Paper 2405

Luca Arnaboldi & collaborators

August 26, 2026

Abstract

Neural networks can identify low-dimensional relevant structures within high-dimensional noisy data, yet our mathematical understanding of how they do so remains scarce. Here, we investigate the training dynamics of two-layer shallow neural networks trained with gradient-based algorithms, and discuss how they learn pertinent features in multi-index models, that is target functions with low-dimensional relevant directions. In the high-dimensional regime, where the input dimension d diverges, we show that a simple modification of the idealized single-pass gradient descent training scenario, where data can now be repeated or iterated upon twice, drastically improves its computational efficiency. In particular, it surpasses the limitations previously believed to be dictated by the Information and Leap exponents associated with the target function to be learned. Our results highlight the ability of networks to learn relevant structures from data alone without any pre-processing. More precisely, we show that (almost) all directions are learned with at most $O(d \log d)$ steps. Among the exceptions is a set of hard functions that includes sparse parities. In the presence of coupling between directions, however, these can be learned sequentially through a hierarchical mechanism that generalizes the notion of staircase functions. Our results are proven by a rigorous study of the evolution of the relevant statistics for high-dimensional dynamics.

1 Introduction

Gradient Descent-based algorithms such as Stochastic Gradient Descent (SGD) and its variations, are a fundamental tool in training neural networks, and are therefore the subject of intense theoretical scrutiny. Recent years have witnessed significant advancements in understanding their dynamics and the learning mechanisms of neural nets. Significant progress has been made, in particular, in the case of two-layer networks thanks, in part, to the so-called mean-field analysis [MMN18; CB18; RV22; SS20]. A large part of the theoretical approach focused on *one-pass* optimization algorithms, where *each iteration involves a new fresh batch of data*. In particular, in the mathematically solvable case of high-dimensional synthetic Gaussian data, and a low-dimensional multi-index target function model, the class of functions efficiently learned by these *one-pass* methods has been thoroughly analyzed in a series of recent works, and has been shown to be limited by the so-called information exponent [BGJ21] and leap exponent [AAM22; AAM23]. These analyses have sparked many follow-up theoretical works over the last few months, see, e.g. [DLS22; Dam+23; Dan+24a; BBP23; Ba+23; Mon+24; Mou+23; ZB24], leading to a picture where SGD dynamics is governed by the information exponent, that is, the order of the first non-zero Hermite coefficient of the target.

A common point between all these works, however, is that they considered the idealized situation where one processes data one sample at a time, all of them independently identically distributed, without any repetition. This is, however, not a realistic situation in practice. Indeed, most datasets contain similar data-points, so that repetition occurs (see e.g. for repetition and duplicate in CIFAR [BD20]). Additionally, it is common in machine learning to repeatedly go through the same mini-batch of data multiple times over epochs. Finally, many gradient algorithms, often used in machine learning in Lookahead optimizers [Zha+19], Extragradient methods [Kor76], or Sharpness Aware Minimization (SAM) [For+21], *explicitly* take a few gradient steps over the exact same data-point. It is thus natural to wonder if using such extra-gradient algorithms, or simply reusing many times some data, would change the picture with respect to the current consensus.

There are good reasons to believe this to be the case. Indeed, it has been observed very recently [Dan+24b] that iterating twice over large ($O(d)$) datasets was enough to alter the picture reached in [BGJ21; AAM22; AAM23]. Importantly, [Dan+24b] showed that there exist functions that can be learned over few (just two) batch repetitions, while they require a much larger (i.e. polynomial in d) number of steps otherwise. While the set of function considered in this work was limited, the conclusion is indeed surprising, thus motivating the following question:

Can data repetition increase the efficiency of SGD when learning any multi-index functions?

We establish a positive answer to the above question by the analysis of more realistic optimization schemes with respect to the standard One-Pass SGD. Indeed, the hardness exponents developed in the seminal works [Abb+21; BGJ21] are substantially bound to the idealized training scenario that considers i.i.d data. Slight modifications to this training scenario toward a more realistic setting starkly change the global picture. Such slight modifications include:

(i) non-vanishing correlations between different data points, that is a natural requirement for real datasets;

(ii) processing the same data point multiple times in the optimization routine, i.e., a standard step in any algorithmic procedure when looping over different epochs.

We consider a simple yet analytically tractable single-pass training scheme: processing one Gaussian sample at a time, but *repeating the gradient step twice*. This minor adjustment, common in Lookahead optimizers [Zha+19], Extragradient methods [Kor76], and Sharpness-Aware Minimization (SAM) [For+21], significantly boosts efficiency. In fact, we will see that this approach matches the performance of our best algorithms in most cases.

We show in particular that most multi-index functions are learned with total sample complexity either $O(d)$ or $O(d \log d)$ if they are related to the presence of a symmetry, without any need for preprocessing of the data that is performed in various works (See e.g., [MM18; LAL19; Mai+20; CM20]). Our results thus demonstrate that shallow neural networks turn out to be significantly more powerful than previously believed [BGJ21; AAM22; AAM23; Arn+23] in identifying relevant features in the data. In fact, they are shown in many (but not all) cases to saturate the bounds predicted by statistical queries (SQ) [Dam+24] (rather than the more limited correlation statistical queries (CSQ) bounds that were characteristic of the former consensus [BGJ21; AAM23]). Equivalently, it roughly mirrors the prediction of the approximate message passing analysis in

[Tro+25]. Our conclusions follow from the rigorous analysis of the overlaps between the trained weight vectors with the low-dimensional relevant subspace that draws inspiration from the pivotal works of [SS95; CS00; CSX00; BGJ21].

