A (very) biased overview of Gaussian multi-index models

Bruno Loureiro[1]

[1]Département d'Informatique, École Normale Supérieure, PSL & CNRS, Paris, France

August 7, 2025

Abstract

Notes for a 3h lecture at the Statistical Physics & Machine Learning: moving forward summer school that took place at the Institut d'Études Scientifiques de Cargèse on August 5-15 2025.

1 Motivation

Consider the classical supervised learning regression problem with training data $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ drawn i.i.d. from a joint distribution $\rho(x, y)$ over $\mathbb{R}^{d+1}$.

Given a notion of risk, say the quadratic risk

$$R(f) = \mathbb{E}[(y - f(x))^2] \tag{1.1}$$

one can define the minimal risk achievable by a measurable function of the data (the *Bayes risk*)

$$R_\star = \inf_f R(f). \tag{1.2}$$

For the quadratic risk, the Bayes risk is attained by the *Bayes predictor* $f_\star(x) = \mathbb{E}[y|x]$, i.e. $R_\star = R(f_\star)$. The function $f_\star$ is also known as the *target function*, as it is what we would like to achieve in supervised learning.

One of the key problems in learning theory is to understand how much data is required to approximate $f_\star$. In full generality, fitting an arbitrary target might require exponential number of samples in the dimension d — a problem known as the *curse of dimensionality* (CoD). Therefore, to do better we necessarily need structural assumptions on ρ.

At first, one might think that regularity conditions on $f_\star$. It turns out that regularity is not enough, as even some quite well-behaved functions $f_\star : \mathbb{R}^d \to \mathbb{R}$ suffer from the CoD, as illustrated by the follow theorem.

Theorem 1.1 (Tsybakov (2008), informal)**.** Assume $x_i \sim \text{Unif}([0, 1]^d)$ and that $y_i = f_\star(x_i) + \varepsilon_i$ with $f_\star$ a 1-Lipschitz function and ε_i a sufficiently well-behaved noise.[1] Then:

$$\inf_f \sup_{f_\star \in \text{Lip}(1)} \mathbb{E}[(f(x) - f_\star(x))^2] \gtrsim n^{-\frac{2}{2+d}}, \tag{1.3}$$

where the expectation is over the marginal distribution $\rho(x)$. Equivalently, we need $n(\delta) \gtrsim \delta^{-\frac{2+d}{2}}$ in order to approximate $f_\star$ to precision δ.

Remark 1.1. A few remarks are in order.

- Note that the infimum is taken over all measurable functions of the data, i.e. it makes no assumption on the hypothesis class. This is known as a *minimax rate*.
- The proof of this result is based on a ε-net argument, which involve covering $[0, 1]^d$ and using the regularity of $f_\star$ to approximate it to a given precision.

[1] For instance, Gaussian with finite variance.

This result suggests that stronger structural assumptions are needed to dodge the CoD. A common folklore in machine learning is that even though the data distribution has some underlying low-dimensional structure, known as the manifold hypothesis. In supervised learning, there are two ways to consider this hypothesis: (a) the covariate marginal distribution has a low-dimensional structure; (b) the target function $f_\star : \mathbb{R}^d \to \mathbb{R}$ is a low-dimensional function of the covariates. The class of multi-index models is an instance of (b).

2 Gaussian multi-index models

A *multi-index function* is a function $f : \mathbb{R}^d \to \mathbb{R}$ that depends on the covariates $x \in \mathbb{R}^d$ only through a k-dimensional subspace of $\mathbb{R}^d$.

Definition 2.1 (Multi-index function)**.** Let $w_1, \dots, w_r \in \mathbb{S}^{d-1}(\sqrt{d})$ denote an orthonormal family of r vectors. A multi-index function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as:

$$f(x) = g\left(\langle w_1, x\rangle, \dots, \langle w_r, x\rangle\right) \tag{2.1}$$

where $g : \mathbb{R}^k \to \mathbb{R}$ is a non-linear function known as the *link function.* Equivalently, we denote $y = g(Wx)$ where $W \in \mathbb{R}^{r\times d}$ is the row-orthogonal matrix $WW^\top = dI_k$ obtained by stacking the vectors $(w_k)_{k\in[r]}$ row-wise. In particular, when $r = 1$ we say f is a *single-index function.*

Remark 2.1 (Random link function)**.** A common variation consists of allowing g to be a (possibly) stochastic function. In this case, we denote $y \sim \mathsf{P}_y(y|Wx)$, where P_y denote the model likelihood. Note that the case of a deterministic g is a particular case given by $\mathsf{P}_y(y|Wx) = \delta\left(y - g(Wx)\right)$.

Example 2.1. Many classical functions studied in learning theory and signal processing can be written as multi-index functions:

- Linear functions ($r = 1$): $g(z) = z$.
- Phase retrieval ($r = 1$): $g(z) = |z|$
- Perceptron ($r = 1$): $g(z) = \text{sign}(z)$
- Polynomials ($r > 1$): $g(z) = z_1 \dots z_r$.
- Parities ($r > 1$): $g(z) = \text{sign}(z_1 \dots z_r)$
- Two-layer neural networks ($r > 1$): $g(z) = \sum_{k=1}^{r} a_k \sigma(z_k)$.

where we denoted $z_k = \langle w_k, x\rangle$, $k \in [r]$.

In the following, our main object of study will be supervised learning problems where the training data $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ is drawn according to *Gaussian multi-index model* (GMIM): the covariates are drawn i.i.d. from a isotropic Gaussian distribution, and the labels are given by a multi-index function:

$$y_i \sim \mathsf{P}_y(\cdot|W_\star x_i), \qquad x_i \sim \mathcal{N}(0, {}^1\!/_d I_d), \quad \text{i.i.d..} \tag{2.2}$$

We will further assume that the weights are uniform vectors in the sphere $w_{\star,k} \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, i.i.d. on $k \in [r]$.[2]

Remark 2.2. Information theoretically, learning a GMIM should be considerably easier than a general smooth function in $\mathbb{R}^d$. Indeed, if the linear subspace span$(W_\star)$ is known, we can apply theorem 1.1 directly to g, changing $d \to r$, which requires $n = O(r^{-2/(2+r)})$. On the other hand, estimating the linear subspace span$(W_\star)$ requires $n = O(d)$. Therefore, if $d \gg r$ we can expect that information theoretically learning a MIM to require $n = O(d)$.

Finally, let's define what we mean when we say "to learn" a multi-index function. In particular, two notions will be useful.

[2] For some computations it will be more convenient to assume $W_\star$ is a Gaussian matrix with i.i.d. entries $\mathcal{N}(0, 1)$. Although this can be technically different, we will only use these interchangeably when they are equivalent in the asymptotic limits studied here.

Definition 2.2 (Learnability)**.** Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ denote n samples from a GMIM. Denote by $\hat{W}(\mathcal{D})$ an estimator of $W_\star$,[3] which we assume has norm $\left\|\hat{W}\right\|_{\mathrm{F}} = \Theta(d)$.

- We say that $\hat{W}$ has *weakly learned* or *weakly recovered* a subspace $V_\star \subset \operatorname{span}(W_\star)$ if:

$$\inf_{\substack{v \in V_\star \\ \|v\|_2 = 1}} \left\| \frac{\hat{W} W_\star^\top v}{d} \right\|_2 = \Theta_d(1), \qquad \text{w.h.p. as } d \to \infty \tag{2.3}$$

- Similarly, we say that $\hat{W}$ has *fully learned* or *fully recovered* a subspace $V_\star \subset \operatorname{span}(W_\star)$ if:

$$\inf_{\substack{v \in V_\star \\ \|v\|_2 = 1}} \left\| \frac{\hat{W} W_\star^\top v}{d} \right\|_2 = 1, \qquad \text{w.h.p. as } d \to \infty \tag{2.4}$$

  Note that full-recovery is equivalent to $\hat{W} = W_\star$ up to rotational symmetry.

