in Appendix A.1 is a lower bound to the spectral one, therefore $\alpha_k^{\mathrm{sp}} = \Omega(\alpha_k^{\mathrm{oracle}})$ and diverges with $k$. The second equation of (41), due to the monotonicity of the RHS, implies that also the solution $t$ must diverge with $k$. Using the bound we have derived for $\mathcal{E}_k$,

$$\begin{cases}(\alpha_k^{\mathrm{sp}})^{-1} = -t^{-1}a_k^\star\mathbb{E}_z[g_k(z)(z^2-1)]\int \mathrm{d}y\mathcal{T}(y)\mathsf{Z}'(y) + O(a_k^\star t^{-2} + (a_k^\star)^2 t^{-1}) \\ (\alpha_k^{\mathrm{sp}})^{-1} = t^{-2}\mathbb{E}[\mathcal{T}(y)^2] + O(t^{-3}).\end{cases} \qquad (60)$$

The leading term on the right-hand side of the first equation is positive, so that the informative outliers emerge from the right edge of the bulk. Indeed, $\int \mathrm{d}y\,\mathcal{T}(y)\mathsf{Z}'(y) = -\int \mathsf{Z}(y)\,\mathrm{d}\mathcal{T}(y) < 0$, since $\mathcal{T}$ is strictly increasing, and $\mathbb{E}_z[g_k(z)(z^2-1)] > 0$, since $g_k$ is even and strictly increasing on $(0,\infty)$. Note that this is the only place where the monotonicity of $g_k$ enters. In the previous computations, we have used

$$\begin{aligned}
\left|\int \mathrm{d}y\left(\frac{\mathcal{T}(y)\mathsf{Z}'(y)}{t-\mathcal{T}(y)} - t^{-1}\mathcal{T}(y)\mathsf{Z}'(y)\right)\right| &= t^{-2}\left|\int \mathrm{d}y\frac{\mathcal{T}^2(y)\mathsf{Z}'(y)}{1-\mathcal{T}(y)/t}\right| \qquad (61)\\
&\le t^{-2}\sup_{y\in\mathbb{R}}\left(\frac{\mathcal{T}^2(y)}{1-\mathcal{T}(y)/t}\right)\int \mathrm{d}y|\mathsf{Z}'(y)| \qquad (62)\\
&\overset{(44)}{=} t^{-2}\sup_{y\in\mathbb{R}}\left(\frac{\mathcal{T}^2(y)}{1-\mathcal{T}(y)/t}\right)\int \mathrm{d}y\frac{e^{-y^2/2}}{\sqrt{2\pi\Delta}}|y| = O(t^{-2}), \qquad (63)
\end{aligned}$$

where, for $t$ sufficiently large, the supremum is bounded by a constant independent of $t$. The system (60) is thereby solved for

$$t^{-1} = \Theta(a_k^\star), \qquad \alpha_k^{\rm sp} = \Theta(\alpha_k^{\rm oracle}). \tag{64}$$

Moreover, for $\alpha \gg 1$, the second line of eq. (60) gives $\bar{t}_\alpha = \Theta(\sqrt{\alpha})$, and the first one gives $u_k(\alpha) \geq c\, a_k^\star \sqrt{\alpha}$ for $k$ large enough, with $c > 0$ independent of $k$ and $\alpha$. Therefore, as $u_k(\alpha) > 1$ for all $\alpha > (c\, a_k^\star)^{-2} = \Theta(\alpha_k^{\rm sp})$, once direction $k$ is weakly recovered, it remains recovered at larger sample complexities.

**Remark A.3** (Outliers at the left edge of the bulk)**.** *If $\mathbb{E}_z[g_k(z)(z^2-1)] < 0$, the leading term in the first line of eq. (60) is negative, and eq. (38) has no solution $t_k > \bar{t}_\alpha$, which implies that direction $k$ produces no outlier at the right edge of the bulk. Since the smallest eigenvalues of $\boldsymbol{T}$ are the largest eigenvalues of $-\boldsymbol{T}$, i.e. of the matrix in eq. (13) with preprocessing $-\mathcal{T}$, applying Theorem A.2 to $-\mathcal{T}$ and repeating the computation above, the leading term in the first line of eq. (60) changes sign, while the second line is unchanged. Direction $k$ then produces an outlier at the left edge of the bulk, with the same scaling of the threshold and of the overlap. Hence, without the monotonicity of $g_k$, Theorem 3.6 holds for the spectral estimator that retains the outliers at both edges of the bulk.*

We can now derive the scaling laws for the spectral weighted mean-squared error. Taking $\alpha \gg \alpha_k^{\rm sp}$, for $k$ large enough, the solution $t_k$ of eq. (38) scales as $t_k = \Theta(\alpha a_k^\star)$. As a consequence of Theorem 4.2 in Kovačević et al. (2025), specialized in our setting – where we can exploit the diagonality of $\boldsymbol{G}(y)$, each solution of eq. (31) is associated with a single direction, and we denote by $t_h$ the solution associated with $\boldsymbol{w}_h^\star$. Let $\pi : [r] \to [m_\star]$ be the injective map such that the $k^{\rm th}$ largest eigenvalue $\lambda_k^{\boldsymbol{T}}$ converges to $\zeta_\alpha(t_{\pi(k)})$. Then, 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

$$m_{hk}^2 := \frac{1}{d^2}\langle \boldsymbol{w}_h^\star, \boldsymbol{v}_k\rangle^2 = \delta_{h\pi(k)} \frac{\zeta_\alpha'(t_h)}{\zeta_\alpha'(t_h) - \frac{\rm d}{{\rm d}t} R_h(t_h)}, \tag{65}$$

where $R_h(t) = \alpha t\, \mathbb{E}[z_h^2 \mathcal{T}(y)/(t - \mathcal{T}(y))]$. For $\alpha \gg \alpha_k^{\rm sp}$ we have that $t_k \gg 1$ and eq. (38) is solved by $t_k = \Theta(a_k^\star \alpha)$. Putting all together

$$m_{hk}^2 = \delta_{h\pi(k)} \left(1 - \Theta\left(\frac{(a_h^\star)^{-2}}{\alpha}\right)\right). \tag{66}$$

Recall that

$$\mathrm{mse}_h(\boldsymbol{v}_k) = \frac{1}{d^2}\mathbb{E}\left[\|\boldsymbol{v}_k\boldsymbol{v}_k^\top\|_{\rm F}^2 + \|\boldsymbol{w}_h^\star \boldsymbol{w}_h^{\star\top}\|_{\rm F}^2 - 2\langle \boldsymbol{w}_h^\star, \boldsymbol{v}_k\rangle^2\right] \to 2(1 - m_{hk}^2), \tag{67}$$

which tends to 2 for $h \neq \pi(k)$, while $\mathrm{mse}_h(\boldsymbol{0}) = 1$. Therefore, the minimum over permutations in Definition 2.7 is attained, up to constant factors, by pairing each $\boldsymbol{v}_k$ with $\boldsymbol{w}_{\pi(k)}^\star$ and the zero rows with the remaining directions. By repeating computations analogous to the ones in Appendix A.1, one finds, setting $\hat{\boldsymbol{W}} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_r)^\top$,

$$\mathrm{MSE}_\gamma(\hat{\boldsymbol{W}} := (\hat{\boldsymbol{W}}^{\rm sp}, \boldsymbol{0}_{(m_\star - r)\times d})) = \Theta(\mathrm{MMSE}_\gamma^{\rm oracle}), \tag{68}$$

which proves Theorem 3.2 and 3.6.

Figure 4 illustrates the sequential recovery underlying Theorem 3.6. The features are recovered one at a time, each at its own threshold $\alpha_k^{\rm sp}$, where its error $\mathrm{mse}_k$ drops.

## B Proof of Theorem 3.9

In this section we prove Theorem 3.9, stating that a two-layer neural network, trained according to Algorithm 1, can achieve an excess risk bottlenecked by the information-theoretic error on the recovery of the representations $\boldsymbol{W}_\star$. Throughout, we work conditionally on the dataset $\mathcal{D}_1$ used to compute the spectral estimator $\hat{\boldsymbol{W}}^{\rm sp} \in \mathbb{R}^{r\times d}$ of Definition 3.3. To lighten the notation, in this section we denote by $n$ the number of samples in $\mathcal{D}_2$, i.e. $n/2$ in Algorithm 1. The activation $\sigma$ is assumed to be bounded, $|\sigma| \leq B$, continuous and nonconstant. Moreover, we assume that each link function $g_k$, $k \in [m_\star]$, is weakly differentiable with $\mathbb{E}_z[g_k'(z)^2] \leq C'$, for a constant $C'$ independent of $k$. This holds, for instance, for Lipschitz link functions and for polynomials of bounded degree.

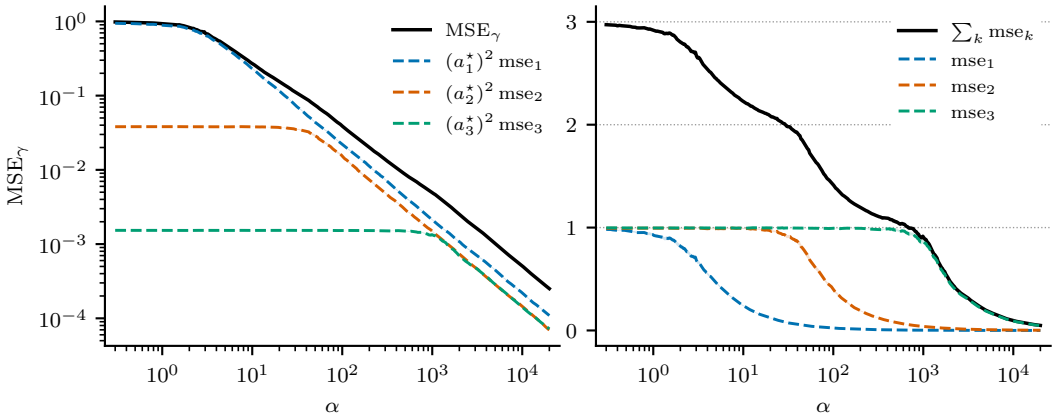



Figure 4: Illustration of the sequential recovery of the features by the spectral estimator of Definition 3.3, with $\mathcal{T}(y) = y/(1+|y|)$. The target is a hierarchical multi-index model with $m_\star = 3$, $g_k(z) = |z| - \sqrt{2/\pi}$ for all $k$, $a^\star \propto (1, 0.2, 0.04)$ with $\sum_k (a_k^\star)^2 = 1$, and $\Delta = 0.5$; $d = 300$. (**left**) $\mathrm{MSE}_\gamma$ and its contributions $(a_k^\star)^2 \mathrm{mse}_k$ (dashed). (**right**) Unweighted error $\sum_k \mathrm{mse}_k$ and the individual $\mathrm{mse}_k$ (dashed); horizontal dotted lines mark the plateau values of $\sum_k \mathrm{mse}_k$ when $j$ features are not yet recovered, $j \in [m_\star]$. Averages over 20 instances; shaded bands represents one standard error.