Concurrently with the present work, [Lee+24] also studies the properties of SGD under data repetition in neural networks. We discuss the differences between our works in Section 3.2.

2 Setting and Main Contributions

Given a dataset of N labeled examples $\{\boldsymbol{z}^\nu, y^\nu\}_{\nu\in[N]}$ in d dimensions, we analyze the learning properties of fully connected two-layer networks with first layer weights $W = \{\boldsymbol{w}_j\}_{j\in[p]}$, second layer $\boldsymbol{a} \in \mathbb{R}^p$, and activation function $\sigma : \mathbb{R} \to \mathbb{R}$

In this chapter the input covariate is denoted $\boldsymbol{z}$, playing the role of $\boldsymbol{x}$ in the rest of the thesis.

$$f(\boldsymbol{z}; W, \boldsymbol{a}) = \frac{1}{p}\sum_{j=1}^{p} a_j \sigma\left(\langle \boldsymbol{z}, \boldsymbol{w}_j\rangle\right) \tag{1}$$

We consider target functions that are dependent only on few orthogonal relevant directions $k = O_d(1)$ encoded in the target's weight matrix $W^\star = \{\boldsymbol{w}_r^\star\}_{r\in[k]} \in \mathbb{R}^{k\times d}$:

The target $f^\star$ plays the role of the teacher and the network f that of the student from Section **??**; we switch to *target/network* here because $f^\star$ and f no longer share the same architecture.

$$y^\nu = f^\star(\boldsymbol{z}^\nu) = h^\star(W^\star \boldsymbol{z}^\nu) \tag{2}$$

This model for structured data is usually referred to as a *multi-index* model. Our main objective will be to analyze how efficiently two-layer nets adapt to this low-dimensional structure during training. To ensure that all necessary Gaussian expectations exist, we make the following non-restrictive assumption on $\sigma, h^\star$:

Assumption 0.1. *The functions σ and $h^\star$ have at most polynomial growth, in the sense that there exists a constant $C > 0$ such that for all $x \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^k$,*

$$|\sigma(x)| \le C(1+|x|)^C \quad \textit{and} \quad |h^\star(\boldsymbol{x})| \le C(1+\|\boldsymbol{x}\|)^C. \tag{3}$$

Training algorithm – Our goal is to minimize the following objective:

$$\mathcal{R}(W, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{z},y}\left[\mathcal{L}(f(\boldsymbol{z}, W, \boldsymbol{a}), y)\right] \tag{4}$$

where $\mathcal{L} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is a loss function, and the expectation is over the data distribution.

This objective is non-convex, and we do not have direct access to the expectation over the data. A central question in Machine Learning is therefore to quantify the properties of the weights attained at convergence by different algorithms. In this manuscript we consider a specific class of one-pass training routines to minimize the empirical risk. First, we partition the dataset in disjoint mini-batches of size n_b, i.e. $\mathcal{D} = \cup_{t\in[T]}\{\boldsymbol{z}^\nu, y^\nu\}_{\nu\in\{tn_b+1,\dots,(t+1)n_b\}}$ where T is the total number of iterations considered. For each algorithmic step, we update the hidden layer neurons $W = \{\boldsymbol{w}_j\}_{j\in[p]}$ with the program $\mathcal{A} : (\boldsymbol{w}_{j,t}, \gamma, \rho, n_b) \in \mathbb{R}^{d+3} \to \boldsymbol{w}_{t+1,j} \in \mathbb{R}^d$.

This chapter departs from the discrete/continuous time convention of Section **??**: here t denotes the discrete SGD step (called ν there), since ν is reserved for indexing the individual samples $\boldsymbol{z}^\nu$.

Algorithm 0.1 (Optimizer Main Step)**.**

$$\begin{aligned}\boldsymbol{w}_{t+1,j} &= \mathcal{A}(\boldsymbol{w}_t, \gamma, \rho, n_b) = \boldsymbol{w}_{t,j} - \frac{\gamma}{n_b} \sum_{\nu \in [n_b]} \nabla_{\boldsymbol{w}_{t,j}} \mathcal{L}\left(f(\boldsymbol{z}^\nu; \tilde{W}_t(\rho), \boldsymbol{a}_0), y^\nu\right) \\ \tilde{\boldsymbol{w}}_{t,j}(\rho) &= \boldsymbol{w}_{t,j} - \rho \nabla_{\boldsymbol{w}_{t,j}} \mathcal{L}\left(f(\boldsymbol{z}^\nu; W_t, \boldsymbol{a}_0), y^\nu\right)\end{aligned} \tag{5}$$

This class of algorithms defined by the optimizer step $\mathcal{A}(\cdot, \gamma, \rho, n_b)$, in which we consider the final gradient update after taking a linear combination of the current iterate and its current gradient (noted as $\tilde{W}_t(\rho)$ in the above), is broadly used in different contexts. Indeed, routines with positive ρ parameters correspond to Extragradient methods (EgD) [Kor76], while with negative ρ have been recently used in the context of Sharpness Aware Minimization (SAM) [For+21]. For clarity, we present our theoretical results with the above, easily interpretable, Algorithm 0.1. However, our theoretical claims are valid in a more general setting; we refer to Appendices B and D for additional results on more general optimizer steps.