Remark 2.3. Note that weakly learning a subspace implies that the risk in eq. (1.1) is smaller than the risk associated to a estimator, i.e. $\hat{w}_k \sim \operatorname{Unif}(\mathbb{S}^{d-1})$ i.i.d. Similarly, full-recovery implies $R = R_\star$.

Although both notions are important, our main focus in the following will be on weak learnability.

3 Fundamental computational limits

As we saw in section 1, the best estimator in quadratic error is given by the posterior mean. Given the training data, the posterior is given by:

$$p(W|\mathcal{D}) = \frac{p(W)}{Z_d(\mathcal{D})} \prod_{i=1}^{n} \mathsf{P}_y(y_i|Wx_i) \tag{3.1}$$

Therefore, understanding the fundamental limits of learning a GMIM in the high-dimensional limit $d \to \infty$ is equivalent to understanding the marginals of the posterior distribution in the high-dimensional limit.

3.1 Approximate message passing

Our main tool will be the *approximate message passing* algorithm. This is an iterative algorithm that seeks to approximate the marginals of the posterior eq. (3.1) from an initial guess $\hat{W}_0$:

$$\Omega_t = X f_{\text{in}}(B_t, A_t) - g_{\text{out}}(y, \Omega_{t-1}, V_t) V_t^\top \tag{3.2}$$

$$B_{t+1} = X^\top g_{\text{out}}(y, \Omega_t, V_t) - f_{\text{in}}(B_t, A_t) A_t^\top \tag{3.3}$$

where $\Omega_t \in \mathbb{R}^{n \times r}$ and $B_t \in \mathbb{R}^{d \times r}$ are matrices with rows $\omega_i, b_j \in \mathbb{R}^r$, respectively, and $f_{\text{in}}(;A) : \mathbb{R}^r \to \mathbb{R}^r$ and $g_{\text{out}}(y, ;V) : \mathbb{R} \times \mathbb{R}^r \to \mathbb{R}$ are two vector-valued functions acting row-wise on the matrices B_t, Ω_t:

$$g_{\text{out}}(y, \omega, V) = \mathbb{E}_{z \sim \mathcal{N}(0, I_r)} \left[V^{-1}(z - \omega) \mathsf{P}(y|z) \right], \qquad f_{\text{in}}(b, A) = (I_r - A)^{-1} b \tag{3.4}$$

and A_t, V_t are given by:

$$A_t = \frac{1}{d} \sum_{i=1}^{n} \nabla_{\omega_i} g_{\text{out}}(y_i, \omega_i, V_i), \qquad V_t = \frac{1}{d} \sum_{j=1}^{d} \nabla_{b_j} f_{\text{in}}(b_j, A_j) \tag{3.5}$$

Finally, the estimate of $W_\star$ after T steps is obtained by $\hat{W}_{\text{amp}} = f_{\text{in}}(B_T, A_T)^\top$

[3] In other words, any measurable function of the training data.

Remark 3.1. A few comments on this algorithm are in order.

- Intuitively, AMP can be seen as a two-step process, where one first seeks to estimate $z \in \mathbb{R}^r$ from $y = g(z)$ and then $W \in \mathbb{R}^{r\times d}$ from $z = Wx$. Indeed, the functions $f_{\text{in}}, g_{\text{out}}$ defined in eq. (3.4) are precisely the optimal Bayesian denoisers for these two estimation problems.
- AMP is a first order method: it involves only evaluations of r-dimensional functions and matrix multiplication operations by $X \in \mathbb{R}^{n\times d}$ and $X^\top \in \mathbb{R}^{d\times n}$. Therefore, for $r = O(1)$ its running time complexity is dominated by the matrix-vector product, which linear in the size of X: $\Theta(nd)$.
- Indeed, AMP is asymptotically the optimal first order method when $n, d \to \infty$ at fixed $^n/_d \to \alpha = \Theta(1)$. This has been proven by Celentano et al. (2020); Montanari and Wu (2024). Therefore, AMP provides a fundamental computational lower-bound for the class of first order algorithms.

The key fact that makes AMP a powerful theoretical tool is that its asymptotic performance can be tracked by a set of state evolution equations.

Lemma 3.1 (State evolution (Aubin et al., 2019; Gerbelot and Berthier, 2023))**.** Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ denote i.i.d. samples from a GMIM eq. (2.2). Run AMP from random initialisation $\hat{W}_0 \in \mathbb{R}^{r\times d}$ with $\hat{w}_{0,k} \sim \mathcal{N}(0, I_d)$ i.i.d. Denote by $\hat{W}_t$ the resulting estimator at time $t \in [T]$. Then, in the high-dimensional limit $n, d \to \infty$ with fixed ratio $^n/_d \to \alpha = \Theta_d(1)$, constant $r, T = \Theta_d(1)$, the limiting overlaps satisfy:

$$\frac{1}{d}\hat{W}_t\hat{W}_t^\top \xrightarrow{P} M_t, \qquad \frac{1}{d}\hat{W}_t W_\star^\top \xrightarrow{P} M_t, \tag{3.6}$$

with M_t satisfying the *state evolution equations* from initial condition M_0 iterated with:

$$M_{t+1} = F(M_t) \tag{3.7}$$

where:

$$F(M_t) = G\left(\alpha\mathbb{E}\Big[g_{\text{out}}\left(Y_t, \sqrt{M_t}\xi, I_r - M_t\right)^{\otimes 2}\Big]\right). \tag{3.8}$$

where $G(M) = (I_r + M)^{-1}M \in \mathbb{R}^{r\times r}$ and the expectation is taken over the following effective process

$$Y_t = g\left(\sqrt{I_r - M_t}Z + \sqrt{M_t}\xi\right), \tag{3.9}$$

with $Z, \xi \sim \mathcal{N}(0, I_r)$ independently. The asymptotic mean-squared error on the label prediction is then given by:

$$\mathbb{E}\left[\left(y - g\left(\hat{W}_t(X, y)x\right)\right)^2\right] \xrightarrow{P} \mathbb{E}[(Y_t - g(Z))^2],$$

where the expectation is taken over the effective estimation process eq.(3.9) and $\xrightarrow{P}$ denotes convergence in probability w.r.t the training data as $n, d \to \infty$.

Remark 3.2. It can be shown that F preserves the symmetry and psd properties. Therefore, if $M_0 \succeq 0$ is a symmetric psd matrix, so is $M_{t+1} = F(M_t)$ for every $t \in [T]$.

The state evolution equations reduce the problem of understanding the fundamental computational bottlenecks of learning GMIM in the high-dimensional limit to studying a dynamical system on psd matrices $M \in \mathbb{R}^{r\times r}$

3.2 Weak recovery threshold

Note that if we randomly initialise AMP, independently of $W_\star$, the initial overlap matrix will be element-wise asymptotically small:

$$\frac{\langle \hat{w}_{k,0}, w_{\star,l}\rangle}{d} = \Theta(^1/_{\sqrt{d}}) \tag{3.10}$$

This means that in the $d \to \infty$ limit, state evolution from a so-called *uninformed initialisation* is strictly the zero matrix $M_0 = 0$. Therefore, we must study where do we go from 0 under the state evolution eq. (3.7). In particular, we first need to understand if zero is a fixed point.

Lemma 3.2 (Existence of uninformed fixed point)**.** $M = 0 \in \mathbb{R}^{r\times r}$ is a fixed point of the state evolution eq. (3.7) if and only if the following condition holds almost surely over the effective process $Y = g(Z)$ for $Z \sim \mathcal{N}(0, I_r)$:

$$g_{\text{out}}(Y, 0, I_r) = \mathbb{E}[Z|Y] = \int_{\mathbb{R}^r} \frac{\mathrm{d}z}{(2\pi)^{d/2}} e^{-\frac{1}{2}||z||_2^2} \mathsf{P}_y(Y|z) = 0 \tag{3.11}$$

Note this is an intrinsic property of $\mathsf{P}_y(Y|\cdot)$, and it is satisfied for instance if P_y is an odd function of z.