**Reduction to the projected inputs.** Denote by $\hat{\boldsymbol{w}}_j$ the $j^{\text{th}}$ row of $\hat{\boldsymbol{W}}^{\text{sp}}$. Given a covariate $\boldsymbol{x} \in \mathbb{R}^d$, define the projected inputs $\boldsymbol{s} = \hat{\boldsymbol{W}}^{\text{sp}} \boldsymbol{x} \in \mathbb{R}^r$ and the target indices $\boldsymbol{z}^\star = \boldsymbol{W}_\star \boldsymbol{x} \in \mathbb{R}^{m_\star}$. The first layer pre-activations are given by

$$\boldsymbol{W}\boldsymbol{x} = \boldsymbol{Z}\hat{\boldsymbol{W}}^{\text{sp}}\boldsymbol{x} = \boldsymbol{Z}\boldsymbol{s}, \tag{69}$$

so that the network with second-layer weights $\boldsymbol{c} \in \mathbb{R}^p$, $f(\boldsymbol{x}; \boldsymbol{\Theta}) = \boldsymbol{c}^\top \sigma(\boldsymbol{Z}\boldsymbol{s} + \boldsymbol{b})/\sqrt{p}$, depends on $\boldsymbol{x}$ only through $\boldsymbol{s}$. The rows of $\hat{\boldsymbol{W}}^{\text{sp}}$ are eigenvectors of the symmetric matrix $\boldsymbol{T}$, normalized as $\|\hat{\boldsymbol{w}}_j\|^2 = d$, hence they are orthogonal. Moreover, $\hat{\boldsymbol{W}}^{\text{sp}}$ is independent of $\mathcal{D}_2$ and of the test point. Therefore, for $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, d^{-1}\boldsymbol{I}_d)$, the pair $(\boldsymbol{s}, \boldsymbol{z}^\star)$ is jointly Gaussian with

$$\boldsymbol{s} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r), \qquad \boldsymbol{z}^\star \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{m_\star}), \qquad \mathbb{E}[\boldsymbol{s}\boldsymbol{z}^{\star\top}] = \boldsymbol{M}_d := \frac{1}{d}\hat{\boldsymbol{W}}^{\text{sp}}\boldsymbol{W}_\star^\top \in \mathbb{R}^{r \times m_\star}, \tag{70}$$

and the joint law of $(\boldsymbol{s}, y)$ depends on $d$ only through the overlap matrix $\boldsymbol{M}_d$. Throughout this section, the subscript $d$ denotes a quantity at finite size, which depends on $\mathcal{D}_1$ through $\boldsymbol{M}_d$, and its absence the corresponding deterministic limit as $n, d \to \infty$. In Appendix A.2 we have shown that, as $n, d \to \infty$, each row $j$ of the spectral estimator has asymptotic overlap $m_{\pi(j)}$ with a single signal component $\boldsymbol{w}^\star_{\pi(j)}$, and vanishing overlap with the others, where $\pi : [r] \to [m_\star]$ is injective. Since flipping the sign of $\hat{\boldsymbol{w}}_j$ amounts to flipping the $j$-th column of $\boldsymbol{Z}$, whose law is symmetric, we fix the sign of each $\hat{\boldsymbol{w}}_j$ so that its overlap with $\boldsymbol{w}^\star_{\pi(j)}$ is nonnegative. Then $\boldsymbol{M}_d$ converges in probability to $\boldsymbol{M}$, with $M_{jh} = m_{\pi(j)}\delta_{h\pi(j)}$. We denote by $\mathcal{K} := \pi([r])$ the set of recovered directions. Its size $r$ is the number of outliers of $\boldsymbol{T}$,[3] so that $r \asymp k_\alpha$, with $k_\alpha$ defined in Appendix A.1.

With respect to the projected input space, the effective target function is

$$f^{\text{eff}}_{\star,d}(\boldsymbol{s}) := \mathbb{E}[f_\star(\boldsymbol{x})|\boldsymbol{s}] = \mathbb{E}[y|\boldsymbol{s}]. \tag{71}$$

We denote by $\|h\|^2_{L^2} := \mathbb{E}_{\boldsymbol{s}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_r)}[h(\boldsymbol{s})^2]$ the squared norm of $L^2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r))$. The effective target depends on $d$ only through $\boldsymbol{M}_d$, and we denote by $f^{\text{eff}}_\star$ the effective target computed at the limit, i.e. with $\mathbb{E}[\boldsymbol{s}\boldsymbol{z}^{\star\top}] = \boldsymbol{M}$. Below, we replace $f^{\text{eff}}_{\star,d}$ by $f^{\text{eff}}_\star$, both in the irreducible error and in the choice of the second-layer weights. This is allowed by the following lemma, proven at the end of Appendix B.

**Lemma B.1.** *As $n, d \to \infty$, $\|f^{\text{eff}}_{\star,d} - f^{\text{eff}}_\star\|_{L^2}$ converges to zero in probability.*

[3] More precisely, $r$ equals the number of outliers with probability tending to one as $n, d \to \infty$.

For any network whose weights are independent of the test point, $f(\boldsymbol{x};\boldsymbol{\Theta})$ is a function of $\boldsymbol{s}$, while $f^{\rm eff}_{\star,d}(\boldsymbol{s})$ is the conditional expectation of $f_\star(\boldsymbol{x})$ given $\boldsymbol{s}$. The cross term then vanishes, and the excess risk eq. (10) decomposes as

$$R(\boldsymbol{\Theta}) = \underbrace{\mathbb{E}\left[\left(f_\star(\boldsymbol{x}) - f^{\rm eff}_{\star,d}(\boldsymbol{s})\right)^2\right]}_{=:R_{\star,d}} + \underbrace{\left\|f^{\rm eff}_{\star,d} - p^{-1/2}\boldsymbol{c}^\top\sigma(\boldsymbol{Z}\,\cdot\,+\boldsymbol{b})\right\|^2_{L^2}}_{=:R_{\rm ro}(\boldsymbol{c})} \tag{72}$$

where $\boldsymbol{c}\in\mathbb{R}^p$ denotes the second-layer weights of the network. The first term is the irreducible error due to the imperfect recovery of the representations, and does not depend on the second layer. The second term is the readout error. We now show that the former scales as $\mathrm{MMSE}_\gamma$, and that the latter can be made smaller than the former by taking the width $p$ large enough and the regularization $\lambda$ small enough.

**Irreducible error.** We expand each link function in Hermite polynomials,

$$g_k(z) = \sum_{\ell\geq 0}\frac{g_{k,\ell}}{\ell!}\mathrm{He}_\ell(z), \qquad g_{k,\ell} := \mathbb{E}_z[g_k(z)\mathrm{He}_\ell(z)], \tag{73}$$

where the series converges in $L^2(\mathcal{N}(0,1))$ and $\mathbb{E}_z[g_k(z)^2] = \sum_\ell g_{k,\ell}^2/\ell!$. Since $\mathrm{He}_\ell' = \ell\mathrm{He}_{\ell-1}$, by Stein's lemma $\mathbb{E}_z[g_k'(z)\mathrm{He}_{\ell-1}(z)] = g_{k,\ell}$, hence $\mathbb{E}_z[g_k'(z)^2] = \sum_{\ell\geq 1}\ell\, g_{k,\ell}^2/\ell!$. If $(u,v)$ are standard Gaussian variables with correlation $m$, then $\mathbb{E}[\mathrm{He}_\ell(u)\mathrm{He}_{\ell'}(v)] = \delta_{\ell\ell'}\,\ell!\,m^\ell$ (O'Donnell, 2014). Since $u|v\sim\mathcal{N}(mv, 1-m^2)$, $\mathbb{E}[\mathrm{He}_\ell(u)|v]$ is a polynomial of degree at most $\ell$ in $v$, and the previous identity gives its expansion in Hermite polynomials,

$$\mathbb{E}[\mathrm{He}_\ell(u)|v] = m^\ell\mathrm{He}_\ell(v). \tag{74}$$

At the limit $\boldsymbol{M}$, for $h=\pi(j)\in\mathcal{K}$ the index $z_h^\star$ is correlated with $s_j$ only, with correlation $m_{\pi(j)}$, while for $h\notin\mathcal{K}$ it is independent of $\boldsymbol{s}$. By eq. (74),

$$f^{\rm eff}_\star(\boldsymbol{s}) = \sum_{j=1}^r a^\star_{\pi(j)}\sum_{\ell\geq 0}\frac{g_{\pi(j),\ell}}{\ell!}m^\ell_{\pi(j)}\mathrm{He}_\ell(s_j) + \sum_{h\notin\mathcal{K}}a^\star_h g_{h,0}. \tag{75}$$

As $f^{\rm eff}_{\star,d}$ is a conditional expectation, $R_{\star,d} = \mathbb{E}[f_\star(\boldsymbol{x})^2] - \|f^{\rm eff}_{\star,d}\|^2_{L^2}$. We denote by $R_\star := \mathbb{E}[f_\star(\boldsymbol{x})^2] - \|f^{\rm eff}_\star\|^2_{L^2}$ its value at the limit $\boldsymbol{M}$. Using the independence of the $z_h^\star$, the independence of the $s_j$ and the orthogonality of Hermite polynomials, we find

$$R_\star = \sum_{h\in\mathcal{K}}(a^\star_h)^2\sum_{\ell\geq 1}\frac{g_{h,\ell}^2}{\ell!}\left(1-m_h^{2\ell}\right) + \sum_{h\notin\mathcal{K}}(a^\star_h)^2\sum_{\ell\geq 1}\frac{g_{h,\ell}^2}{\ell!}. \tag{76}$$

For $m\in[0,1]$ and $\ell\geq 1$, $1-m^2\leq 1-m^{2\ell}\leq\ell(1-m^2)$. Moreover, by Assumption 2.3,

$$\sum_{\ell\geq 1} g_{h,\ell}^2/\ell! \geq g_{h,2}^2/2 = \mathbb{E}_z[g_h(z)(z^2-1)]^2/2 > D^2/2$$

and

$$\sum_{\ell\geq 1}\ell\, g_{h,\ell}^2/\ell! = \mathbb{E}_z[g_h'(z)^2] \leq C',$$

which implies

$$\begin{aligned} R_\star &\geq \sum_{h\in\mathcal{K}}(a^\star_h)^2\left(1-m_h^2\right)\sum_{\ell\geq 1}\frac{g_{h,\ell}^2}{\ell!} + \sum_{h\notin\mathcal{K}}(a^\star_h)^2\sum_{\ell\geq 1}\frac{g_{h,\ell}^2}{\ell!} \\ &\geq \frac{D^2}{2}\left(\sum_{h\in\mathcal{K}}(a^\star_h)^2\left(1-m_h^2\right) + \sum_{h\notin\mathcal{K}}(a^\star_h)^2\right) \end{aligned}$$

and, denoting $C'' := \max\{\mathbb{E}_z[g_h(z)^2] : h \in [m_\star]\}$,

$$\begin{aligned} R_\star &\le \sum_{h\in\mathcal{K}} (a_h^\star)^2 \left(1-m_h^2\right) \sum_{\ell\ge 1} \frac{\ell g_{h,\ell}^2}{\ell!} + \sum_{h\notin\mathcal{K}} (a_h^\star)^2 \sum_{\ell\ge 1} \frac{g_{h,\ell}^2}{\ell!} \\ &\le C' \sum_{h\in\mathcal{K}} (a_h^\star)^2 \left(1-m_h^2\right) + C'' \sum_{h\notin\mathcal{K}} (a_h^\star)^2 \\ &\le \max(C', C'') \left( \sum_{h\in\mathcal{K}} (a_h^\star)^2 \left(1-m_h^2\right) + \sum_{h\notin\mathcal{K}} (a_h^\star)^2 \right). \end{aligned}$$