As previously stated, the central object of our analysis is the efficiency of the network to adapt to the low-dimensional structure identified by $W^\star$. Therefore, we focus on the learned representations by the first layer weights W_t while keeping the second layer weights $\boldsymbol{a}_t = \boldsymbol{a}_0$ fixed during training. This assumption is favorable to performing the theoretical analysis and is largely used in the theoretical community (e.g., [DLS22; Ba+22; Bie+22]).

Numerous theoretical efforts have been devoted to understanding the sample complexity needed to learn multi-index targets. However, an important aspect of many algorithmic routines considered is to exploit a clever *warm start* to initialize the iterates [MM18; LAL19; Mai+20; CM20]. We show in this work that such preprocessing of the data is not needed to learn all polynomial multi-index functions in $T = O(d \log d)$ with the routine in Algorithm 0.1. Instead, we consider a fully agnostic initialization of the network:

Assumption 0.2. *For each $j \in [p]$, the initial first-layer weight $\boldsymbol{w}_{0,j}$ is initialized according to the uniform measure on the sphere:*

$$\boldsymbol{w}_{0,j} \overset{i.i.d}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1}).$$

The second-layer weights are drawn i.i.d in $\{-1, 1\}$:

$$a_{0,j} \overset{i.i.d}{\sim} \mathrm{Unif}(\{-1, 1\}).$$

Weak recovery in high dimensions – The essential object in our analysis is the evolution of the correlation between the hidden neurons $W = \{\boldsymbol{w}_j\}_{j \in [p]}$ and the target's weights $W^\star$, as a function of the algorithmic time steps. More precisely, we are interested in the number of algorithmic steps needed to achieve an order one correlation with the target's weights $W^\star$, known as *weak recovery*.

Definition 0.1 (Weak recovery)**.** *The target subspace $V^\star$ is defined as the span of the rows of the target weights $W^\star$:*

$$V^\star = \mathrm{span}(\boldsymbol{w}_1^\star, \dots, \boldsymbol{w}_k^\star) \tag{6}$$

We define the weak recovery stopping time for a parameter $\eta \in (0,1)$ independent from d as:

$$t_\eta^+ = \min\{t \geq 0 : \|WW^{\star\top}\|_F \geq \eta\} \tag{7}$$

From Information Exponents to Generative Exponents – Recent works have sharply theoretically characterized the weak recovery stopping time for a variety of multi-index targets learned with one-pass SGD. [BGJ21] for the single-index case, then [AAM22] for the multi-index one have unveiled the presence of an Information Exponent ℓ characterizing the time complexity needed to weakly recover the target subspace $V^\star$, defined in the single-index case as

$$\ell = \min\{k \in \mathbb{N} : \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[h^\star(x)H_k(x)\right] \neq 0\}, \tag{8}$$

where H_k is the k-th Hermite polynomial [ODo14]. Under this framework, the sample complexity required to achieve weak recovery when the link function has Information Exponent ℓ is [BGJ21]

$$T(\ell) = \begin{cases} O(d^{\ell-1}) & \text{if } \ell > 2 \\ O(d\log d) & \text{if } \ell = 2 \\ O(d) & \text{if } \ell = 1. \end{cases} \tag{9}$$

These bounds have been improved by [Dam+23] up to order $d^{\ell/2}$; the latter matches the so-called *Correlated Statistical Query* lower bound, which considers queries of the form $\mathbb{E}\left[y\phi(\boldsymbol{z})\right]$.

However, this Information Exponent is not the right measure of complexity for low-dimensional weak recovery. Some iterative algorithms, for instance, are known to perform drastically better[Bar+19; CCM21; MM18; LAL19; Mai+20; Tro+25]. Recently [Dam+24] introduced a new Generative Exponent $\ell^\star$ (generative exponent) governing the weak recovery time scaling for algorithms in the Statistical Query (SQ) or Low Degree Polynomial (LDP) family. This exponent is defined as

$$\ell^\star = \min\{k \in \mathbb{N} : \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[f(h^\star(x))H_k(x)\right] \neq 0 \text{ for some } f : \mathbb{R} \to \mathbb{R}\}, \tag{10}$$

which allows for the application of arbitrary transformations of the labels before performing the queries. Under this definition, the "hard" problems with $\ell^\star > 2$ are also impossible to learn for the best first-order algorithm, *Approximate Message Passing* [Tro+25].

We discuss in this work the extension of the generative exponent to the multi-index setting and investigate the dynamics of specific gradient-based programs (Algorithm 0.1) as a function of the freshly defined hardness exponent $\ell^\star$. These theoretical findings match the recent one of [Kou+24] on the computational efficiency of SGD in learning sparse $k-$parity (associated with $\ell^\star = k$).

When the optimal transformation f is known, a simple SGD algorithm on $(\boldsymbol{z}, f(y))$ achieves the Generative Exponent lower bound. However, one problem remains: does there exist a class of SGD-like algorithms that can achieve this lower bound agnostically?

Achieving lower bounds with data repetition – Although the performance of single-pass SGD is limited by the Information Exponent of $h^\star$, the situation drastically changes when multiple-pass algorithms are considered. Recently, [Dan+24b] proved that non-even single-index targets are weakly recovered in $T = O(d)$ when considering extensive batch sizes with multiple pass. This begs the question of how good can a "reusing" algorithm be. We answer this question for the class of polynomial link functions:

Theorem 0.1 (Informal)**.** *There is a choice of hyperparameters such that if $h^\star$ is a polynomial function, Algorithm 0.1 achieves weak recovery in $O(d \log d)$ samples.*

We quote the sharp bound; the additional, not-provably-optimal polylogarithmic factor at generative exponent ≥ 2 is suppressed here and carried in full ($d \log(d)^2$) by the formal Theorem 0.2.