Trivial functions — Note that if $M = 0$ is not a fixed point of eq. (3.7), we have M_1, i.e. a single iteration of AMP away from initialisation, *no matter how small is the sample complexity* $\alpha > 0$ suffices to weakly recover a subspace $T_\star$ of dimension $\dim(T_\star) = \text{rank}(M_1)$. We call this a *trivial subspace*. More formally, we can define:

Definition 3.1 (Trivial subspace)**.** Let $H_\star \subset \text{span}(W_\star)$ denote the subspace spanned by the vectors $v \in \mathbb{R}^r$ satisfying:

$$\langle g_{\text{out}}(Y, 0, I_r), v\rangle = \lim_{d\to\infty} \mathbb{E}[\langle W_\star^\top v, x\rangle | Y = y] = 0 \tag{3.12}$$

where equality holds almost surely over $Y = g(Z)$ with $Z \sim \mathcal{N}(0, I_r)$. The trivial subspace $T_\star$ is define as the orthogonal complement of $H_\star$, i.e. $\text{span}(W_\star) = T_\star \oplus H_\star$

Given this definition, we can show that:

Theorem 3.1. For any $\alpha > 0$, with high-probability as $d \to \infty$, the AMP algorithm eq. (3.2) weakly recovers $T_\star$ as per eq. (2.3) in a single iteration.

Example 3.1. To get some intuition, we can have a look at a few examples of trivial subspaces.

- For single-index models ($r = 1$), T_* is one dimensional if and only if g is non-even, e.g. $g(z) = \text{He}_3(z)$. This follows from requiring that $g_{\text{out}}(y, 0, 1) \neq 0$ for at least one value of y. In particular, on any open interval where g_{out} is invertible we have $g_{\text{out}} = g^{-1}$.
- For a linear multi-index model, $g(z) = \sum_{k=1}^r z_i$, $T_\star$ is spanned by $1_r \in \mathbb{R}^r$ (all-one vector).
- For a committee $g(z) = \sum_{k=1}^r \text{sign}(z_k)$, the trivial subspace $T_\star$ is again 1d, spanned by $1_r \in \mathbb{R}^p$.
- For monomials $g(z) = z_1 \dots z_r$, the trivial subspace $T_\star$ is non-empty if and only if $p = 1$.
- For leap one staircase functions (Abbe et al., 2023b):

$$g(z) = z_1 + z_1 z_2 + z_1 z_2 z_3 + \dots \tag{3.13}$$

  The trivial subspace is $T_\star = \mathbb{R}^r$ and is spanned by the canonical basis. In other words, AMP learns all the directions with a *single step* for any $\alpha > 0$.

Remark 3.3 (Optimal label pre-processing)**.** The condition on lemma 3.2 can be related to computational models based on queries, such as SQ learning (Kearns, 1998). Indeed, the denoiser g_{out} can be interpreted as a non-linear transformation on the labels $y \mapsto g_{\text{out}}(y, 0, I_r)$. From this perspective, the statement on the condition for the existence of a non-empty trivial subspace translates to the condition

$$\mathbb{E}[g_{\text{out}}(y, 0, I_r)^\top v \langle W_\star^\top v, x\rangle] = \mathbb{E}\left[\mathbb{E}[\langle W_\star^\top v, x\rangle | Y = y]^2\right] \neq 0 \tag{3.14}$$

where $v \in T_\star$. The left-hand side can be seen as a statistical query of the type $\mathbb{E}[\varphi(y)\psi(x)]$ with label pre-processing $\varphi = g_{\text{out}}$. In fact the denoiser g_{out} is the optimal such transformation in the sense that when g_{out} fails to obtain a linear correlation along $\boldsymbol{v}$, i.e when $v \in T_\star$, then no transformation can.

Easy functions — On the other hand, if $M = 0$ is a fixed point of eq. (3.7), we can have two types of behaviour: it can be an *unstable* fixed point (repeller) or *stable* fixed point (attractor) of the dynamics. These will behave very differently: if we initialise $\hat{W}_0$ such that $\|M_0\|_{\mathrm{F}} \approx \epsilon > 0$ arbitrarily small, a repeller implies that M_1 will move away from 0, while an attractor will attract back 0. In other words, if $M = 0$ is an attractor, a random initialisation will not be enough for AMP to meaningfully correlate with $W_\star$ at larger times. To characterise the stability of a fixed point, we need to look at the Jacobian of F around $M = 0$:

$$F(M) \approx \alpha \mathcal{F}(\delta M) + O(\|\delta M\|^2) \tag{3.15}$$

where $\mathcal{F}(\delta M)$ is a linear operator on the cone $\mathcal{S}_r^+$ of psd matrices of dimension r:

$$\mathcal{F}(M) = \mathbb{E}\left[G(Y)MG(Y)^\top\right]. \tag{3.16}$$

where the expectation is with respect to $Y = g(Z)$ with $Z \sim \mathcal{N}(0, I_r)$, and the operator $\hat{G}$ is given by:

$$\hat{G}(y) = \nabla_\omega g_{\mathrm{out}}(y, 0, I_r) = \mathbb{E}[ZZ^\top - I_r|y] \in \mathbb{R}^{r\times r} \tag{3.17}$$

The stability of the $M = 0$ fixed point is then closely related to the operator norm of $\mathcal{F}$ on the psd cone:

Lemma 3.3 (Stability of the uninformed fixed point)**.** If $M = 0 \in \mathbb{R}^{r\times r}$ is a fixed point of the state evolution equations. Then, it is an unstable fixed point if and only if $\|\mathcal{F}(M)\|_{\mathrm{F}} > 0$ and $n > \alpha_c d$, where the critical sample complexity α_c, known as the *weak recovery threshold*, is given by:

$$\frac{1}{\alpha_c} = \sup_{\substack{M\in\mathbb{R}^{r\times r}\\ \|M\|_F=1}} \|\mathcal{F}(M)\|_{\mathrm{F}}, \tag{3.18}$$

Moreover, if $\mathcal{F}(M) \neq 0$, there exists at least one $M_\star \neq 0 \in \mathcal{S}_p^+$ achieving the above supremum. While if $\mathcal{F}(M) = 0$, then $M = 0$ is a stable fixed point for any $n = \Theta(d)$.

This implies that for $\alpha > \alpha_c$, iterating the state evolution equations eq. (3.7) will eventually move us away from initialisation, provided we have $M_0 \neq 0$. We can further characterise the subspace learned in this case, which we denote the *easy subspace*:

Definition 3.2 (Easy subspace $E_\star$)**.** Let $H_* \subset \mathrm{span}(W_\star)$ denote a subspace of directions $v \in \mathbb{R}^r$ satisfying

$$\langle v, \hat{G}(Y)v\rangle = 0, \tag{3.19}$$

almost surely over $Y = g(Z)$, $Z \sim \mathcal{N}(0, I_r)$. We define the easy subspace $E_\star$ as the orthogonal complement of $H_\star$, i.e. $\mathrm{span}(W_\star) = E_\star \oplus H_\star$

The challenge in showing AMP effectively learns the easy subspace for $\alpha > \alpha_c$ from the state evolution equations and random initialisation is that, strictly speaking, $M = 0$ which is by construction a fixed point of the state equations. A standard approach to dodge that is to assume an arbitrarily small (but $\Theta_d(1)$) initial overlap, and showing that AMP converges to the easy subspace in a time that does not diverge too fast as this initial correlation vanishes. Formally, this equivalent to assuming we have access to an arbitrarily noisy version of $W_\star$, also known as *side information* channel:

$$S = \sqrt{\lambda}W_\star + \sqrt{1-\lambda}Z \tag{3.20}$$

where $Z \in \mathbb{R}^{r\times d}$ is a random Gaussian matrix with $\mathcal{N}(0, 1)$ entries, and $\lambda > 0$ quantifies the amount of side information.