Therefore,

$$R_\star = \Theta\left( \sum_{h\in\mathcal{K}} (a_h^\star)^2 (1-m_h^2) + \sum_{h\notin\mathcal{K}} (a_h^\star)^2 \right) = \Theta(\mathrm{MMSE}_\gamma), \tag{77}$$

where the last equality follows from Theorem 3.6, since, as shown at the end of Appendix A.2, the weighted mean-squared error of the zero-padded spectral estimator is of the same order as the sum in the middle term. As for the spectral method mean-squared error, the two terms correspond to the approximation error of learned features and underfitting of unlearned ones. Finally, $\mathbb{E}[f_\star(\boldsymbol{x})^2]$ does not depend on $\boldsymbol{M}_d$, while $\|f_{\star,d}^{\mathrm{eff}}\|_{L^2}^2$ converges in probability to $\|f_\star^{\mathrm{eff}}\|_{L^2}^2$ by Lemma B.1, so that $R_{\star,d}$ converges in probability to $R_\star$ as $n, d \to \infty$.

**Readout error.** Define the normalized features $\boldsymbol{\phi}(\boldsymbol{s}) := \sigma(\boldsymbol{Z}\boldsymbol{s} + \boldsymbol{b})/\sqrt{p} \in \mathbb{R}^p$, and, for $\boldsymbol{c} \in \mathbb{R}^p$, denote by $f_{\boldsymbol{c}}(\boldsymbol{s}) := \boldsymbol{c}^\top \boldsymbol{\phi}(\boldsymbol{s})$ the predictor with second-layer weights $\boldsymbol{c}$, so that $R_{\mathrm{ro}}(\boldsymbol{c}) = \|f_{\star,d}^{\mathrm{eff}} - f_{\boldsymbol{c}}\|_{L^2}^2$. Since $\|\boldsymbol{\phi}(\boldsymbol{s})\|^2 \le B^2$ for all $\boldsymbol{s}$, independently of $p$, the Cauchy–Schwarz inequality gives

$$f_{\boldsymbol{c}}(\boldsymbol{s})^2 \le B^2 \|\boldsymbol{c}\|^2 \qquad \text{for all } \boldsymbol{s} \in \mathbb{R}^r. \tag{78}$$

Given $\mathcal{D}_2 = \{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$ and $\boldsymbol{s}_i = \hat{\boldsymbol{W}}^{\mathrm{sp}} \boldsymbol{x}_i$, the second layer $\hat{\boldsymbol{a}}$ returned by Algorithm 1 is the minimizer of $\frac{1}{n}\sum_i (y_i - f_{\boldsymbol{c}}(\boldsymbol{s}_i))^2 + \lambda\|\boldsymbol{c}\|^2$ over $\boldsymbol{c} \in \mathbb{R}^p$. Its population counterpart is the minimizer $\boldsymbol{a}_\lambda$ of

$$J(\boldsymbol{c}) := \mathbb{E}\left[(y - f_{\boldsymbol{c}}(\boldsymbol{s}))^2\right] + \lambda\|\boldsymbol{c}\|^2, \tag{79}$$

where the expectation is over a fresh pair $(\boldsymbol{s}, y)$, at fixed $\boldsymbol{Z}, \boldsymbol{b}$. We split the readout error of the trained network into an approximation and an estimation term,

$$\mathbb{E}_{\mathcal{D}_2} R_{\mathrm{ro}}(\hat{\boldsymbol{a}}) \le 2 \underbrace{\|f_{\star,d}^{\mathrm{eff}} - f_{\boldsymbol{a}_\lambda}\|_{L^2}^2}_{\text{approximation}} + 2 \underbrace{\mathbb{E}_{\mathcal{D}_2} \|f_{\boldsymbol{a}_\lambda} - f_{\hat{\boldsymbol{a}}}\|_{L^2}^2}_{\text{estimation}}, \tag{80}$$

and bound the two terms separately.

*Estimation term.* Setting the gradients of the two objectives to zero,

$$\left(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}_p\right) \hat{\boldsymbol{a}} = \hat{\boldsymbol{v}}, \qquad \left(\boldsymbol{\Sigma} + \lambda \boldsymbol{I}_p\right) \boldsymbol{a}_\lambda = \boldsymbol{v}, \tag{81}$$

where

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{s}_i)\boldsymbol{\phi}(\boldsymbol{s}_i)^\top, \qquad \hat{\boldsymbol{v}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{s}_i)\, y_i, \tag{82}$$

$$\boldsymbol{\Sigma} := \mathbb{E}\left[\boldsymbol{\phi}(\boldsymbol{s})\boldsymbol{\phi}(\boldsymbol{s})^\top\right], \qquad \boldsymbol{v} := \mathbb{E}\left[\boldsymbol{\phi}(\boldsymbol{s})\, y\right]. \tag{83}$$

Subtracting $(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}_p)\boldsymbol{a}_\lambda$ from the first identity in (81) and using the second one,

$$\left(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}_p\right)(\hat{\boldsymbol{a}} - \boldsymbol{a}_\lambda) = \hat{\boldsymbol{v}} - \hat{\boldsymbol{\Sigma}}\boldsymbol{a}_\lambda - (\boldsymbol{v} - \boldsymbol{\Sigma}\boldsymbol{a}_\lambda) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{u}_i - \mathbb{E}[\boldsymbol{u}], \tag{84}$$

where $\boldsymbol{u}_i := \boldsymbol{\phi}(\boldsymbol{s}_i)\,(y_i - f_{\boldsymbol{a}_\lambda}(\boldsymbol{s}_i))$ are i.i.d. copies of $\boldsymbol{u} := \boldsymbol{\phi}(\boldsymbol{s})\,(y - f_{\boldsymbol{a}_\lambda}(\boldsymbol{s}))$. Since $\hat{\boldsymbol{\Sigma}}$ is positive semi-definite, $\|(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}_p)^{-1}\|_{\mathrm{op}} \le \lambda^{-1}$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_2}\|f_{\boldsymbol{a}_\lambda} - f_{\hat{\boldsymbol{a}}}\|_{L^2}^2 &\le B^2\,\mathbb{E}_{\mathcal{D}_2}\|\hat{\boldsymbol{a}} - \boldsymbol{a}_\lambda\|^2 && \text{(by eq. (78))}\\
&\le \frac{B^2}{\lambda^2}\,\mathbb{E}_{\mathcal{D}_2}\Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{u}_i - \mathbb{E}[\boldsymbol{u}]\Big\|^2 \\
&\le \frac{B^2}{\lambda^2 n}\,\mathbb{E}\|\boldsymbol{u}\|^2 && \text{(variance of an empirical mean)}\\
&\le \frac{B^4}{\lambda^2 n}\,\mathbb{E}\left[(y - f_{\boldsymbol{a}_\lambda}(\boldsymbol{s}))^2\right] && \text{(since } \|\boldsymbol{\phi}(\boldsymbol{s})\|^2 \le B^2\text{)}\\
&\le \frac{B^4}{\lambda^2 n}\,J(\boldsymbol{a}_\lambda) \le \frac{B^4}{\lambda^2 n}\,J(\boldsymbol{0}) && \text{(since } \boldsymbol{a}_\lambda \text{ minimizes } J\text{).}
\end{aligned}$$

Since $J(\boldsymbol{0}) = \mathbb{E}[y^2] =: C_y = \mathbb{E}[f_\star(\boldsymbol{x})^2] + \Delta$,

$$\mathbb{E}_{\mathcal{D}_2}\|f_{\boldsymbol{a}_\lambda} - f_{\hat{\boldsymbol{a}}}\|_{L^2}^2 \le \frac{B^4 C_y}{\lambda^2 n}, \tag{85}$$

where the right-hand side does not depend on $p$, $r$, $d$, $\boldsymbol{M}_d$, $\boldsymbol{Z}$ and $\boldsymbol{b}$.

*Approximation term.* Recalling the definition of $f^{\mathrm{eff}}_{\star,d}(\boldsymbol{s}) = \mathbb{E}[y|\boldsymbol{s}]$ and the fact that $f_{\boldsymbol{c}}$ depends only on $\boldsymbol{s}$,

$$J(\boldsymbol{c}) = \mathbb{E}[(y - f^{\mathrm{eff}}_{\star,d}(\boldsymbol{s}))^2] + \|f^{\mathrm{eff}}_{\star,d} - f_{\boldsymbol{c}}\|_{L^2}^2 + \lambda\|\boldsymbol{c}\|^2,$$

where the first term does not depend on $\boldsymbol{c}$. As $\boldsymbol{a}_\lambda$ minimizes $J$, for any $\boldsymbol{c} \in \mathbb{R}^p$, possibly depending on $\boldsymbol{Z}$ and $\boldsymbol{b}$, the approximation term can be bounded as follows

$$\|f^{\mathrm{eff}}_{\star,d} - f_{\boldsymbol{a}_\lambda}\|_{L^2}^2 \le \|f^{\mathrm{eff}}_{\star,d} - f_{\boldsymbol{c}}\|_{L^2}^2 + \lambda\|\boldsymbol{c}\|^2. \tag{86}$$

We now construct second-layer weights $\boldsymbol{c}$ for which the right-hand side is small enough to produce the result. The pairs $(\boldsymbol{z}_j, b_j)$, formed by the rows of $\boldsymbol{Z}$ and the biases, are i.i.d. copies of $(\boldsymbol{z}, b)$, with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, r^{-1}\boldsymbol{I}_r)$ and $b \sim \mathcal{N}(0,1)$ independent. We denote by $L^2_{\boldsymbol{z},b}$ the space of functions $w : \mathbb{R}^{r+1} \to \mathbb{R}$ with $\|w\|^2_{L^2_{\boldsymbol{z},b}} := \mathbb{E}_{\boldsymbol{z},b}[w(\boldsymbol{z},b)^2] < \infty$. For $w \in L^2_{\boldsymbol{z},b}$, define the infinite-width predictor

$$f^\infty_w(\boldsymbol{s}) := \mathbb{E}_{\boldsymbol{z},b}\left[w(\boldsymbol{z},b)\,\sigma(\boldsymbol{z}^\top \boldsymbol{s} + b)\right], \tag{87}$$

which satisfies $|f^\infty_w(\boldsymbol{s})| \le B\|w\|_{L^2_{\boldsymbol{z},b}}$. We construct $\boldsymbol{c}$ in two steps: we first approximate $f^{\mathrm{eff}}_\star$ by an infinite-width predictor $f^\infty_w$, and then approximate $f^\infty_w$ by a finite-width network with the random features $(\boldsymbol{z}_j, b_j)$, with weights of controlled norm. The first step relies on the following density result, which adapts the argument of Hornik (1991).