This shows in particular that a simple gradient-based algorithm, which only requires seeing the data twice, is optimal—in terms of SGD performance—for weak recovery of any polynomial target.

Learning representations – Learning relevant features from data is a fundamental property of neural networks. However, much theoretical work has focused on lazy training in two-layer networks [JGH18], where features are not effectively learned. In this regime, shallow networks behave as kernel machines, which struggle to adapt to low-dimensional structures in data [Gho+20]. Understanding representation learning and corrections to the kernel regime is thus a key challenge in machine learning theory [DH18; ABP22; Bie+22; DLS22; Pet+22]. In this manuscript we provably characterize the number of iterations needed from Algorithm 0.1 to learn the relevant features of the data, i.e., attain weak recovery of the target subspace. Our results provide the scaling of the relevant hyperparameters (γ, ρ) with the diverging input dimension d. The minibatch size will be fixed to $n_b = 1$ in the main body, and we refer to Appendix D for the extension to larger batch sizes.

Summary of Main Results

- We prove for single-index targets that one-pass gradient-based algorithms surpass the Correlation Statistical Query (CSQ) limits established by the Information Exponent [BGJ21].
- We unveil that the dynamics of the family of Algorithm 0.1 are governed by the generative exponent [Dam+24], with an additional polynomial restriction on the transformation of the output.
- We generalize the notion of generative exponent to multi-index targets. As in the single-index case, one-pass gradient-based algorithms can overcome CSQ performance dictated by the Leap Coefficient [AAM22]. Our definition is consistent with the one of [Tro+25].
- We prove that all polynomial multi-index functions are learned by Algorithm 0.1 either with total sample complexity $O(d)$ or $O(d \log(d)^2)$ if associated with the presence of a symmetry.
- The implementation of Algorithm 0.1 does not require pre-processing to initially correlate the estimation with the ground truth, and learns the meaningful representations from data alone. This is in contrast numerous findings (e.g., [MM18; LAL19; Mai+20] for single index targets or [CM20] for multi-index ones).

- We characterize the class of hard functions not learned in (almost) linear sample complexity. We show, however, that such functions can be learned through a hierarchical mechanism that extends the CSQ staircase first developed in [Abb+21] to a different, larger set, of functions.
- We validate the formal theoretical claims with detailed numerical illustrations. The code to reproduce representative figures is available in the project GitHub repository.

3 Single-Index Model

3.1 Main results

We first consider, for simplicity, the class of single-index models, in which $k = p = 1$. This corresponds to a mismatched setting of Generalized Linear Models, in which we want to learn

$$y = h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle), \quad \text{with} \quad f(\boldsymbol{z}, \boldsymbol{w}, a) = a\sigma(\langle \boldsymbol{w}, \boldsymbol{z}\rangle).$$

In this section we rigorously characterize the number of iterations needed for the class of Algorithms 0.1 to perform weak recovery (as in Definition 0.1) in the context of single-index targets. We establish a clear separation between the learning efficiency of the algorithmic family 0.1 and One-Pass SGD, limited by the Information Exponent of the target to be learned [BGJ21].

We start by introducing a restriction of the generative exponent in [Dam+24] to polynomial transformations.

Definition 0.2 (Polynomial Generative Exponent)**.** *We define $\ell_p^\star$ as the smallest integer k such that there exists a polynomial $p : \mathbb{R} \to \mathbb{R}$ with:*

$$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[p(h^\star(x))H_k(x)\right] \neq 0, \tag{11}$$

Our assumptions on the SGD implementation and the target function are as follows:

Assumption 0.3. *Algorithm 0.1 is run with single-sample steps ($n_b = 1$), and uses the correlation loss*

$$\mathcal{L}(y, \hat{y}) = 1 - y\hat{y}.$$

We also consider a spherical version of Algorithm 0.1, with

$$\boldsymbol{w}_{t+1} = \frac{\boldsymbol{w}_t - \gamma\nabla_{\boldsymbol{w}}^\perp \mathcal{L}\left(f(\boldsymbol{z}^\nu; \tilde{\boldsymbol{w}}_t(\rho), a_0), y^\nu\right)}{\left\|\boldsymbol{w}_t - \gamma\nabla_{\boldsymbol{w}}^\perp \mathcal{L}\left(f(\boldsymbol{z}^\nu; \tilde{\boldsymbol{w}}_t(\rho), a_0), y^\nu\right)\right\|} \tag{12}$$

where $\nabla^\perp f = \nabla f - \langle \nabla f, \boldsymbol{w}\rangle \boldsymbol{w}$ is the spherical gradient.

Assumption 0.4. *The activation σ is analytic almost everywhere, and $\sigma^{(n)}$ satisfies the polynomial growth assumption (3) for $n \in \mathbb{N}$ (possibly for different constants C_n). Further, for all $n \geq 1$, we have*

$$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma^{(n)}(x)\sigma'(x)^{n-1}\right] \neq 0 \quad \textit{and} \quad \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[x\sigma^{(n)}(x)\sigma'(x)^{n-1}\right] \neq 0$$

We are now ready to state our main result.