Theorem 3.2. Let $M_{d,t} := {}^1\!/\!{}_d\hat{W}_t W^{\star\top}$ denote the model-target overlap matrix at any finite time t. Suppose that $T_\star = 0$ and consider the AMP algorithm eq. (3.2). Then, with high probability as $d \to \infty$:

(i) For $\alpha \geq \alpha_c$, $\exists\delta > 0$ such that for sufficiently small λ ,$M_d^t \succ \delta M_\star$ for $t = \mathcal{O}(\log {}^1\!/\!{}_\lambda)$, where $M_\star$ is any of the extremisers defined in eq. (3.18). Furthermore, there exists an $\alpha \geq \alpha_c$ and a $\delta > 0$ such that $M_d^t \succ \delta M_{E_*}$ in $t = \mathcal{O}(\log {}^1\!/\!{}_\lambda)$ iterations, where $M_{E^\star} \in \mathcal{S}_r^+$ spans $E_\star$.

(ii) For $\alpha < \alpha_c$ however, $M_d^t = 0$ is asymptotically stable i.e. there exist constants $\lambda' < 1$ and $C > 0$ such that for $\lambda < \lambda'$, $\sup_{t\geq 0} \|M_d^t\| \leq C\sqrt{\lambda}$.

Theorem 3.2 formalises our intuition about the stability of the uninformed fixed point. It implies that for $\alpha < \alpha_c$, not only does AMP fail to find any pertinent directions, but it also fails to improve on the small side-information. For $\alpha > \alpha_c$, however, AMP will develop a growing overlap M_t along a non-empty subspace starting with itarbitrarily small (but $\Theta_d(1)$) side information.

Remark 3.4 (Warm start)**.** Heuristically, we would like to identify $\lambda = \Theta(1/d)$. Of course, this is not justified, since in the state evolution equations we have already taken the high-dimensional limit, and would imply $\lambda = 0$ in this limit. To make sense of this in the AMP framework, we would need to a non-asymptotic control of the state evolution equations. This is mathematically challenging, and only available for a few simpler problems, see (Rush and Venkataramanan, 2018; Li and Wei, 2022; Li et al., 2023).

Remark 3.5 (Spectral methods)**.** An alternative approach is to construct other algorithms achieving the same asymptotic weak recovery threshold α_c, which by construction is the optimal computational threshold for first order methods. A standard approach is to construct spectral methods from linearising AMP, which maps weak recovery into a BBP transition problem for a matrix. This idea was pioneered by Mondelli and Montanari (2019); Luo et al. (2019); Maillard et al. (2022) in the context of single-index models ($r = 1$), and has been recently generalised by Kovačević et al. (2025); Defilippis et al. (2025) to the multi-index case. On a high-level, this consists of linearising AMP around the uninformed fixed point and closing the equations on $\delta\Omega_t$:

$$\delta\Omega_{t+1} = T\delta\Omega_t \tag{3.21}$$

where $T \in \mathbb{R}^{(rd)\times(rd)}$ is an operator with entries

$$T_{(k\mu),(l\nu)} = \sum_{i=1}^{n} X_{i\mu}X_{i\nu}\left[G(y_i)\left(G(y_i) + I_r\right)\right]_{kl}^{-1}, \qquad \mu,\nu \in [d], \quad k,l \in [r]. \tag{3.22}$$

where $G(y)$ was defined in eq. (3.17). Based on a state evolution argument, Defilippis et al. (2025) shown that this matrix has a BBP transition precisely at the optimal weak-recovery threshold α_c of theorem 3.2. This was rigorously proven by Kovačević et al. (2025); Defilippis et al. (2025) in the particular case in which the matrix $G(y_i)$ is jointly diagonalisable for all y_i, which is the case for many of the cases of interest.
Open problem: Prove this for general G.

Example 3.2. Two examples of easy multi-index functions.

- The monomials $g(z) = \prod_{k=1}^{r} z_k$ with $r > 1$ can always be learned with $\alpha > \alpha_c(r)$ (Chen and Meka, 2020). For instance, we have $\alpha_c(2) \approx 0.5937$, $\alpha_c(3) \approx 3.725$, $\alpha_c(4) \approx 4.912$ and $\alpha_c(r) \sim r^{1.2}$ for large r.
- The 2-sparse parity $g(z) = \mathrm{sign}(z_1 z_2)$ is easy, and can be learned with $\alpha_c = \pi^2/4$.

Hard functions — Finally, there are functions for which $M = 0$ is a stable fixed point for all $\alpha = \Theta_d(1)$, meaning $\alpha_c \to \infty$, i.e. the supremum in the right-hand side of eq. (3.18) is zero. This is what we call a *hard function*. A canonical example of a hard function is the r-sparse parity for $r > 2$:

$$g(z) = \prod_{k=1}^{r} \mathrm{sign}(z_k) \tag{3.23}$$

Remark 3.6. A few final remarks.

- **Single-index models:** The computational weak learnability threshold for single-index models were first established in (Mondelli and Montanari, 2019; Luo et al., 2019; Barbier et al., 2019). Barbier et al. (2019); Maillard et al. (2020) have also studied the computational and information theoretical thresholds for full-recovery in this simpler case.

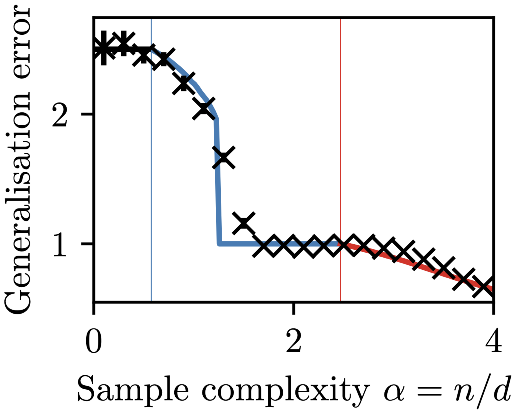

Figure 1: Weak learnability phase transitions for $g(z_1, z_2, z_3) = z_1^2 + \mathrm{sign}(z_1 z_2 z_3)$.

- **SQ and LDP:** Interestingly, Damian et al. (2024) have shown that the AMP weak recovery thresholds for single-index functions coincide exactly with the ones predicted by SQ lower-bounds and the low-degree polynomial method for the so-called generative exponent 2 functions, which are by definition functions for which $\alpha_c < \infty$. For SIM with generative exponent > 2, we have $\alpha_c = \infty$ — these are, by definition, the hard single-index functions from the AMP perspective. Recent work by Damian et al. (2025) extended some of these notions to multi-index functions.
- **Hierarchical Learning:** The discussion so far has focused on the the analysis of the initial condition $M_0 = 0$. The state evolution equations, however, can have other fixed points. The notion of easy and hard subspaces can be promptly generalised conditionally on a learned subspace. This allows to study the flow of AMP from fixed point to fixed point, which can lead to interesting hierarchical learning phenomena. For instance, a hard multi-index function $\mathrm{sign}(z_1 z_2 z_3)$ can become easy once it is coupled to an easy function z_1^2, i.e. $g(z) = z_1^2 + \mathrm{sign}(z_1 z_2 z_3)$. In this case, AMP first learn the function z_1^2 at sample complexity $\alpha_{c,1} \approx 0.575$, and after it is able to learn the hard part at $\alpha_{c,2} = \pi^2/4$. This was studied by Troiani et al. (2025), and to contrast with the staircase phenomenon for SGD is was named the *grand staircase*. See fig. 2

4 Learning with 2LNN

In the last section, we saw that (most of) GMIM can be (weakly) learned efficiently and with $n = \Theta(d)$ with a tailored AMP algorithm in the high-dimensional setting $d \to \infty$. But what about "standard" machine learning algorithms?

We will now consider the complexity of learning a GMIM with a two-layer neural network:

$$f(x; a, W) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} a_j \sigma(\langle w_j, x \rangle) \tag{4.1}$$

where σ is a non-linear activation. Our main question here will be to understand how adapting to the data during training, i.e. *feature learning*, allow the network to efficiently learn a GMIM.

However, before attacking this question it will be important to understand first what non-adaptive networks, i.e. networks with fixed features $\varphi(x) = 1/\sqrt{p}\sigma(W_0 x)$ can learn. Indeed, two-layer networks with fixed feature maps can be seen as a truncation of a kernel method, the so-called random features approximation. Since kernels are universal approximators (Micchelli et al., 2006), any $f_\star \in L^2(\rho_x)$ can be learned with enough data — and this will hold for random features provided p is large enough and σ is not poorly chosen.