**Lemma B.2.** *Let $\sigma$ be bounded, continuous and nonconstant. Then the set $\{f^\infty_w : w \in L^2_{\boldsymbol{z},b}\}$ is dense in $L^2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r))$.*

The proof is given at the end of Appendix B. For the second step, given $w \in L^2_{\boldsymbol{z},b}$, let $\boldsymbol{c} \in \mathbb{R}^p$ with $c_j := w(\boldsymbol{z}_j, b_j)/\sqrt{p}$. Since the pairs $(\boldsymbol{z}_j, b_j)$ are i.i.d. copies of $(\boldsymbol{z}, b)$ and $\mathbb{E}_{\boldsymbol{z},b}[w(\boldsymbol{z},b)\sigma(\boldsymbol{z}^\top\boldsymbol{s}+b)] = f^\infty_w(\boldsymbol{s})$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z},\boldsymbol{b}}\|f^\infty_w - f_{\boldsymbol{c}}\|^2_{L^2} &= \mathbb{E}_{\boldsymbol{s}}\,\mathrm{Var}_{\boldsymbol{Z},\boldsymbol{b}}\Big[\frac{1}{p}\sum_{j=1}^p w(\boldsymbol{z}_j, b_j)\sigma(\boldsymbol{z}_j^\top \boldsymbol{s} + b_j)\Big] \\
&= \frac{1}{p}\,\mathbb{E}_{\boldsymbol{s}}\,\mathrm{Var}_{\boldsymbol{z},b}\left[w(\boldsymbol{z},b)\sigma(\boldsymbol{z}^\top\boldsymbol{s}+b)\right] \\
&\le \frac{1}{p}\,\mathbb{E}_{\boldsymbol{s}}\mathbb{E}_{\boldsymbol{z},b}\left[w(\boldsymbol{z},b)^2\sigma(\boldsymbol{z}^\top\boldsymbol{s}+b)^2\right] \le \frac{B^2}{p}\mathbb{E}_{\boldsymbol{z},b}[w(\boldsymbol{z},b)^2].
\end{aligned}$$

Therefore,

$$\mathbb{E}_{\boldsymbol{Z},\boldsymbol{b}}\|f^\infty_w - f_{\boldsymbol{c}}\|^2_{L^2} \le \frac{B^2\|w\|^2_{L^2_{\boldsymbol{z},b}}}{p}, \qquad \mathbb{E}_{\boldsymbol{Z},\boldsymbol{b}}\|\boldsymbol{c}\|^2 = \frac{1}{p}\sum_{j=1}^p \mathbb{E}_{\boldsymbol{z},b}[w(\boldsymbol{z}_j,b_j)^2] = \|w\|^2_{L^2_{\boldsymbol{z},b}}. \tag{88}$$

Fix $\epsilon > 0$. By Lemma B.2, there exists $w_\epsilon \in L^2_{\boldsymbol{z},b}$ such that $\|f^{\rm eff}_\star - f^\infty_{w_\epsilon}\|_{L^2} \le \epsilon$, and we take $\boldsymbol{c}$ as above with $w = w_\epsilon$. Then, averaging eq. (86) over $\boldsymbol{Z}, \boldsymbol{b}$ and using twice $(a+b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{b}}\|f^{\rm eff}_{\star,d} - f_{\boldsymbol{a}_\lambda}\|^2_{L^2} &\le 2\|f^{\rm eff}_{\star,d} - f^{\rm eff}_\star\|^2_{L^2} + 4\|f^{\rm eff}_\star - f^\infty_{w_\epsilon}\|^2_{L^2} \\ &\quad + 4\,\mathbb{E}_{\boldsymbol{Z},\boldsymbol{b}}\|f^\infty_{w_\epsilon} - f_{\boldsymbol{c}}\|^2_{L^2} + \lambda\,\mathbb{E}_{\boldsymbol{Z},\boldsymbol{b}}\|\boldsymbol{c}\|^2 \\ &\le 2\|f^{\rm eff}_{\star,d} - f^{\rm eff}_\star\|^2_{L^2} + 4\epsilon^2 + \left(\frac{4B^2}{p} + \lambda\right)\|w_\epsilon\|^2_{L^2_{\boldsymbol{z},b}} \qquad \text{(by eq. (88))}.\end{aligned}$$

**Conclusion.** Inserting the two bounds in eq. (80), the readout error of the trained network satisfies

$$\mathbb{E}_{\mathcal{D}_2,\boldsymbol{Z},\boldsymbol{b}} R_{\rm ro}(\hat{\boldsymbol{a}}) \le \frac{2B^4 C_y}{\lambda^2 n} + 4\|f^{\rm eff}_{\star,d} - f^{\rm eff}_\star\|^2_{L^2} + 8\epsilon^2 + \left(\frac{8B^2}{p} + 2\lambda\right)\|w_\epsilon\|^2_{L^2_{\boldsymbol{z},b}}, \tag{89}$$

for every $n$, $p$ and $\lambda > 0$. Note that the dependence on $r$ enters only through $w_\epsilon$, and $r$ is fixed as $n, d \to \infty$. At fixed $\lambda$, as $n, d \to \infty$, the first term vanishes, whatever the dependence of $p$ on $n$, and the second converges to zero in probability, by Lemma B.1. For $p \ge 8B^2\|w_\epsilon\|^2_{L^2_{\boldsymbol{z},b}}/\epsilon^2$ and $\lambda \le \epsilon^2/(2\|w_\epsilon\|^2_{L^2_{\boldsymbol{z},b}})$, the expected readout error is then at most $10\epsilon^2 + o_{\mathbb{P}}(1)$. Take $\epsilon$ such that $\epsilon^2 = R_\star/10$. By eq. (72), for every $\alpha$ there exist a finite width $p(\alpha)$ and a regularization $\lambda(\alpha) > 0$ such that, for $p \ge p(\alpha)$ and $\lambda \le \lambda(\alpha)$,

$$R_\star + o_{\mathbb{P}}(1) \le R_{\star,d} + \mathbb{E}_{\mathcal{D}_2,\boldsymbol{Z},\boldsymbol{b}} R_{\rm ro}(\hat{\boldsymbol{a}}) \le 2R_\star + o_{\mathbb{P}}(1), \tag{90}$$

with no condition relating $p$ and $n$. Taking the expectation over $\mathcal{D}_1$, $R_\star + o(1) \le R_{\rm NN} \le 2R_\star + o(1)$. Since $R_\star = \Theta(\mathrm{MMSE}_\gamma)$, the readout introduces no additional bottleneck, and $R_{\rm NN} = \Theta(\mathrm{MMSE}_\gamma)$.

Moreover, eq. (89) shows that the readout error vanishes if the width grows and the regularization decreases slowly enough. If $p_n \to \infty$ and $\lambda_n \to 0$ with $n\lambda_n^2 \to \infty$, the first and last terms vanish for every $\epsilon$, so that the expected readout error is at most $8\epsilon^2 + o_{\mathbb{P}}(1)$ for any $\epsilon > 0$, and $R_{\rm NN}$ converges to $R_\star$.

**Proofs of the Lemmas B.1 and B.2.**

*Proof of Lemma B.1.* For $\boldsymbol{Q} \in \mathbb{R}^{r\times m_\star}$ with $\|\boldsymbol{Q}\|_{\rm op} \le 1$, denote by $f^{\rm eff}_\star(\cdot;\boldsymbol{Q})$ the effective target computed with $\mathbb{E}[\boldsymbol{s}\boldsymbol{z}^{\star\top}] = \boldsymbol{Q}$, so that $f^{\rm eff}_{\star,d} = f^{\rm eff}_\star(\cdot;\boldsymbol{M}_d)$ and $f^{\rm eff}_\star = f^{\rm eff}_\star(\cdot;\boldsymbol{M})$. Since $\boldsymbol{M}_d$ converges in probability to $\boldsymbol{M}$, by the continuous mapping theorem it suffices to show that the map $\boldsymbol{Q} \mapsto f^{\rm eff}_\star(\cdot;\boldsymbol{Q})$ is continuous in $L^2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r))$.

By Assumption 2.3, $f_\star(\boldsymbol{x}) = \sum_k a^\star_k g_k(z^\star_k)$ is a square-integrable function of $\boldsymbol{z}^\star \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{m_\star})$. Fix $\epsilon > 0$. Since polynomials are dense in $L^2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{m_\star}))$, there exists a polynomial $q_\epsilon$ with $\mathbb{E}[(f_\star - q_\epsilon)^2] \le \epsilon^2$. We denote by $q^{\rm eff}_\epsilon(\boldsymbol{s};\boldsymbol{Q}) := \mathbb{E}[q_\epsilon(\boldsymbol{z}^\star)|\boldsymbol{s}]$ its conditional expectation computed with $\mathbb{E}[\boldsymbol{s}\boldsymbol{z}^{\star\top}] = \boldsymbol{Q}$.

Since the law of $\boldsymbol{z}^\star$ does not depend on $\boldsymbol{Q}$, by Jensen's inequality

$$\|f^{\rm eff}_\star(\cdot;\boldsymbol{Q}) - q^{\rm eff}_\epsilon(\cdot;\boldsymbol{Q})\|_{L^2} \le \epsilon \qquad \text{for all } \boldsymbol{Q}. \tag{91}$$

Then, since $\boldsymbol{z}^\star|\boldsymbol{s} \sim \mathcal{N}(\boldsymbol{Q}^\top\boldsymbol{s}, \boldsymbol{I}_{m_\star} - \boldsymbol{Q}^\top\boldsymbol{Q})$ and the moments of a Gaussian vector are polynomials in its mean and covariance, $q^{\rm eff}_\epsilon(\boldsymbol{s};\boldsymbol{Q})$ is a polynomial in $\boldsymbol{s}$ whose coefficients are polynomials in the entries of $\boldsymbol{Q}$. As $\boldsymbol{s}$ has finite moments of all orders, the map $\boldsymbol{Q} \mapsto q^{\rm eff}_\epsilon(\cdot;\boldsymbol{Q})$ is continuous in $L^2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r))$. Fix $\boldsymbol{Q}$. There exists $\delta > 0$ such that

$$\|q^{\rm eff}_\epsilon(\cdot;\boldsymbol{Q}') - q^{\rm eff}_\epsilon(\cdot;\boldsymbol{Q})\|_{L^2} \le \epsilon \qquad \text{for all } \boldsymbol{Q}' \text{ with } \|\boldsymbol{Q}' - \boldsymbol{Q}\|_{\rm op} \le \delta. \tag{92}$$