Theorem 0.2. *Suppose that Assumptions 0.1-0.4 hold, and let $\rho = \rho_0 d^{-1}$. Then, under an event of probability $1/2$ on the initialization $(a_0, \boldsymbol{w}_0)$, for each $\delta > 0$, there exist constants $\gamma_0(\delta)$ and $C(\delta)$ such that the following holds:*

- *If $\ell_p^\star = 1$, choosing $\gamma = \gamma_0(\delta) d^{-1}$, then $\mathbb{P}(t_\eta^+ \leq C(\delta) \cdot d) \geq 1 - \delta$,*
- *If $\ell_p^\star = 2$, choosing $\gamma = \gamma_0(\delta)[d \log(d)]^{-1}$, then $\mathbb{P}(t_\eta^+ \leq C(\delta) \cdot d \log(d)^2) \geq 1 - \delta$.*

The above holds for almost every choice of ρ_0 under the Lebesgue measure. In particular, in a network with $p = O(\log \frac{1}{\delta})$ neurons, at least one of them achieves weak recovery with probability $1 - \delta$.

Proof Sketch – We provide here the proof sketch for Theorem 0.2. For any vector $\boldsymbol{w}$, under the correlation loss, we have

$$\nabla_{\boldsymbol{w}} \mathcal{L}(f(\boldsymbol{z}; \boldsymbol{w}, a_0), y) = -a_0 y \sigma'(\langle \boldsymbol{w}, z \rangle) \cdot \boldsymbol{z}, \tag{13}$$

so the gradient is aligned with $\boldsymbol{z}$. As a result, we have

$$\nabla_{\boldsymbol{w}} \mathcal{L}\left(f(\boldsymbol{z}; \tilde{\boldsymbol{w}}(\rho), a_0), y\right) = -a_0 y \sigma'\left(\langle \boldsymbol{w}, \boldsymbol{z} \rangle + a_0 \rho y \sigma'(\langle \boldsymbol{w}, \boldsymbol{z} \rangle) \cdot \underbrace{\|\boldsymbol{z}\|^2}_{\approx d}\right) \cdot \boldsymbol{z} \tag{14}$$

This expression for the gradient exhibits two important properties:

- Even though $\rho \asymp d^{-1}$, since the gradient lies along $\boldsymbol{z}$, the additional dot product with $\boldsymbol{z}$ amplifies the signal by a factor of d,
- The resulting gradient (14) is a *non-linear* function of y, as opposed to the linear one of (13).

The latter property enables us to show that the dynamics driven by equation (14) can implement all polynomial transformations of the output y, for almost all choices of ρ_0. Subsequently, the SGD algorithm on the transformed input $p(y)$ can be studied with the same techniques as [BGJ21], but with an information exponent equal to $\ell_p^\star$.

3.2 Discussion

Significance and necessity of the assumptions – Assumption 0.3 simplifies Algorithm 0.1 to a more tractable version. In particular, the normalization step allows us to keep track of only one parameter, the overlap between $\boldsymbol{w}_t$ and $\boldsymbol{w}^\star$. We expect however Theorem 0.2 to hold in the general setting of Algorithm 0.1.

Assumption 0.4 is a regularity Assumption, ensuring that some quantities of interest in the proof are non-zero. The first part is satisfied by virtually every activation function; the second is more restrictive, but we show that it is satisfied for biased activations:

Lemma 0.1. *Let σ be a non-polynomial analytic function. Then the second part of Assumption 0.4 is true for the function $x \mapsto \sigma(x + b)$ for almost every choice of b (according to the Lebesgue measure).*

Finally, although we state our result as an almost sure event over ρ_0, it can be reformulated into an event on the second layer weight a_0, in which the randomness comes from the initialization step.

Learning polynomial functions – Although our definition of the Polynomial Generative Exponent is more restrictive than the one in [Dam+24], we show that it is enough to allow learning of a large class of functions, namely polynomials:

Theorem 0.3. *Assume that $h^\star$ is a polynomial function. Then the Polynomial Generative Exponent of $h^\star$ is always at most 2, and is equal to 1 whenever $h^\star$ is not even.*

This theorem implies that the Extragradient algorithm class allows us to match the sample complexity of the algorithm of [CM20], without the need for *ad hoc* preprocessing.

Beyond Algorithm 0.1 – A succinct summary of the proof goes as follows: the first time we see a sample $\boldsymbol{z}$, we store information about y along the direction of $\boldsymbol{z}$. The second time we see the same sample, the component along $\boldsymbol{z}$ in $\boldsymbol{w}$ interact *non-linearly* with y, which bypasses the CSQ framework. Importantly, we expect that this behavior also appears in SGD algorithms where we do not see the data twice *in a row*, but *across multiple epochs*! Indeed, upon small enough correlation between the $\boldsymbol{z}^\nu$, seeing multiple examples before the repetition of $\boldsymbol{z}$ should not interfere with the data stored along this direction. We are therefore not claiming that Algorithm 0.1 is superior to other classically-used methods; rather, it provides a tractable and realistic setting to study data repetition in SGD, which was ignored in previous theoretical works on the topic.