4.1 Initialisation

Given the discussion above, consider a two-layer neural networks with a random, fixed first layer weights $w_{0,j} \sim \mathcal{N}(0, I_d)$, where we only train the second-layer:

$$\hat{a}_\lambda(X, y) = \arg\min_{a \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i; a, W_0))^2 + \frac{\lambda}{2} ||a||_2^2 \tag{4.2}$$

where the training data has been sampled from a GMIM. Note that this admits a closed-form solution, which can be written in terms of the feature matrix $\Phi_{ij} = {}^{\sigma(\langle w_{0,j}, x_i \rangle)}/_{\sqrt{p}}$:

$$\hat{a}_\lambda(X, y) = \left(\Phi^\top \Phi + n\lambda I_p\right)^{-1} \Phi^\top y = \Phi^\top \left(\Phi\Phi^\top + n\lambda I_n\right)^{-1} y \tag{4.3}$$

The linear operator $y \in \mathbb{R}^n \mapsto \hat{a}_\lambda(X, y) \in \mathbb{R}^p$ maps the training data into the column-space of the feature matrix $\text{Image}(\Phi^\top) \subset \mathbb{R}^p$, which is a linear subspace of the feature space (of dimension $\text{rank}(\Phi) = \min(n, p)$). To see this explicitly, consider the SVD of $\Phi = \sum_{j=1}^{\min(n,p)} \hat{\sigma}_j u_j v_j^\top$:

$$\hat{a}_\lambda(X, y) = \sum_{j=1}^{\min(n,p)} \frac{\hat{\sigma}_j}{\lambda + \hat{\sigma}_j^2} \langle u_j, y \rangle v_j \tag{4.4}$$

which has the form of a shrinkage operator. Therefore, in order to understand what can be fitted by our RF method in the high-dimensional limit, we need to understand the singular values $\hat{\sigma}$ of the feature matrix Φ, or equivalently the eigenvalues of the covariance $\Phi^\top\Phi$ or kernel matrix $\Phi\Phi^\top$ in this limit.

Since the data is Gaussian, it is useful to decompose the target function in an orthonormal basis with respect to the distribution of the covariates. Since we assume $x_i \sim \mathcal{N}(0, {}^1/_d I_d)$, this is given by the Hermite tensors:

$$f_\star(x) = \sum_{m=0}^{\infty} \sum_{\substack{\alpha \in \mathbb{Z}_+^d \\ |\alpha| = m}} C_\alpha H_\alpha(x) \tag{4.5}$$

where $\alpha \in \mathbb{N}^d$ is a multi-index and $H_\alpha(x) = \prod_{i=1}^{d} h_{\alpha_j}(x_j)$ are the Hermite tensors. This basis induces an orthogonal decomposition of $L^2(\mathbb{R}^d, \rho_x) = \bigoplus_{m \geq 1} V_m$, where V_m is the linear space spanned by polynomials of degree $m = |\alpha|$. The dimension of each subspace is given by:

$$\dim(V_m) = \binom{d + m - 1}{m} \tag{4.6}$$

The coefficients c_α quantify how much of the total energy of the target $||f_\star||_{\rho_x}^2 = \sum_\alpha c_\alpha^2$ lies in each subspace.

Since the ridge predictor in eq. (4.3) spans a linear subspace of dimension $r = \min(n, p)$, a naive power counting suggests that to learn the component of the target in subspace V_m requires $r = O(m)$, with the minimum between the number of samples n and the width p being the bottleneck for approximating V_m. Therefore, to fit the components up to $\leq \kappa$ of the target, one requires:

$$\min(n, p) \sim \sum_{m=0}^{\kappa} \binom{d + m - 1}{m} = O(d^\kappa) \tag{4.7}$$

This intuition was made precise by Mei et al. (2022), who proved the following theorem:

Theorem 4.1 (Mei et al. (2022), informal)**.** With $\min(n, p) = O(d^\kappa)$, the random features predictor in eq. (4.2) can learn at best a degree κ approximation of the target function $f_\star$. In other words:

$$\mathbb{E}\left[||f(x; \hat{a}_\lambda, W_0) - f_\star(x)||_2^2\right] = ||P_{>\kappa} f_\star||_{L^2(\rho_x)} + o_d(1) \tag{4.8}$$

Remark 4.1. The key ingredient in the argument above is that the data distribution is isotropic. This implies that the target will have energy uniformly spread over the different frequencies in the orthogonal decomposition. Similarly, since the features are not adapted to the structure of the data, to fit a given frequency it requires spanning the full space to be fitted.

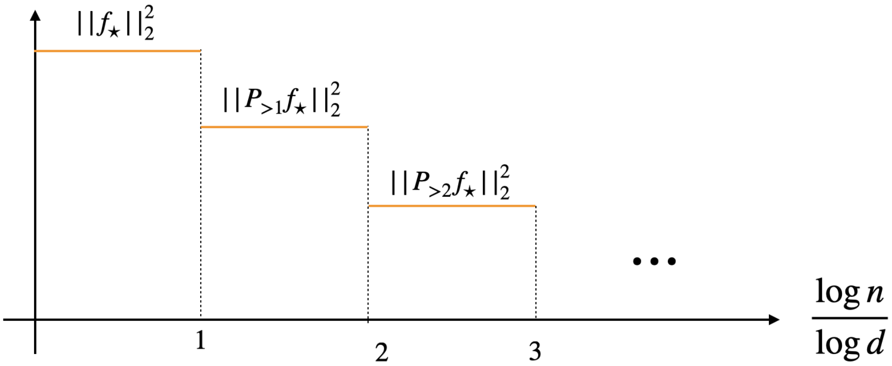

Figure 2: Illustration of Theorem 4.1

.

Gaussian universality — What is nice about the above result is that it can be made more quantitative in the high-dimensional limit. Consider the high-dimensional regime we studied in Section 3, where we take $d \to \infty$ with $n, p = \Theta(d)$. In this limit, theorem 4.1 tell us that the best we can hope is to fit a linear approximation of the target:

$$f_\star(x) \simeq \langle \theta_\star, x \rangle + f_{\rm nl}(x) \tag{4.9}$$

where $\theta_\star \in \mathbb{R}^d$ are the coefficients in the linear part of the decomposition in eq. (4.5), while $f_{\rm nl}$ are all the remaining, non-linear part that lead to the irreducible error — playing a role equivalent to label noise. Given this, an exact characterisation of the population risk associated to the predictor eq. (4.3) is a random matrix theorem problem. The key challenge in this problem is to deal with the Wishart matrix $\Phi^\top\Phi$ from non-linear features $\Phi = {}^1\!/\!\sqrt{p}\,\sigma(XW_0^\top)$. Again, the key idea is to consider the Hermite decomposition of the feature map:

$$\sigma(z) = \sum_{m\geq 0} b_m h_m(z) \tag{4.10}$$

Using this, the feature (population) covariance matrix is given by:

$$\Sigma_{jk} = \mathbb{E}[\sigma(\langle w_{0,j}, x\rangle)\sigma(\langle w_{0,k}, x\rangle)] = \sum_{m=0}^{\infty} b_m^2 \left(\frac{\langle w_{0,j}, w_{0,k}\rangle}{d}\right)^m \tag{4.11}$$

Note that since the neurons are independent, the scalar product above is of $\Theta(1)$ for $j = k$ and $O({}^1\!/\!\sqrt{d})$ for $j \neq k$, meaning that the frequency m has weight $\Theta(d^{-m/2})$ in the decomposition of the off-diagonal components. Therefore, to characterise the error in the high-dimensional limit, it suffices to keep the leading order terms in this expansion:

$$\Sigma \simeq b_0^2 1_p 1_p^\top + b_1^2 WW^\top + b_\star^2 I_p \tag{4.12}$$

where $b_\star^2 = \sum_{m\geq 2} b_m^2$. Higher-order terms, as well as corrections to the diagonal term, are negligible in the limit. With matrix concentration arguments, this approximation can be carried over to the empirical feature matrix.