Therefore, for all such $\boldsymbol{Q}'$,

$$\begin{aligned}\|f^{\rm eff}_\star(\cdot;\boldsymbol{Q}') - f^{\rm eff}_\star(\cdot;\boldsymbol{Q})\|_{L^2} &\le \|f^{\rm eff}_\star(\cdot;\boldsymbol{Q}') - q^{\rm eff}_\epsilon(\cdot;\boldsymbol{Q}')\|_{L^2} + \|q^{\rm eff}_\epsilon(\cdot;\boldsymbol{Q}') - q^{\rm eff}_\epsilon(\cdot;\boldsymbol{Q})\|_{L^2} \\ &\quad + \|q^{\rm eff}_\epsilon(\cdot;\boldsymbol{Q}) - f^{\rm eff}_\star(\cdot;\boldsymbol{Q})\|_{L^2} && \text{(triangle inequality)} \\ &\le 3\epsilon && \text{(by eqs. (91) and (92))}.\end{aligned}$$

As $\epsilon > 0$ is arbitrary, the map $\boldsymbol{Q} \mapsto f^{\rm eff}_\star(\cdot;\boldsymbol{Q})$ is continuous at $\boldsymbol{Q}$. □

*Proof of Lemma B.2.* Since $w \mapsto f_w^\infty$ is linear, the set $\{f_w^\infty : w \in L^2_{\boldsymbol{z},b}\}$ is a linear subspace of $L^2(\mathcal{N}(\mathbf{0}, \boldsymbol{I}_r))$, and it is dense if and only if its orthogonal complement is $\{0\}$. Let $f \in L^2(\mathcal{N}(\mathbf{0}, \boldsymbol{I}_r))$ be such that $\mathbb{E}_{\boldsymbol{s}}[f(\boldsymbol{s}) f_w^\infty(\boldsymbol{s})] = 0$ for all $w \in L^2_{\boldsymbol{z},b}$. We show that $f = 0$. Define

$$\psi_f(\boldsymbol{z}, b) := \mathbb{E}_{\boldsymbol{s}}\left[f(\boldsymbol{s})\, \sigma(\boldsymbol{z}^\top \boldsymbol{s} + b)\right], \qquad (\boldsymbol{z}, b) \in \mathbb{R}^{r+1}. \tag{93}$$

We use two properties of $\psi_f$. First, since $|\sigma| \le B$, the Cauchy–Schwarz inequality gives $|\psi_f(\boldsymbol{z}, b)| \le B\|f\|_{L^2}$, so that $\psi_f \in L^2_{\boldsymbol{z},b}$. Second, $\psi_f$ is continuous on $\mathbb{R}^{r+1}$, by dominated convergence, since $\sigma$ is continuous and $|f(\boldsymbol{s})\sigma(\boldsymbol{z}^\top \boldsymbol{s} + b)| \le B|f(\boldsymbol{s})|$, which is integrable. For any $w \in L^2_{\boldsymbol{z},b}$,

$$\begin{aligned} 0 = \mathbb{E}_{\boldsymbol{s}}\left[f(\boldsymbol{s}) f_w^\infty(\boldsymbol{s})\right] &= \mathbb{E}_{\boldsymbol{s}}\mathbb{E}_{\boldsymbol{z},b}\left[f(\boldsymbol{s}) w(\boldsymbol{z}, b)\sigma(\boldsymbol{z}^\top \boldsymbol{s} + b)\right] \\ &= \mathbb{E}_{\boldsymbol{z},b}\left[w(\boldsymbol{z}, b)\psi_f(\boldsymbol{z}, b)\right], \end{aligned}$$

where the last step follows from Fubini's theorem, as $|f(\boldsymbol{s})w(\boldsymbol{z}, b)\sigma(\boldsymbol{z}^\top \boldsymbol{s} + b)| \le B|f(\boldsymbol{s})||w(\boldsymbol{z}, b)|$. Choosing $w = \psi_f$ gives $\mathbb{E}_{\boldsymbol{z},b}[\psi_f(\boldsymbol{z}, b)^2] = 0$, and, since $\psi_f$ is continuous and the law of $(\boldsymbol{z}, b)$ has a positive density on $\mathbb{R}^{r+1}$, $\psi_f(\boldsymbol{z}, b) = 0$ for all $(\boldsymbol{z}, b) \in \mathbb{R}^{r+1}$. In other words, $f$ is orthogonal to all the functions $\boldsymbol{s} \mapsto \sigma(\boldsymbol{z}^\top \boldsymbol{s} + b)$, hence to their linear span. Since $\sigma$ is bounded and nonconstant, Theorem 1 in Hornik (1991) ensures that this span is dense in $L^2(\mu)$ for any finite measure $\mu$ on $\mathbb{R}^r$, in particular for $\mu = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_r)$. Therefore, $f$ is orthogonal to the whole space $L^2(\mathcal{N}(\mathbf{0}, \boldsymbol{I}_r))$, i.e., $f = 0$. □

# C Proof of Theorem 3.13

In order to prove Theorem 3.13, we consider the following matching upper and lower bounds for the optimal computational thresholds $\alpha_k^{\mathrm{AMP}}$ and the weighted mean-squared error in Definition 2.7.

## C.1 Lower bound

In this section, we make use of the Bayes-AMP algorithm (Aubin et al., 2018; Troiani et al., 2025). Consider the modified dataset $\mathcal{D}'_k = \{(\boldsymbol{x}_i, y'_i)\}_{i\in[n]}$, with single-index labels $y'_i = k^{-\gamma} g(\langle \boldsymbol{w}_k^\star, \boldsymbol{x}_i\rangle) + \sqrt{\Delta}\xi_i$ and the problem of learning $\boldsymbol{w}_k^\star$ from $\mathcal{D}'_k$. In Celentano et al. (2020); Montanari and Wu (2024) it is shown that the Bayes-AMP algorithm (Algorithm 1) for this single-index problem is provably optimal among first order algorithms, leading to the smallest mean-squared error. Consider now a second algorithm, which, knowing $\{\boldsymbol{w}_h\}_{h\neq k}$, maps $\mathcal{D}'_k$ to the original dataset $\mathcal{D}$ with multi-index labels and applies the Bayes-AMP for multi-index models, in order to compute $\hat{\boldsymbol{W}}^{\mathrm{AMP}}$ as an estimator for $\boldsymbol{W}^\star$ (Algorithm 2). Note that this procedure is equivalent to applying Bayes-AMP to the original dataset. Concerning merely the estimation of $\boldsymbol{w}_k^\star$, the Algorithm 2 is sub-optimal with respect to Algorithm 1, as the information on $\boldsymbol{w}_k^\star$ is equivalent for both procedures. Therefore, the critical threshold and mean-squared error of Algorithm 1 are lower-bounds of the ones of Algorithm 2 in the optimal computational weak recovery of the $k^{\mathrm{th}}$ feature. Note that dataset $\mathcal{D}'_k$ corresponds to the single-index problem studied in Defilippis et al. (2026b), with SNR $\lambda = k^{-2\gamma}$. In particular, by Theorem 1 in Defilippis et al. (2026b), for $k$ large enough, $\alpha_k^{\mathrm{AMP}}$ is lower bounded by

$$\alpha_k^{\mathrm{SIM}} = \Theta(k^{2\gamma\beta_\star}). \tag{94}$$

Define the lower bound estimator as $\boldsymbol{W}^{\mathrm{low}} \in \mathbb{R}^{m_\star \times d}$ having each row $k$ obtained from Bayes-AMP applied to dataset $\mathcal{D}'_k$. For $\alpha$ large enough, the estimator $\hat{\boldsymbol{W}}^{\mathrm{low}}$ has weakly recovered the first $\hat{k}(\alpha) = \Theta(\alpha^{1/(2\gamma\beta_\star)})$ (if $\alpha \ll {m_\star}^{2\gamma\beta_\star}$) and $\hat{k}(\alpha) = m_\star$ (if $\alpha \gg m_\star^{2\gamma\beta_\star}$) features, with an asymptotic overlap $m_k^{\mathrm{low}}$. We can further lower bound the correspondent mean-squared error by considering, for each $k \in [\hat{k}(\alpha)]$, its statistical bottleneck, established in Corollary D.4. Consequently, the asymptotic error for the recovered features satisfies $1 - (m_k^{\mathrm{low}})^2 \ge Ck^{2\gamma}/\alpha$ for $k \in [\hat{k}(\alpha)]$ and a

certain constant $C$. For $\alpha \ll m_\star^{2\gamma\beta_\star}$, the weighted mean-squared error is given by:

$$\begin{aligned}
\mathrm{MSE}_\gamma(\hat{\boldsymbol{W}}^{\mathrm{low}}) &= \sum_{k=1}^{\hat{k}(\alpha)} k^{-2\gamma}(1-(m_k^{\mathrm{low}})^2) + \sum_{k=\hat{k}(\alpha)+1}^{m_\star} k^{-2\gamma} \\
&\geq C\sum_{k=1}^{\hat{k}(\alpha)} k^{-2\gamma}\frac{k^{2\gamma}}{\alpha} + \Theta\left(\alpha^{(1-2\gamma)/(2\gamma\beta_\star)}\right) \\
&= C\frac{1}{\alpha}\sum_{k=1}^{\hat{k}(\alpha)} 1 + \Theta\left(\alpha^{(1-2\gamma)/(2\gamma\beta_\star)}\right) \\
&= \Theta\left(\alpha^{-1+1/(2\gamma\beta_\star)}\right) + \Theta\left(\alpha^{-1/\beta_\star(1-1/(2\gamma))}\right).
\end{aligned}$$

Instead, if $\alpha \gg m_\star^{2\gamma\beta_\star}$,

$$\begin{aligned}
\mathrm{MSE}_\gamma(\hat{\boldsymbol{W}}^{\mathrm{low}}) &= \sum_{k=1}^{m_\star} k^{-2\gamma}(1-(m_k^{\mathrm{low}})^2) \\
&\geq C\sum_{k=1}^{m_\star} k^{-2\gamma}\frac{k^{2\gamma}}{\alpha} \\
&= \Theta(m_\star/\alpha).
\end{aligned}$$

### C.2 Upper bound

An upper bound can be obtained as a direct adaptation of the proof in Appendix A.2 to the case of generic noise sensitivity exponent $\beta_\star \geq 1$.

Note that the assumptions of Theorem 3.13 do not fix the sign of the leading term in the first line of eq. (60), nor of its analogue for $\beta_\star > 1$, eq. (95) below. The informative outliers can therefore emerge from either edge of the bulk (see Remark A.3). Here we consider the spectral method applied to both edges of the spectrum. Denoting by $\hat{\boldsymbol{W}}_+^{\mathrm{sp}} \in \mathbb{R}^{r_+\times d}$ and $\hat{\boldsymbol{W}}_-^{\mathrm{sp}} \in \mathbb{R}^{r_-\times d}$ the spectral estimators of Definition 3.3 computed from $\boldsymbol{T}$ and from $-\boldsymbol{T}$, respectively, we set $\hat{\boldsymbol{W}}^{\mathrm{sp}} := (\hat{\boldsymbol{W}}_+^{\mathrm{sp}}; \hat{\boldsymbol{W}}_-^{\mathrm{sp}}) \in \mathbb{R}^{(r_++r_-)\times d}$. Its rows are the eigenvectors of $\boldsymbol{T}$ associated with its $r_+$ largest and $r_-$ smallest eigenvalues that are separated from the bulk, selected by the criterion of Definition 3.3. Since $r_+ + r_- \leq m_\star$, we evaluate $\mathrm{MSE}_\gamma$ on the zero-padded estimator $(\hat{\boldsymbol{W}}^{\mathrm{sp}}, \boldsymbol{0}_{(m_\star-r_+-r_-)\times d})$, as in Theorem 3.6.