3.3 Numerical illustrations

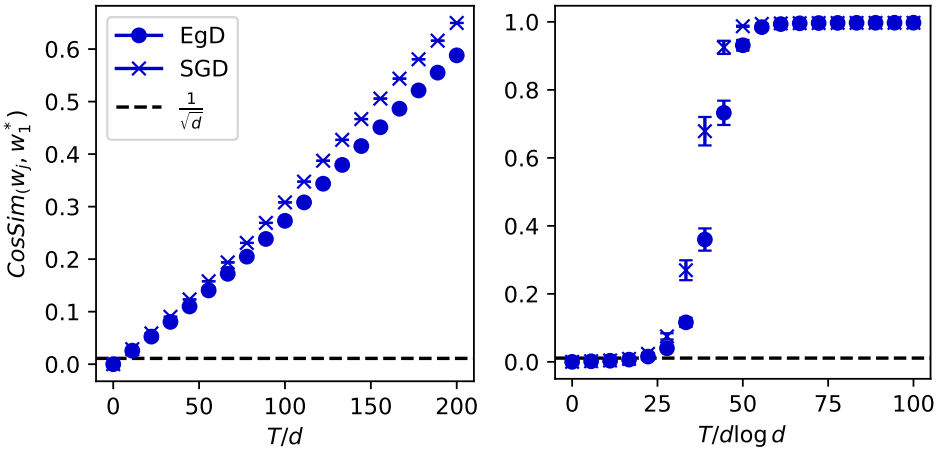
We illustrate in Figure 1 the stark difference between the weak recovery dynamics of One-Pass SGD and Algo. 0.1 for single-index targets. The Cosine Similarity between the learned weight $\boldsymbol{w}_t$ and the ground truth informative direction $\boldsymbol{w}_\star$ is shown as a function of the time steps t for different target functions. Two optimization routines are considered to exemplify the learning behavior:

(i) Extragradient Descent (EgD), corresponding to the family of Algorithms 0.1 with positive ρ parameters;

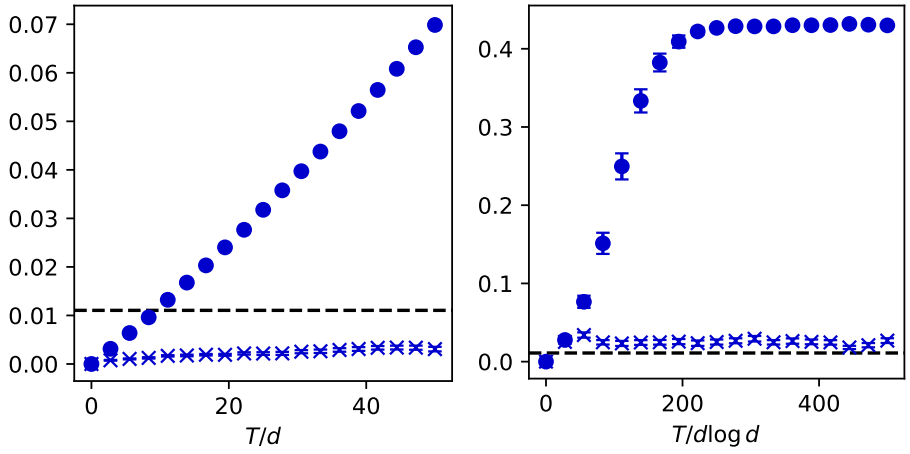
(ii) One-Pass SGD, corresponding to vanilla SGD, or equivalently Algorithm 0.1 associated with $\rho = 0$ hyperparameter.

The scaling of (γ, ρ) as a function of the input dimension d and exponent $\ell^\star$ are given in Theorem 0.2, while the mini-batch size is fixed to $n_b = 1$ (See Appendix D for extension to larger n_b). For all plots, we take $\sigma =$ relu and we refer to Appendix E for details.

- **SGD-easy non-symmetric targets ($\ell = \ell^\star = 1$) –** The Left section of Figure 1 shows $h^\star(x) = \mathrm{relu}(x)$. Here both SGD and Algorithm 0.1 learn in $T = O(d)$.
- **SGD-easy symmetric targets ($\ell = \ell^\star = 2$) –** The Center-Left section of Figure 1 shows $h^\star(x) = \mathrm{He}_2(x)$. Here both SGD and Algorithm 0.1 learn in $T = O(d \log d)$.
- **SGD-hard non-symmetric targets ($\ell > 2, \ell^\star = 1$) –** The Center-Right section of Figure 1 shows $h^\star(x) = \mathrm{He}_3(x)$. Here Algorithm 0.1 learns in $T = O(d)$ while One Pass SGD suffers from the limitations detailed in Equation (9), i.e. $T = O(d^{\ell-1})$ with $\ell = 3$ for this case.



(a) $h^\star(x) = \mathrm{relu}(x)$ (b) $h^\star(x) = \mathrm{He}_2(x)$



(c) $h^\star(x) = \mathrm{He}_3(x)$ (d) $h^\star(x) = \mathrm{He}_4(x)$

Figure 1: **Learning single-index targets** – Evolution of the Cosine Similarity attained by EgD (dots) and SGD (crosses) as a function of the normalized iteration time. (dashed horizontal line $\frac{1}{\sqrt{d}}$ is a visual guide to place random performance). **(a)** $(\ell, \ell^\star) = (1, 1)$**:** both algorithms learn in linear time. **(b)** $(\ell, \ell^\star) = (2, 2)$**:** both algorithms learn in $O(d \log d)$ time. **(c)** $(\ell, \ell^\star) = (3, 1)$**:** EgD learns in linear time, while SGD requires $O(d^2)$ time. **(d)** $(\ell, \ell^\star) = (4, 2)$**:** EgD learns in $O(d \log d)$ time, SGD requires $O(d^3)$ time (Details in Appendix E).