Overall, this argument implies that studying the model with non-linear features $\Phi = {}^1\!/\!\sqrt{p}\,\sigma(XW_0)^\top$ is equivalent to studying a model with Gaussian features:

$$\Phi_{\rm lin} = b_0 1_p + b_1 XW_0^\top + b_\star Z \tag{4.13}$$

where Z is a Gaussian matrix with i.i.d. Gaussian entries $\mathcal{N}(0,1)$. This result, known as *Gaussian university*, was formulated in (Mei et al., 2022; Goldt et al., 2020, 2022; Hu and Lu, 2022; Montanari and Saeed, 2022), and was used to derive an exact asymptotic description of the population risk for the random features model in Mei et al. (2022); Gerace et al. (2020); Dhifallah and Lu (2020). Complementary RMT results on the asymptotic spectral density of the empirical covariance matrix of non-linear features appeared in (Pennington and Worah, 2017; Liao and Couillet, 2018; Fan and Wang, 2020; Benigni and Péché, 2021).

Remark 4.2. Generalisations of this result in different directions abound.

- Similar universality results were derived for deep random features model in the proportional regime by Schröder et al. (2023); Bosch et al. (2023); Schröder et al. (2024).
- Universality results beyond the proportional regime were derived by Lu and Yau (2022); Xiao et al. (2022); Hu et al. (2024).
- Random matrix theory analysis under general concentration assumptions on the feature map Misiakiewicz and Saeed (2024); Defilippis et al. (2024)

4.2 Away from initialisation

The result of theorem 4.1 holds for any target function $f_\star$, and is completely agnostic to the underlying low-dimensional structure of the GMIM. Indeed, in order to do better than polynomial fit, the features of our 2LNN neural networks must adapt to the structure of $f_\star$, i.e. *learn features.* Ideally, we would like to understand the joint training of both layers of $f(x; a, W)$. However, this is a challenging problem in general. Therefore, we will start with a simplified setting, which nevertheless will allow us to see the benefits of learning features.

Consider the following two-step training procedure:

1. Initialise a_0, W_0 randomly but in a balanced way, such that $f(x; a_0, W_0) \approx 0$.[4]
2. Take a single GD step on the first-layer weights W on a batch of training data $\mathcal{D}_0$ with n_0 samples drawn from the GMIM defined in eq. (2.2):

$$W_1 = W_0 - \eta \frac{1}{2n_0} \sum_{i=1}^{n_0} \nabla_W \left(y_i - f(x_i; a_0, W_0)\right)^2 \tag{4.14}$$

3. Train the second layer with fixed first layer W_1 on a fresh batch of data $\mathcal{D}_1$ with n samples drawn from a GMIM:

$$\hat{a}_\lambda(X, y) = \underset{a \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f(x_i; a, W_1)\right)^2 + \frac{\lambda}{2} ||a||_2^2 \tag{4.15}$$

Given this procedure, our goal is to understand how the error of $f(x; \hat{a}_\lambda, W_1)$ compares with the initialisation baseline studied in section 4.1, as a function of the learning rate η, batch sizes n_0, n and the dimensions p, d.

Again, eq. (4.15) admits a closed-form expression in terms of the feature matrix $\Phi = \sigma(XW_1^\top)$. The key difference is that now the features are correlated with the underlying data structure, since both batches were generated from the same multi-index target with weights $W_\star \in \mathbb{R}^{r \times d}$.

4.2.1 Weak recovery

A first natural question consists of quantifying this correlation. Since $f(x; a_0, W_0) \approx 0$, the first-step gradient is given by:

$$g_j = \frac{a_{0,j}}{n_0 \sqrt{p}} \sum_{i=1}^{n_0} g(W_\star x_i) x_i \sigma'(\langle w_{0,j}, x_i \rangle) \tag{4.16}$$

Again, it is natural to proceed by considering the Hermite decomposition of the activation σ as in eq. (4.10) and of the MIM link function in the Hermite tensor basis. For simplicity of exposition, let's consider the single-index case $r = 1$, where:

$$g(z) = \sum_{m=0}^{\infty} C_m h_m(z) \tag{4.17}$$

[4] Either exactly or in the high-dimensional limits considered in the following.

Then, the correlation with $w_\star$ satisfies in expectation:

$$\mathbb{E}[\langle g_j, w_\star \rangle] = \frac{a_{0,j}}{\sqrt{p}} \mathbb{E}[g(z_\star)\sigma'(z_j)z_\star] = \frac{a_{0,j}}{\sqrt{p}} \mathbb{E}[g'(z_\star)\sigma'(z_j)]$$
$$= \frac{a_{0,j}}{\sqrt{p}} \sum_{m=0}^{\infty} c_{m+1} b_{m+1} \left(\frac{\langle w_{0,j}, w_\star \rangle}{d} \right)^m \tag{4.18}$$

Again, the correlation in brackets is $\Theta(d^{-1/2})$, and therefore the sum will be dominated by the lowest non-zero frequency term. Provided σ is not degenerated and $a_j = \Theta(1)$, this is given by the so-called *information exponent* (Arous et al., 2021):

$$m_\star = \min\{m \in \mathbb{Z}_+ : \mathbb{E}[g(z)h_m(z)]\} \tag{4.19}$$

Therefore, the correlation with the first step ηg_j remains $\Theta(1)$ as long as:

$$\frac{\eta}{\sqrt{p}} = \Theta\left(d^{\frac{m_\star - 1}{2}}\right) \tag{4.20}$$

Otherwise, for any scaling below this one we will have vanishing correlation in the high-dimensional limit. Dandi et al. (2024a) has shown that this correlation concentrates as soon as $n = \Omega(d^{m_\star})$, with $n = \Theta(d^{m_\star - \delta})$ leading to zero asymptotic correlation for arbitrary $\delta > 0$.

Theorem 4.2 (Dandi et al. (2024a), informal)**.** Consider a single gradient step on the first-layer weights of a 2LNN trained on n_0 samples drawn from Gaussian single index model with learning rate $\eta/\sqrt{p} = \Theta(d^{\frac{m_\star - 1}{2}})$, where $m_\star$ is the information exponent of the target link function. Then, for any $j \in [p]$, with high-probability:

$$\lim_{d\to\infty} \frac{\langle w_{1,j}, w_\star \rangle}{||w_1||_2 ||w_\star||_2} \begin{cases} = 0 & n_0 = \Theta(d^{m_\star - \delta}) \\ \gtrsim 0 & n_0 = \Omega(d^{m_\star}) \end{cases} \tag{4.21}$$

Remark 4.3 (Previous work)**.** A few remarks on related works in the literature.

- The information exponent was introduced by Arous et al. (2021) in the context of analysing one-pass SGD for learning a well-specified Gaussian multi-index function. In this case, it was shown that the sample complexity is given instead by $n = \Theta(d^{m_\star - 1})$.
- Indeed, both large-step GD and one-pass SGD can be seen as CSQ-type algorithms. CSQ lower-bounds imply that $n_0 = \Theta(d^{m_\star/2})$ samples are required to weakly learn in this class. Leveraging landscape smoothing ideas from Biroli et al. (2020), Damian et al. (2023) showed that one-pass SGD on a smoothed loss can achieve this lower-bound.
- The result that with constant learning rate $\eta = \Theta(1)$ the correlation between the step and the features vanish for $p = \Theta(d)$ first appeared in Ba et al. (2022). Indeed, in this work it was shown that the Gaussian universality discussed in eq. (4.34) still holds, and therefore the error remains bounded by the best linear predictor.