Following the same steps as in Appendix A.2, one finds that, for any $h : \mathbb{R} \to \mathbb{R}$ bounded, there exists a constant $C$ such that, defining $G_k(y) = \mathbb{E}[z_k^2 - 1|Y = y]$, $Y \sim \mathcal{N}(\sum_{k=1}^{m_\star} k^{-\gamma} g(z_k),\, \Delta)$, $z_k \sim \mathcal{N}(0,1)$ for $k \in [m_\star]$,

$$\mathbb{E}[h(Y)G_k(Y)] = Ck^{-\gamma\beta_\star}\mathbb{E}_z[g_k^{\beta_\star}(z)(z^2-1)] + O_k(k^{-\gamma(\beta_\star+1)}). \tag{95}$$

Using this result, Theorem 4.1 and 4.2 in Kovačević et al. (2025) imply that $\frac{1}{d}|\langle \hat{\boldsymbol{w}}_k^{\mathrm{sp}}, \boldsymbol{w}_h^\star\rangle| = m_k^{\mathrm{sp}}\delta_{kh}$ and that $m_k^{\mathrm{sp}} > 0$ if and only if $\alpha > \alpha_k^{\mathrm{sp}}$ where, for $k$ large enough, $\alpha_k^{\mathrm{sp}}$ is the solution of the following system of equations in $(\alpha_k^{\mathrm{sp}}, t)$,

$$\begin{cases} (\alpha_k^{\mathrm{sp}})^{-1} = \Theta(t^{-1}k^{-\gamma\beta_\star}) \\ (\alpha_k^{\mathrm{sp}})^{-1} = \Theta(t^{-2}) \end{cases} \quad \Longrightarrow \quad \begin{cases} \alpha_k^{\mathrm{sp}} = \Theta(k^{2\gamma\beta_\star}) \\ t = \Theta(k^{\gamma\beta_\star}). \end{cases} \tag{96}$$

Using Theorem 4.2 in Kovačević et al. (2025), we can also derive the leading order behavior of $m_k^{\mathrm{sp}}$. Consider the auxiliary function $\zeta_\alpha(t)$ eq. (33)

$$\zeta_\alpha(t) = t\left(1 + \alpha\mathbb{E}_Y\left[\frac{\mathcal{T}(Y)}{t-\mathcal{T}(Y)}\right]\right) \tag{97}$$

Then, for $t \gg 1$

$$\zeta'_\alpha(t) = 1 + \alpha \mathbb{E}_Y \left[ \frac{\mathcal{T}(Y)}{t - \mathcal{T}(Y)} \right] - \alpha t \mathbb{E}_Y \left[ \frac{\mathcal{T}(Y)}{(t - \mathcal{T}(Y))^2} \right] \tag{98}$$

$$= 1 - \alpha \mathbb{E}_Y \left[ \frac{\mathcal{T}^2(Y)}{(t - \mathcal{T}(Y))^2} \right] \tag{99}$$

$$= 1 - \Theta(\alpha t^{-2}). \tag{100}$$

Further, defining $R_k(t) = \alpha t \, \mathbb{E}[z_k^2 \mathcal{T}(Y)/(t - \mathcal{T}(Y))]$, for $t \gg 1$,

$$\begin{aligned} R'_k(t) &= -\alpha \mathbb{E} \left[ \frac{(z_k^2 - 1)\mathcal{T}^2(Y)}{(t - \mathcal{T}(Y))^2} \right] - \alpha \mathbb{E} \left[ \frac{\mathcal{T}^2(Y)}{(t - \mathcal{T}(Y))^2} \right] \\ &= -\Theta(\alpha k^{-\beta_\star \gamma} t^{-2} + \alpha t^{-2}) \qquad \text{(by eq. (95))} \\ &= -\Theta(\alpha t^{-2}) \end{aligned}$$

Then, define $\hat{t}$ as the solution of $\alpha \mathbb{E}[G_k(Y)\mathcal{T}(Y)/(t - \mathcal{T}(Y))] = 1$, which, for $\alpha \gg 1$, $k$ large enough, using eq. (95), is given by $\hat{t} = \Theta(\alpha k^{-\gamma\beta_\star})$. By Theorem 4.2 in Kovačević et al. (2025) applied to hierarchical models

$$\begin{aligned} (m_k^{\mathrm{sp}})^2 &= \frac{\zeta'_\alpha(\hat{t})}{\zeta'_\alpha(\hat{t}) - R'_k(\hat{t})} \\ &= \frac{1 - \Theta(k^{2\gamma\beta_\star}/\alpha)}{1 - \Theta(k^{2\gamma\beta_\star}/\alpha) + \Theta(k^{2\gamma\beta_\star}/\alpha)} \\ &= 1 - \Theta(k^{2\gamma\beta_\star}/\alpha). \end{aligned}$$

As a consequence, denoting by $\hat{k}(\alpha)$ the number of weakly recovered features at sample complexity $\alpha \gg 1$, where $\hat{k}(\alpha) = \Theta(\alpha^{1/(2\gamma\beta_\star)})$ if $\alpha \ll {m_\star}^{2\gamma\beta_\star}$ and $\hat{k}(\alpha) = m_\star$ if $\alpha \gg {m_\star}^{2\gamma\beta_\star}$, the weighted mean-squared error satisfies, for $\alpha \ll {m_\star}^{2\gamma\beta_\star}$

$$\begin{aligned} \mathrm{MSE}_\gamma(\hat{\boldsymbol{W}}^{\mathrm{sp}}) &= \sum_{k=1}^{\hat{k}(\alpha)} k^{-2\gamma}(1 - (m_k^{\mathrm{sp}})^2) + \sum_{k=\hat{k}(\alpha)+1}^{m_\star} k^{-2\gamma} \\ &= \sum_{k=1}^{\hat{k}(\alpha)} k^{-2\gamma} \Theta\left(\frac{k^{2\gamma\beta_\star}}{\alpha}\right) + \Theta\left(\alpha^{(1-2\gamma)/(2\gamma\beta_\star)}\right) \\ &= \Theta\left(\alpha^{(1-2\gamma)/(2\gamma\beta_\star)}\right); \end{aligned}$$

while, for $\alpha \gg {m_\star}^{2\gamma\beta_\star}$

$$\begin{aligned} \mathrm{MSE}_\gamma(\hat{\boldsymbol{W}}^{\mathrm{sp}}) &= \sum_{k=1}^{m_\star} k^{-2\gamma}(1 - (m_k^{\mathrm{sp}})^2) \\ &= \sum_{k=1}^{m_\star} k^{-2\gamma} \Theta\left(\frac{k^{2\gamma\beta_\star}}{\alpha}\right) \\ &= \Theta({m_\star}^{1+2\gamma(\beta_\star - 1)}/\alpha). \end{aligned}$$

## D Useful results on single-index models

In this section we present various results applied in the derivation of the lower bound to $\mathrm{MMSE}_\gamma$ in Appendix A.1. In particular, we focus on the following single-index model setting.

**Definition D.1.** *Let $\boldsymbol{w}^\star \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$, $g : \mathbb{R} \to \mathbb{R}$ satisfy Assumption 2.3. 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*

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d/d), \qquad y_i = \sqrt{\lambda} g(\langle \boldsymbol{x}_i, \boldsymbol{w}^\star \rangle) + \xi_i, \tag{101}$$

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

Since $\|\boldsymbol{w}^\star\|^2/d \to 1$ almost surely, in the high-dimensional limit this Gaussian prior is equivalent to the uniform prior on the sphere of radius $\sqrt{d}$ followed by the rows of $\boldsymbol{W}_\star$, and the results below apply to both. 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:

$$\mathrm{mse}(\boldsymbol{w}) := \frac{1}{d^2}\mathbb{E}[\|\boldsymbol{w}\boldsymbol{w}^\top - \boldsymbol{w}^\star\boldsymbol{w}^{\star\top}\|_{\mathrm{F}}^2], \tag{102}$$

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

$$\mathrm{mmse} := \frac{1}{d^2}\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top|\mathcal{D}] - \boldsymbol{w}^\star\boldsymbol{w}^{\star\top}\|_{\mathrm{F}}^2\right] = \min_{\hat{\boldsymbol{Q}}} \frac{1}{d^2}\mathbb{E}\left[\|\hat{\boldsymbol{Q}}(\mathcal{D}) - \boldsymbol{w}^\star\boldsymbol{w}^{\star\top}\|_{\mathrm{F}}^2\right]. \tag{103}$$

Intuitively, such optimal estimation error decreases with the SNR.

**Lemma D.2.** *In the setting of Definition D.1, the optimal matrix-MSE is a non-increasing function of the SNR $\lambda$.*

*Proof.* Let $0 \le \lambda' < \lambda$ and let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$ be generated at SNR $\lambda$. Define

$$y_i' = \sqrt{\lambda'/\lambda}\, y_i + \sqrt{1-\lambda'/\lambda}\,\tilde{\xi}_i, \qquad \tilde{\xi}_i \sim \mathcal{N}(0,1) \text{ i.i.d., independent of everything else.}$$

Then $y_i' = \sqrt{\lambda'}g(\langle \boldsymbol{x}_i, \boldsymbol{w}^\star\rangle) + \xi_i'$ with $\xi_i' \sim \mathcal{N}(0,1)$ i.i.d. and independent of $(\boldsymbol{X}, \boldsymbol{w}^\star)$, so that $\mathcal{D}' = \{(\boldsymbol{x}_i, y_i')\}_{i\in[n]}$ has the same joint law with $\boldsymbol{w}^\star$ as a dataset generated at SNR $\lambda'$. Since $\tilde{\boldsymbol{\xi}}$ is independent of $(\mathcal{D}, \boldsymbol{w}^\star)$, we have $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top|\mathcal{D}, \tilde{\boldsymbol{\xi}}] = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top|\mathcal{D}]$, and since $\mathcal{D}'$ is a function of $(\mathcal{D}, \tilde{\boldsymbol{\xi}})$, the optimality of the conditional expectation gives

$$\mathrm{mmse}(\lambda) = \frac{1}{d^2}\mathbb{E}\|\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top|\mathcal{D}, \tilde{\boldsymbol{\xi}}] - \boldsymbol{w}^\star\boldsymbol{w}^{\star\top}\|_{\mathrm{F}}^2 \le \frac{1}{d^2}\mathbb{E}\|\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top|\mathcal{D}'] - \boldsymbol{w}^\star\boldsymbol{w}^{\star\top}\|_{\mathrm{F}}^2 = \mathrm{mmse}(\lambda').$$

□

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 Barbier et al. (2019) to the setting of interest.