- **SGD-hard symmetric targets ($\ell > 2, \ell^\star = 2$) –** The Right section of Figure 1 shows $h^\star(x) = \mathrm{He}_4(x)$. Here Algorithm 0.1 learns in $T = O(d \log d)$ while One Pass SGD suffers from the time scaling law detailed in Equation (9) $T = O(d^{\ell-1})$.

4 Multi-Index Model

We now investigate the superiority of Algorithm 0.1 over the idealized One-Pass SGD training scheme when learning multi-index targets. In this setting, the networks weights can align with only a few select directions from $V^\star$. To quantify this behavior, we define directional versions of the Information and Generative Exponents (see also [Tro+25]):

Definition 0.3. *Let $\boldsymbol{v} \in V^\star$. The Information Exponent $\ell_{\boldsymbol{v}}$ (respectively* Polynomial Generative Exponent *$\ell^\star_{\boldsymbol{v}}$) of y in the direction $\boldsymbol{v}$ is the smallest k such that*

$$\mathbb{E}\left[yH_k(\langle \boldsymbol{v}, \boldsymbol{z}\rangle)\right] \neq 0 \quad (\textit{resp.}\mathbb{E}\left[p_{\boldsymbol{v}}(y)H_k(\langle \boldsymbol{v}, \boldsymbol{z}\rangle)\right] \neq 0)$$

for a polynomial $p_{\boldsymbol{v}}$. We also let $\ell = \min_{\boldsymbol{v}} \ell_{\boldsymbol{v}}$ and $\ell^\star_p = \min_{\boldsymbol{v}} \ell^\star_{\boldsymbol{v}}$. The minimal subspace $U^\star \subseteq V^\star$ is defined as the smallest subspace U such that $\ell_{\boldsymbol{v}} > \ell$ (resp. $\ell^\star_{\boldsymbol{v}} > \ell^\star_p$) for every $v \in U^\perp$.

The identification of the class of hard functions has been subject of intense theoretical scrutiny [Abb+21; AAM22; AAM23; Dan+24a; BBP23; Tro+25]; and for the problem of initial alignment it was shown to depend on the Information exponent defined above. Thus, similarly to the single-index scenario, the time complexity needed for Algorithm 0.1 to weakly recover the target subspace follows the law in Equation (9) for the multi-index information exponent ℓ. A natural question is thus to ask whether the class of Algorithm 0.1 also allows us to bypass this requirement. We show that this is indeed the case. Consider a neural network as in Equation (1); we assume that the internal learning rate ρ is chosen independently for each neuron:

Assumption 0.5. *For each $j \in [p]$, the parameter $\rho_{0,j}$ is chosen independently from all other neurons, according to a* symmetric *distribution which is absolutely continuous w.r.t the Lebesgue measure.*

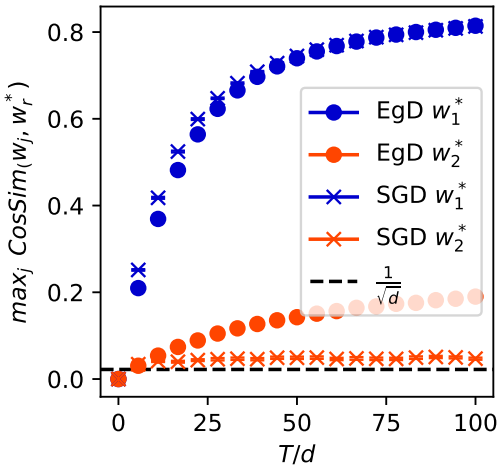
Theorem 0.4. *Let $\delta > 0$, and $\Pi^\star$ the projection operator on $V^\star$. There exists constants $\gamma_0(\delta), C(\delta)$ depending only on δ such that if we take $p \geq C(\delta)k$ and*

- *if $\ell^\star_p = 1$, $\gamma = \gamma_0(\delta)d^{-1}$ and $T = C(\delta)d$,*
- *if $\ell^\star_p = 2$, $\gamma = \gamma_0(\delta)(d\log(d))^{-1}$ and $T = C(\delta)d\log(d)^2$,*

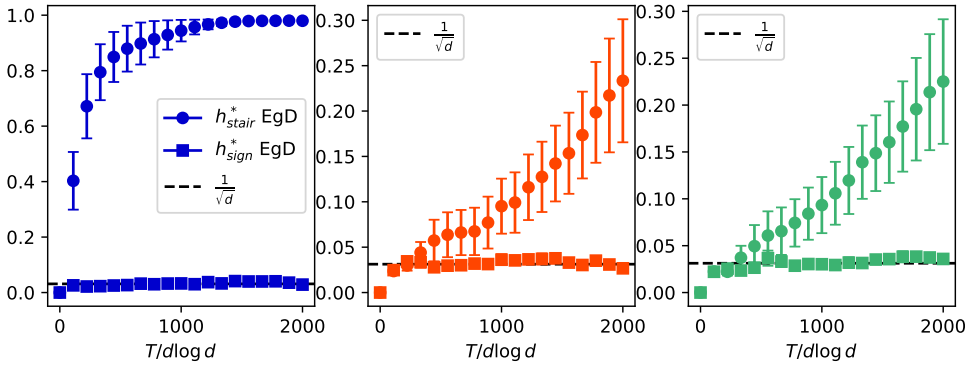
then with probability $1-\delta$, there exists a $t \leq T$ such that at least a positive proportion of the vectors $\boldsymbol{w}_{t,i}$ satisfy $\left\|\Pi^\star \boldsymbol{w}_{t,i}\right\| \geq \eta$. Further, if $\ell^\star_p = 1$, then the vectors $\{\Pi^\star \boldsymbol{w}_{t,i}\}_{i\in[p]}$ span the minimal subspace $U^\star$.