This computation can be generalised to the multi-index case $r > 1$ by considering instead the decomposition of g in the Hermite tensor basis. In this case, the relevant quantity governing the sample complexity is a generalisation of the information exponent to tensors: the lowest non-vanishing sector in the tensor decomposition of g. This was introduced by Abbe et al. (2023a) and is known as the *leap exponent*. The learned subspace then corresponds to the space spanned by the principal directions of the tensor singular value decomposition of the target tensor coefficients $C_{m_\star}$:

$$C_{m_\star} = \sum_{j_1, \dots, j_{m_\star}} S_{j_1 \dots j_{m_\star}} u_{j_1} \otimes \cdots \otimes u_{j_{m_\star}} \tag{4.22}$$

A summary of the full result is given in fig. 3. The main take aways from this result are:

- With $n_0 = \Theta(d)$, one weakly learns at best a linear approximation of the multi-index target. This linear approximation is given by the weighted sum of the multi-index directions, where the relative weights are given by the information exponent associated to each direction.

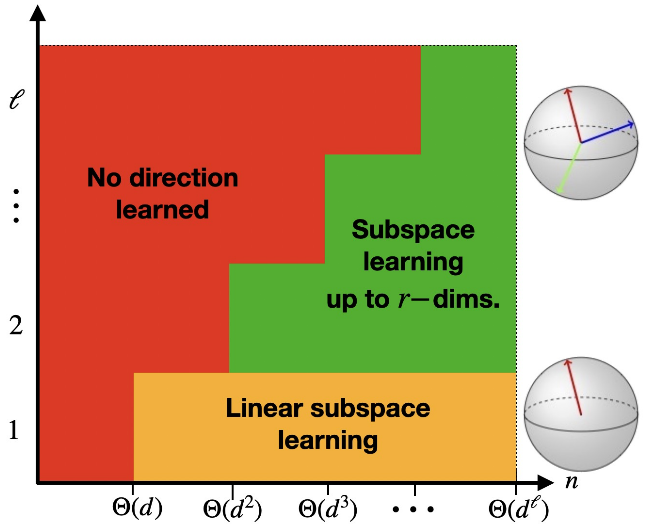

Figure 3: Illustration of the relationship between batch size and target function complexity for learning multi-index functions with a single giant step in the $n_0 = \Theta(d^k)$ regime.

- To weakly learn a subspace of dimension larger than one, one needs at least $n_0 = \Theta(d^2)$ data. This generalises the result of Damian et al. (2022), who showed that with $n = \omega(d^2)$ samples, two-layer neural networks can specialise to more than one direction of a multi-index target function with $\ell = 2$.

Remark 4.4 (Multiple GD steps)**.** The discussion above can be generalised to multiple GD steps, which can lead to the learning of additional directions and interesting hierarchical phenomena.

- Under the assumption that fresh data is used at every step, it can be that that even at limited sample complexity $n_0 = \Theta(d)$ there is a class of functions for which all directions can be learned. The so-called *staircase functions* were first introduced and studied by Abbe et al. (2021, 2022, 2023b) in the context of one-pass SGD. They are functions for which the different directions are conditionally linearly coupled in a progressive fashion, the canonical example being $g(z) = z_1 + z_1 z_2 + z_1 z_2 z_3$, etc — see Abbe et al. (2021, 2022, 2023b) for a formal definition.
- The picture can be drastically different when data is reused. This was first studied by Dandi et al. (2024b), who showed that certain high-information exponent functions such as $g(z) = h_3(z)$ can be weakly learned in two full-batch GD steps under data reuse. Surprisingly, a similar phenomenology holds for a simple SGD algorithm where each data point is seeing exactly twice before being refreshed, and was studied in two concurrent works by Lee et al. (2024); Arnaboldi et al. (2024).

Open problem (Optimality of GD): How does the pre-processing function induced by data reuse compare to the optimal pre-processing from AMP? Can we derive sharp asymptotic weak-recovery thresholds? How do they compare?

4.2.2 Generalisation error

How does the correlation in the features translate to generalisation when training the second layer $\hat{a}_\lambda$? This is a considerably harder problem to analyse. However, we expect that with enough data one can possibly learn any non-linear function in the subspace learned by the gradient step $U \subset \text{span}(W_\star)$. What about the unlearned directions? These should require a sample complexity comparable to the random features model.

To formalise this, let's look at a simple case where $U = \{v\}$, i.e. only a single direction has been learned. Then, for any point $x \in \mathbb{R}^d$, we can decompose $x = x^\perp + v$ where $\langle x^\perp, v\rangle = 0$. Since

x is Gaussian, so is $x^\perp$). Then, we can decompose the target function:

$$f_\star(x) = f_\star(x^\perp + v) = \sum_{m=0}^{\infty} \sum_{\substack{\alpha \in \mathbb{Z}_+^d \\ |\alpha|=m}} C_\alpha(v) H_\alpha(x^\perp) \tag{4.23}$$

where $C_\alpha(v) = \mathbb{E}[f_\star(x^\top + v)H_\alpha(x^\perp)]$ Given a sector $\kappa \in \mathbb{Z}_+$, we define the projector $P_{U,>\kappa}$ as a high-pass filter that only retain frequencies $m > \kappa$:

$$P_{U,>\kappa} f_\star(x) = \sum_{m>\kappa} \sum_{\substack{\alpha \in \mathbb{Z}_+^d \\ |\alpha|=m}} C_\alpha(v) H_\alpha(x^\perp) \tag{4.24}$$

With this notation, we can formalise our conjecture:

Conjecture 4.1 (Dandi et al. (2024a), informal)**.** Assume that $\min(n, p) = \Theta(d^\kappa)$ and that the first layer W_1 correlates with a subspace $U \subseteq \text{span}(W_\star)$. Then, ridge regression on the second layer weights satisfies the following lower-bound

$$\mathbb{E}\left[(f_\star(x) - f(x; \hat{a}_\lambda, W_1))^2\right] \geq \|P_{U,>\kappa} f_\star\|_{\rho_x}^2 - o(1), \tag{4.25}$$

Where $P_{U,>\kappa}$ is the projector defined in eq. (4.24).

Although proving this conjecture in full generality is an open problem, there is one particular case for which we know exactly how the error behaves.

Conditional Gaussian equivalence — The starting point to derive a sharp characterisation of the error is to have a better control of the gradient in eq. (4.16), which we repeat here for convenience:

$$g_j = -\frac{a_{0,j}}{n_0\sqrt{p}} \sum_{i=1}^{n_0} f_\star(x_i) \sigma'(\langle w_{0,j}, x_i \rangle) x_i \tag{4.26}$$

Similarly to how we did for the features in eq. (4.34) we can use the Hermite decomposition of the activation in eq. (4.10) to separate the leading order component:

$$\begin{aligned} -g_j &= \frac{a_{0,j} b_1}{\sqrt{p} n_0} \sum_{i=1}^{n} y_i x_i + \frac{a_j}{\sqrt{p} n_0} \sum_{i=1}^{n} y_i x_i \sigma'_{>1}(\langle w_{0,j}, x_i \rangle) \\ &= u_j v + \Delta_j \end{aligned} \tag{4.27}$$

where $b_1 = \mathbb{E}[\sigma(z) z]$, and we defined the spikes:

$$u = \frac{a_0 b_1}{\sqrt{p}} \in \mathbb{R}^p, \qquad v = \frac{1}{n_0} \sum_{i=1}^{n_0} y_i x_i \tag{4.28}$$

Note that $u \in \mathbb{R}^p$ contains information about the first layer initialisation, while $v \in \mathbb{R}^d$ correlates with the target directions via y_i, and $\Delta \in \mathbb{R}^{p \times d}$ denotes the remaining part. In the proportional limit, it can be shown that Δ is subleading with respect to the spikes, in the following sense.