**Theorem D.3** (Barbier et al., 2019)**.** *Consider the setting defined in D.1. Then, in the limit $n, d \to \infty$, with fixed ratio $n/d = \alpha$,*

$$\frac{1}{d^2}\mathbb{E}\|\boldsymbol{w}^\star\boldsymbol{w}^{\star\top} - \mathbb{E}[\boldsymbol{w}\boldsymbol{w}\top \mid \mathcal{D}]\|_{\mathrm{F}}^2 \to 1 - m^2, \tag{104}$$

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

$$\frac{1}{d}|\boldsymbol{w}^\top\boldsymbol{w}^\star| \xrightarrow{\mathbb{P}} m, \tag{105}$$

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

$$\sup_{m\in[0,1]} f_{\mathrm{RS}}(m), \quad f_{\mathrm{RS}}(m) := \{m + \log(1-m) + 2\alpha\Psi_{\mathrm{out}}(m)\}, \tag{106}$$

$$\Psi_{\mathrm{out}}(m;\lambda) := \mathbb{E}_{W,V,Y}\log\mathbb{E}_{w\sim\mathcal{N}(0,1)}\left[\mathsf{P}(Y|\sqrt{m}V + \sqrt{1-m}w)\right], \tag{107}$$

*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 $\alpha^{\mathrm{IT}}$ is the smallest sample complexity $\alpha$ such that the maximizer of the free entropy eq. (106) is $m \neq 0$. Equivalently, $\alpha^{\mathrm{IT}}$ is the smallest $\alpha$ such that $\mathrm{mmse} < 1$. Lemma D.2 readily implies that $\alpha^{\mathrm{IT}} = \alpha^{\mathrm{IT}}(\lambda)$ is non-increasing in the SNR.

Finally, we characterize the IT weak recovery threshold in the limit of small SNR. By expanding eq. (107) around $\lambda = 0$, we derive the following Corollary.

**Corollary D.4** (IT weak recovery threshold in the large noise regime)**.** *Consider the setting of Theorem D.3. Then, in the limit $\lambda \to 0$, the information theoretic weak-recovery threshold satisfies*

$$\alpha^{\mathrm{IT}} = \Theta(\lambda^{-1}) \tag{108}$$

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

$$\mathrm{mmse} = \Theta\left(\frac{1}{\lambda\alpha}\right). \tag{109}$$

*Proof.* For a fixed $m \in [0,1])$, consider the function

$$F(\kappa) := \mathbb{E}_{W,V,Y} \log \mathbb{E}_{w\sim\mathcal{N}(0,1)} \left[\exp\left(-(y - \kappa g(\sqrt{m}V + \sqrt{1-m}w))^2/2\right)\right], \tag{110}$$

with $V, W \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(\kappa g(\sqrt{m}V + \sqrt{1-m}W), 1)$. It is straightforward to show, by a change of variable $Y \to -Y$, that $F$ is even. Note that, up to constant terms in $m$, $F(\kappa) = \Psi_{\text{out}}(m; \kappa^2)$.

Furthermore, as a consequence of Taylor's theorem, there exists constant $\hat{\kappa} \in [0, \kappa]$, such that $F(\kappa) = F(0) + \frac{1}{2}F''(0)\kappa^2 + \frac{1}{24}F^{(4)}(\hat{\kappa})\kappa(\kappa - \hat{\kappa})^3$. By Lemma D.6, there exists a finite $C > 0$, constant in $\kappa$ and $m$, such that

$$\left|\frac{1}{24}F^{(4)}(\hat{\kappa})\kappa(\kappa - \hat{\kappa})^3\right| < C\kappa^4. \tag{111}$$

Define the auxiliary function

$$Z(\kappa, v, u, y) := \frac{1}{\sqrt{2\pi}}\mathbb{E}_w \left[\exp\left(-(y - \kappa(g(\sqrt{m}v + \sqrt{1-m}w) - g(\sqrt{m}v + \sqrt{1-m}u)))^2/2\right)\right]. \tag{112}$$

In particular, $Z(0, v, u, y) = (2\pi)^{-1/2}e^{-y^2/2}$ and, applying the definition of the probabilist's Hermite polynomials, for all $j \in \mathbb{N}$, using the shorthand $g_u = g(\sqrt{m}v + \sqrt{1-m}u)$

$$\frac{\partial^j}{\partial\kappa^j}Z(\kappa, v, u, y) = \frac{1}{\sqrt{2\pi}}\mathbb{E}_w \left[e^{-(y-\kappa(g_w - g_u))^2/2}\text{He}_j(y - \kappa(g_w - g_u))\frac{(g_w - g_u)^j}{j!}\right] \tag{113}$$

we have, after a change of variable $Y \to Y + \kappa g\sqrt{m}V + \sqrt{1-m}W)$, that $F(\kappa) = \mathbb{E}_{V,W,Y}[\log Z(\kappa, V, W, Y)]$ for $Y \sim \mathcal{N}(0,1)$ and

$$\begin{aligned}
F''(0) &= \mathbb{E}_{V,W,Y}\left[\frac{Z''(0,V,W,Y)}{Z(0,V,W,Y)} - \left(\frac{Z'(0,V,W,Y)}{Z(0,V,W,Y)}\right)^2\right] && (114)\\
&= \mathbb{E}_Y[Y](\mathbb{E}_{V,W}[g_w] - \mathbb{E}_{V,w}[g_w]) \\
&\quad - \frac{1}{2}\mathbb{E}_Y[(Y^2 - 1)]\left(\mathbb{E}_{V,W}[g_w^2] + \mathbb{E}_{V,W}[g_w^2] - 2\mathbb{E}_{V,W,w}[g_W g_w]\right) && (115)\\
&= 0 - \mathbb{E}_{z\sim\mathcal{N}(0,1)}[g^2(z)] + \mathbb{E}_{(z_1,z_2)\sim\mathcal{N}(\mathbf{0}_2,\boldsymbol{C})}[g(z_1)g(z_2)] && (116)
\end{aligned}$$

with

$$\boldsymbol{C} = \begin{pmatrix} 1 & m \\ m & 1 \end{pmatrix}. \tag{117}$$

In the above we used

$$\begin{aligned}
\mathbb{E}_{V,W}[g_W^2] &= \mathbb{E}_{W,V}[g^2(\sqrt{m}V + \sqrt{1-m}W)] = \mathbb{E}_z[g^2(z)], && (118)\\
\mathbb{E}_{V,w,W}[g_w g_W] &= \mathbb{E}_V[\mathbb{E}_W[g(\sqrt{m}V + \sqrt{1-m}W)]\mathbb{E}_w[g(\sqrt{m}V + \sqrt{1-m}w)]] && (119)\\
&= \mathbb{E}_{(z_1,z_2)\sim\mathcal{N}(\mathbf{0}_2,\boldsymbol{C})}[g(z_1)g(z_2)]. && (120)
\end{aligned}$$

Therefore, up to constant terms in $m$,[4]

$$\Psi_{\text{out}}(m; \lambda) = F(\sqrt{\lambda}) = \text{const} + \frac{\lambda}{2}\mathbb{E}_{(z_1,z_2)\sim\mathcal{N}(\mathbf{0}_2,\boldsymbol{C})}[g(z_1)g(z_2)] + O(\lambda^2). \tag{121}$$

Assumption 2.3 ensures that $g$ can be decomposed in the Hermite basis as

$$g(z) = \sum_{k\geq 0}\frac{c_k}{k!}\text{He}_k(z), \qquad c_k := \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[g(z)\text{He}_k(z)\right]. \tag{122}$$

Leveraging Proposition 11.31 in O'Donnell (2014),

$$\mathbb{E}_{(z_1,z_2)\sim\mathcal{N}(\mathbf{0}_2,\boldsymbol{C})}[g(z_1)g(z_2)] = \sum_{k\geq 0}\frac{c_k^2}{k!}m^k \in [0, \mathbb{E}_z[g^2(z)]], \tag{123}$$

[4]Recall that, by Lemma D.6, the correction $O(\lambda^2)$ is uniform in $m \in [0,1]$.

which is a non-decreasing function of $m \in [0,1]$. Note that, since $g$ has generative exponent 2, we have that $c_1 = 0$ necessarily. Moreover, by Assumption 2.3, $c_2 \neq 0$ and bounded. Hence, we are interested in maximizing the quantity[5]

$$f_{\mathrm{RS}}(m) := m + \log(1-m) + \alpha\lambda \sum_{k\geq 2} \frac{c_k^2}{k!} m^k + O(\alpha\lambda^2) \tag{124}$$

$$= \frac{1}{2}\left(\alpha\lambda c_2^2 - 1\right) m^2 + \sum_{k>2} \left(\frac{\alpha\lambda}{k!} c_k^2 - \frac{1}{k}\right) m^k + O(\alpha\lambda^2), \tag{125}$$

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

$$D := \inf_{m\in(0,1]} \frac{-m - \log(1-m)}{\sum_k \frac{c_k^2}{k!} m^k}, \tag{126}$$

which is strictly positive and well-defined. Note that

$$\lim_{m\to 0^+} \frac{-m-\log(1-m)}{\sum_k \frac{c_k^2}{k!} m^k} = c_2^{-2} \implies \frac{1}{\mathbb{E}_z[g^2(z)]} \leq D \leq c_2^{-2} = \frac{1}{\mathbb{E}_z[g''(z)]^2}. \tag{127}$$

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

$$D \leq \alpha^{\mathrm{IT}} \lambda \leq c_2^{-2} \implies \alpha^{\mathrm{IT}} = \Theta(\lambda^{-1}), \tag{128}$$

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]

$$\frac{\mathrm{d}}{\mathrm{d}m} f_{\mathrm{RS}}(m) \overset{!}{=} 0 \implies \frac{1}{1-m} = \alpha\lambda \sum_{k\geq 2} \frac{c_k^2}{k!} m^{k-2} \implies \sum_{k\geq 2} \frac{c_k^2}{k!} \left(m^{k-2} - m^{k-1}\right) = \frac{1}{\alpha\lambda}, \tag{129}$$

The equation is solved by $m = 1 - \Theta\left(\frac{1}{\alpha\lambda}\right)$. Theorem D.3 implies that, in this regime, $\mathrm{mmse}(\lambda) = \Theta((\alpha\lambda)^{-1})$. Together Lemma D.2, the result for arbitrary $\lambda$ is proved. □

**Remark D.5.** *The scaling $\alpha^{\mathrm{IT}} = \Theta(\lambda^{-1})$ in Corollary D.4 also holds when $c_2 = 0$, which is the case of a noise sensitivity exponent larger than one (Defilippis et al., 2026b). The lower bound $\alpha^{\mathrm{IT}} \geq D\lambda^{-1}$ is unchanged, as $D > 0$ also in this case. For the upper bound, fix $\hat{m} \in (0,1)$ and let $C := -\log(1-\hat{m}) / \sum_k \frac{c_k^2}{k!} \hat{m}^k$. Then, at $\alpha = C\lambda^{-1}$, eq. (124) gives $f_{\mathrm{RS}}(\hat{m}) = \hat{m} + O(\lambda) > 0 = f_{\mathrm{RS}}(0)$ for $\lambda$ small enough, so that $m = 0$ is not the global maximizer, and $\alpha^{\mathrm{IT}} \leq C\lambda^{-1}$.*

In the proofs of Corollary D.4 and Remark D.5 we have used the following lemma.