Lemma 4.1. With high-probability as $d \to \infty$ with $n_0, p, \eta = \Theta(d)$:

- **Isotropy** For any target direction $w_{\star,k} \in \mathbb{R}^d$:

$$\langle \Delta_j, w_{\star,k} \rangle = O\left(\frac{\text{polylog}(d)}{p\sqrt{d}}\right) \tag{4.29}$$

- **Small operator norm**:

$$||\Delta||_{\text{op}} = O\left(\frac{\text{polylog}(d)}{\sqrt{d}}\right) \tag{4.30}$$

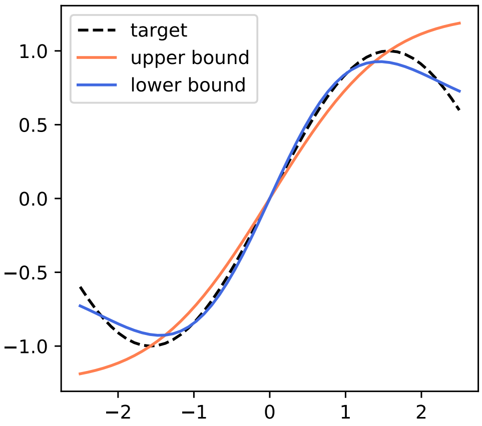

Figure 4: Illustration of the functions realising the upper bound (4.35) (orange) and lower bound (4.36) (blue), for $\sigma = \tanh$, for a target $\sigma_\star = \sin$ (dashed black).

- **Orthogonality**:

$$\langle \Delta_j, \Delta_k \rangle = O\left(\frac{\text{polylog}(d)}{p^2\sqrt{d}}\right) \tag{4.31}$$

This means that we can effectively treat the one-step update as a rank-one correction to the initial weights:

$$W_1 = W_0 + \eta uv + \Delta \tag{4.32}$$

Therefore, understanding the generalisation error of ridge regression after one step is equivalent to analysing a *spiked random features model* with feature matrix:

$$\Phi_{ij} = \frac{1}{\sqrt{p}}\sigma\left(\langle w_{0,j}, x_i\rangle + \eta u_j \langle v, x_i\rangle\right) \tag{4.33}$$

with spike correlated with the target direction $\langle v, w_\star\rangle > 0$.[5], where we have neglected the correction Δ.[6] This problem has been studied with statistical physics techniques by Cui et al. (2024), and later proven using RMT by Dandi et al. (2024b). The attentive reader can probably guess that the key step in this analysis is a linearisation of the feature matrix, similar to eq. (4.34), but taking into account the spike:

$$\Phi^{\text{lin}}_{ij} = b_0(u_j\kappa) + b_1(u_j\kappa)\langle w_{0,j}, x_i\rangle + b_\star(u_j\kappa)Z_{ij} \tag{4.34}$$

where we defined $\kappa = \langle v, x\rangle$, which conditionally on the spike $v \in \mathbb{R}^d$ is a Gaussian random variable. This result is known as the *conditional Gaussian equivalent*, and was first derived in Dandi et al. (2023) in the context of universality results for Gaussian mixture data. It was later adapted for spiked features in (Dandi et al., 2024a; Cui et al., 2024; Dandi et al., 2024b). The asymptotic results are cumbersome to write, but one can draw some interesting lessons from it, such as upper and lower bounds to the asymptotic excess risk.

Theorem 4.3 (Cui et al. (2024), informal)**.** Consider the one step setting with batch size $n_0 = \Theta(d^{1+\delta})$ and homogeneous first layer initialisation $a_0 = 1_p$. In the proportional regime where $d \to \infty$ with fixed $n, d = \Theta(d)$ and $\eta = \Theta(\sqrt{d})$, the excess risk of ridge regression satisfies the following upper and lower bounds under optimal choice of regularisation $\lambda \geq 0$:

$$\inf_{\lambda\geq 0}\lim_{d\to\infty} \mathbb{E}[(f(x;\hat{a}_\lambda, W_1) - g(\langle w_\star, x\rangle))^2] \leq \inf_{\nu_1} \mathbb{E}_\kappa\left[g(\kappa) - \nu_1 b_0(\kappa)\right]^2, \tag{4.35}$$

$$\inf_{\lambda\geq 0}\lim_{d\to\infty} \mathbb{E}[(f(x;\hat{a}_\lambda, W_1) - g(\langle w_\star, x\rangle))^2] \geq \inf_{\nu_1,\nu_2} \mathbb{E}_\kappa\left[g(\kappa) - \nu_1 b_0(\kappa) - \nu_2 b_1(\kappa)\kappa\right]^2. \tag{4.36}$$

[5] Recall that, thanks to theorem 4.2 we can assume without loss of generality that the target is a single-index model, since nothing else can be learned in the proportional regime

[6] Technically, one should treat Δ as a correction to the bulk matrix, or consider the scaling $n_0 = \Theta(d^{1+\delta})$ for arbitrary $\delta > 0$ to ignore it completely, see Dandi et al. (2024b) for a detailed discussion.

Remark 4.5. Theorem 4.3 shows that the asymptotic errors are upper and lower bounded by a two-dimensional optimisation problem over scalars ν_1, ν_2. A consequence of the theory in Cui et al. (2024); Dandi et al. (2024b) is that the effective number of parameters in the high-dimensional limit is proportional to the number of elements of the alphabet in which $a_{0,j}$ is initialised. This implies that adding diversity in the initial weights $a_{0,j}$ increases the expressivity of the network.

Open problem: Derive exact asymptotic results for features under multiple SGD steps. This probably will require a conditional Gaussian equivalence for low-rank matrices.

References

Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34:26989–27002, 2021.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023a.

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

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124023, dec 2019. doi: 10.1088/1742-5468/ab43d2. URL https://dx.doi.org/10.1088/1742-5468/ab43d2.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

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

Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021.

Giulio Biroli, Chiara Cammarota, and Federico Ricci-Tersenghi. How to iron out rough landscapes and get optimal performances: averaged gradient descent and its application to tensor pca. *Journal of Physics A: Mathematical and Theoretical*, 53(17):174003, 2020.

David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4132–4179. PMLR, 2023.

Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In Jacob Abernethy and Shivani 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. URL https://proceedings.mlr.press/v125/celentano20a.html.

Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1161–1227. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/chen20a.html.

Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborova, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9662–9695. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24d.html.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36:752–784, 2023.

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

Alex Damian, Jason D Lee, and Joan Bruna. The generative leap: Sharp sample complexity for efficiently learning gaussian multi-index models. *arXiv preprint arXiv:2506.05500*, 2025.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36:54754–54768, 2023.

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

Yatin Dandi, Luca Pesce, Hugo Cui, Florent Krzakala, Yue M Lu, and Bruno Loureiro. A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities. *arXiv preprint arXiv:2410.18938*, 2024b.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents for random feature regression. *arXiv preprint arXiv:2405.15699*, 2024.

Leonardo Defilippis, Yatin Dandi, Pierre Mergny, Florent Krzakala, and Bruno Loureiro. Optimal spectral transitions in high-dimensional multi-index models. *arXiv preprint arXiv:2502.02545*, 2025.

Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33:7710–7721, 2020.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 09 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaad020. URL https://doi.org/10.1093/imaiai/iaad020.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

Hong Hu, Yue M Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv preprint arXiv:2403.08160*, 2024.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

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

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *Advances in Neural Information Processing Systems*, 37:58716–58756, 2024.

Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*, 2022.

Gen Li, Wei Fan, and Yuting Wei. Approximate message passing from random initialization with applications to z 2 synchronization. *Proceedings of the National Academy of Sciences*, 120(31): e2302930120, 2023.

Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3063–3071. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/liao18a.html.

Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices with polynomial scalings. *arXiv preprint arXiv:2205.06308*, 2022.

Wangyu Luo, Wael Alghamdi, and Yue M Lu. Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Transactions on Signal Processing*, 67(9):2347–2356, 2019.

Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11071–11082. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/7ec0dbeee45813422897e04ad8424a5e-Paper.pdf.

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

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator. *arXiv preprint arXiv:2403.08938*, 2024.

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

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

Andrea Montanari and Yuchen Wu. Statistically optimal firstorder algorithms: a proof via orthogonalization. *Information and Inference: A Journal of the IMA*, 13(4):iaae027, 2024.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *Advances in neural information processing systems*, 30, 2017.

Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In *International Conference on Machine Learning*, pages 30285–30320. PMLR, 2023.

Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. Asymptotics of learning with deep structured (Random) features. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 43862–43894. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/schroder24a.html.

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

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN 9780387790527. URL https://books.google.fr/books?id=mwB8rUBsbqoC.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.