**Lemma D.6.** *Let $m \in [0,1]$ and consider the function $Z : \mathbb{R}^4 \to \mathbb{R}$*

$$Z(\kappa, v, w, y) := \frac{1}{\sqrt{2\pi}} \mathbb{E}_{U\sim\mathcal{N}(0,1)} \left[ e^{-\left(y - \kappa\left(g\left(\sqrt{m}v + \sqrt{1-m}U\right) - g\left(\sqrt{m}v + \sqrt{1-m}w\right)\right)\right)^2/2} \right]. \tag{130}$$

*Then, for all $k \in \mathbb{N}$, there exists a universal constant $C_k$ such that*

$$\mathbb{E}_{W,V,Y} \left| \frac{\partial^k}{\partial \kappa^k} \log Z(\kappa, V, W, Y) \right| < C_k. \tag{131}$$

*with $V, W, Y \sim \mathcal{N}(0,1)$.*

*Proof.* In order to simplify the notation, given $V \sim \mathcal{N}(0,1)$, define the shorthand $g_w = g(\sqrt{m}V + \sqrt{1-m}w)$. By Faà di Bruno's formula, using the shorthand $Z = Z(\kappa, v, w, y)$ we have that

$$\frac{\partial^k}{\partial\kappa^k} \log Z = \sum_{\boldsymbol{m}:\sum_j j m_j = k} \frac{k!}{m_1! \dots m_k!} \frac{(-1)^{|\boldsymbol{m}|}}{|\boldsymbol{m}| Z^{|\boldsymbol{m}|}} \prod_{j=1}^{k} \left(\frac{\partial_\kappa^{(j)} Z}{j!}\right)^{m_j}, \tag{132}$$

$$= \sum_{\boldsymbol{m}:\sum_j j m_j = k} \frac{k!}{m_1! \dots m_k!} \frac{(-1)^{|\boldsymbol{m}|}}{|\boldsymbol{m}|} \prod_{j=1}^{k} \left(\frac{\partial_\kappa^{(j)} Z}{j! Z}\right)^{m_j}, \tag{133}$$

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

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

with $\boldsymbol{m} \in \mathbb{N}^k$ and $|\boldsymbol{m}| = \sum_j m_j$. Therefore, applying the triangle inequality and Holder's inequality

$$\mathbb{E}\left|\frac{\partial^k}{\partial\kappa^k}\log Z\right| \leq \sum_{\boldsymbol{m}:\sum_j jm_j=k} \frac{k!}{m_1!\dots m_k!}\frac{(-1)^{|\boldsymbol{m}|}}{|\boldsymbol{m}|}\prod_{j=1}^{k}\left(\mathbb{E}\left|\frac{\partial_\kappa^{(j)}Z}{j!Z}\right|^{km_j}\right)^{1/k}. \tag{134}$$

Since the sum runs over a finite number of $k$-tuples, the problem can be reduced to showing the finiteness of the quantities

$$\mathbb{E}_{V,W,Y}\left|\frac{\frac{\partial^j}{\partial\kappa^j}Z(\kappa,V,W,Y)}{Z(\kappa,V,W,Y)}\right|^p, \qquad p\in\mathbb{N} \tag{135}$$

Recalling the definition of the probabilist's Hermite polynomials:

$$\frac{1}{Z}\frac{\partial^j}{\partial\kappa^j}Z = \frac{1}{Z\sqrt{2\pi}}\mathbb{E}_U\left[e^{-(y+\kappa(g_W-g_U))^2/2}\mathrm{He}_j(y+\kappa(g_W-g_U))\frac{(g_W-g_U)^j}{j!}\right] \tag{136}$$

$$= \mathbb{E}_{U\sim\mathbb{Q}}\left[\mathrm{He}_j(y+\kappa(g_W-g_U))\frac{(g_W-g_U)^j}{j!}\right], \tag{137}$$

where we have defined the probability density function

$$\mathrm{d}\mathbb{Q}(U) = \frac{\exp\left(-U^2/2-(y+\kappa(g_U-g_W))^2/2\right)}{\mathbb{E}_{U\sim\mathcal{N}(0,1)}\left[\exp\left(-U^2/2-(y+\kappa(g_W-g_U))^2/2\right)\right]}\mathrm{d}U. \tag{138}$$

Applying Jensen's inequality to the convex function $|\cdot|^p$ we obtain

$$\mathbb{E}_{V,W,Y}\left|\frac{\partial_\kappa^{(j)}Z(\kappa,V,W,Y)}{Z(\kappa,V,W,Y)}\right|^p \tag{139}$$

$$\leq \mathbb{E}_{V,W,Y}\mathbb{E}_{U\sim\mathbb{Q}}\left|\mathrm{He}_j(y+\kappa(g_W-g_U))\frac{(g_W-g_U)^j}{j!}\right|^p \tag{140}$$

using the change of variable $y\to y-\kappa g_W$

$$= \mathbb{E}_V\int \mathrm{d}y\mathbb{E}_W\frac{e^{-(y-\kappa g_W)^2/2}}{\sqrt{2\pi}}\frac{\mathbb{E}_{U\sim\mathcal{N}(0,1)}\left[e^{-(y-\kappa g_U)^2/2}\left|\mathrm{He}_j(y-\kappa g_U)\frac{(g_W-g_U)^j}{j!}\right|^p\right]}{\mathbb{E}_U\left[e^{-(y-\kappa g_U)^2/2}\right]} \tag{141}$$

$$\leq 2^{jp-1}\mathbb{E}_V\int \mathrm{d}y\mathbb{E}_W\frac{e^{-(y-\kappa g_W)^2/2}}{\sqrt{2\pi}}\frac{\mathbb{E}_{U\sim\mathcal{N}(0,1)}\left[e^{-(y-\kappa g_U)^2/2}\left|\mathrm{He}_j(y-\kappa g_U)\right|^p\frac{|g_W|^{jp}+|g_U|^{jp}}{j!^p}\right]}{\mathbb{E}_U\left[e^{-(y-\kappa g_U)^2/2}\right]} \tag{142}$$

$$= \frac{2^{jp-1}}{j!^p}\underbrace{\mathbb{E}_V\int \mathrm{d}y\mathbb{E}_{W\sim\mathbb{Q}}\left[|g_W|^{jp}\right]\frac{\mathbb{E}_{U\sim\mathcal{N}(0,1)}\left[e^{-(y-\kappa g_U)^2/2}\left|\mathrm{He}_j(y-\kappa g_U)\right|^p\right]}{\sqrt{2\pi}}}_{=:T_1} \tag{143}$$

$$+\frac{2^{jp-1}}{j!}\underbrace{\mathbb{E}_V\int \mathrm{d}y\mathbb{E}_U\frac{e^{-(y-\kappa g_U)^2/2}}{\sqrt{2\pi}}\frac{\mathbb{E}_{U\sim\mathcal{N}(0,1)}\left[e^{-(y-\kappa g_U)^2/2}\left|\mathrm{He}_j(y-\kappa g_w)\right|^p\ |g_w|^{jp}\right]}{\mathbb{E}_U\left[e^{-(y-\kappa g_U)^2/2}\right]}}_{=:T_2} \tag{144}$$

In the first term $T_1$, applying the Cauchy-Shwartz,

$$T_1 \propto \mathbb{E}_{V,U} \int \mathrm{d}y \mathbb{E}_{W\sim\mathsf{Q}} \left( e^{-(y-\kappa g_U)^2/4} |g_W|^{jp} \right) \left( e^{-(y-\kappa g_U)^2/4} |\mathrm{He}_j(y-\kappa g_U)|^p \right) \tag{145}$$

$$\leq \left( \mathbb{E}_V \int \mathrm{d}y \mathbb{E}_U [e^{-(y-\kappa g_U)^2/2}] \mathbb{E}_{W\sim\mathsf{Q}} \left[ |g_W|^{2jp} \right] \right)^{1/2}$$

$$\times \left( \mathbb{E}_V \int \mathrm{d}y \mathbb{E}_U e^{-(y-\kappa g_U)^2/2} |\mathrm{He}_j(y-\kappa g_U)|^{2p} \right)^{1/2} \tag{146}$$

$$= \left( \mathbb{E}_V \int \mathrm{d}y \mathbb{E}_{W\sim\mathcal{N}(0,1)} \left[ e^{-(y-\kappa g_W)^2/2} |g_W|^{2jp} \right] \right)^{1/2}$$

$$\times \left( \int \mathrm{d}y e^{-y^2/2} |\mathrm{He}_j(y)|^{2p} \right)^{1/2} \tag{147}$$

$$= \left( \mathbb{E}_V \int \mathrm{d}y e^{-y^2/2} \mathbb{E}_{W\sim\mathcal{N}(0,1)} \left[ |g_W|^{2jp} \right] \right)^{1/2} \left( \int \mathrm{d}y e^{-y^2/2} |\mathrm{He}_j(y)|^{2p} \right)^{1/2} \tag{148}$$

$$= \left( \sqrt{2\pi} \mathbb{E}_{z\sim\mathcal{N}(0,1)} [|g(z)|^{2jp}] \right)^{1/2} \left( \int \mathrm{d}y e^{-y^2/2} |\mathrm{He}_j(y)|^{2p} \right)^{1/2} < \infty, \tag{149}$$

where the first factor is bounded as $|g(x)| < C(1+|x|^b)$, for some constants $C \in \mathbb{R}$, $b \in \mathbb{N}$ (Definition D.1). Moreover, we have used

$$\mathbb{E}_{V,W}[g_W^2] = \mathbb{E}_{W,V}[g^2(\sqrt{m}V + \sqrt{1-m}W)] = \mathbb{E}_z[g^2(z)]. \tag{150}$$

Similarly,

$$T_2 \propto \mathbb{E}_V \int \mathrm{d}y \mathbb{E}_U \left[ e^{-(y-\kappa g_U)^2/2} |\mathrm{He}_j(y-\kappa g_U)|^p \ |g_U|^{jp} \right] \tag{151}$$

$$= \left( \mathbb{E}_{z\sim\mathcal{N}(0,1)} [|g(z)|^{jp}] \right) \left( \int \mathrm{d}y e^{-y^2/2} |\mathrm{He}_j(y)|^p \right) < \infty \tag{152}$$

Both $T_1$ and $T_2$ are bounded by quantities independent of $\kappa$ and $m$, and that are functions of the finite moments $\mathbb{E}_z[g^k(z)]$, which completes the proof. □