4.1 Numerical investigation with multi-index target

We illustrate the results of Theorem 0.4 numerically by comparing the performance of One-Pass SGD with Algorithm 0.1 in Figure 3. We measure the maximal cosine similarity between the hidden layer neurons and the different rows of the target's weight matrix $W^\star$ as a function of



(a) $h^\star(\boldsymbol{x}) = x_1 + x_1 \mathrm{He}_3(x_2)$



(b) $h^\star_{\text{sign}}(\boldsymbol{x}) = \text{sign}(x_1x_2x_3), \quad h^\star_{\text{stair}}(\boldsymbol{x}) = \mathrm{He}_2(x_1) + \text{sign}(x_1x_2x_3)$

Figure 2: **Hierarchical learning** – Evolution of the maximum Cosine Similarities with different target directions $\{\boldsymbol{w}_r^\star\}_{r\in[k]}$ are identified by different colors: $\boldsymbol{w}_1^\star$ (blue), $\boldsymbol{w}_2^\star$ (orange), $\boldsymbol{w}_3^\star$ (green) as a function of the normalized iteration time. **(a):** both the algorithms learn $\boldsymbol{w}_1^\star$ in linear time, but only EgD can also learn $\boldsymbol{w}_2^\star$ in $O(d)$ via a staircase mechanism; SGD requires another $O(d^2)$ steps to take advantage of the staircase and learn $\boldsymbol{w}_2^\star$. **(b):** EgD performance for the target function $f^\star_{\text{stair}}$ (dots) and $f^\star_{\text{sign}}$(squares); we plot in separate subplots the overlap with different directions to highlight the presence of hierarchical learning mechanism. Learning the first target direction $\boldsymbol{w}_1^\star$ triggers the hierarchical mechanism and EgD is able to weakly recover the full target subspace $\text{Span}(\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star, \boldsymbol{w}_3^\star)$ in $O(d \log d)$, while this does not happen when removing $\mathrm{He}_2(x_1)$ from the target. See Appendix E for additional details.

the iteration time. Furthermore, we set without loss of generality the target's directions to the standard basis of $\mathbb{R}^d$ to ease the notation. The algorithms considered are again EgD and vanilla One-Pass SGD (See Appendix E for details).

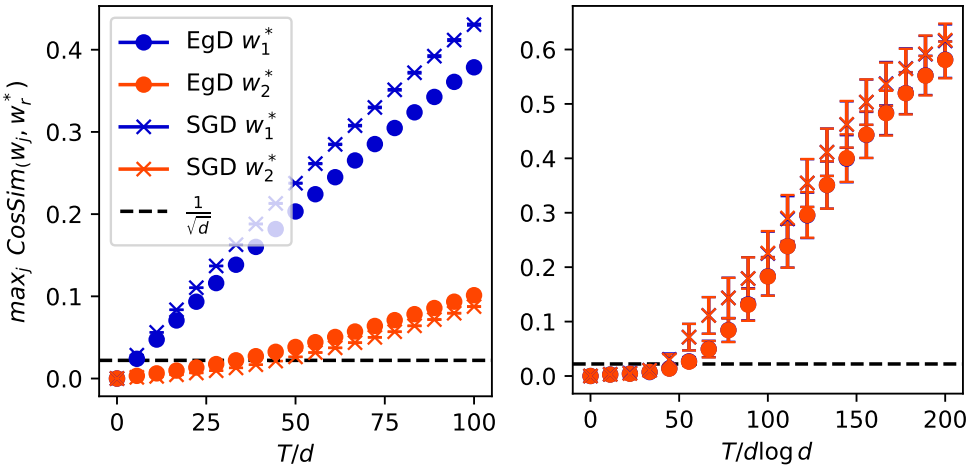
- **SGD-easy non-symmetric targets ($\ell = \ell^\star = 1$)** – Figure 3, left, shows $h^\star(x_1, x_2) = x_1 + x_1 x_2$. Here both SGD and Algorithm 0.1 learn in $T = O(d)$ the directions $(\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star)$.
- **SGD-easy symmetric targets ($\ell = \ell^\star = 2$)** – Figure 3, center left, shows $h^\star(x_1, x_2) = x_1 x_2$. Here both SGD and Algorithm 0.1 learn in $T = O(d \log d)$ the directions $(\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star)$.
- **Non-symmetric targets with SGD-easy and hard directions ($\ell_{\boldsymbol{v}} > 2, \ell_{\boldsymbol{v}}^\star = 1$)** – Figure 3, center-right, shows $h^\star(x_1, x_2) = x_1 + \mathrm{He}_3(x_2)$. Here both SGD and Algorithm 0.1 learn in $T = O(d)$ the direction $\boldsymbol{w}_1^\star$, that satisfies $\ell_{\boldsymbol{w}_1^\star} = \ell^\star_{\boldsymbol{w}_1^\star} = 1$. However, since $\ell_{\boldsymbol{w}_2^\star} = 3$ while $\ell^\star_{\boldsymbol{w}_2^\star} = 1$, One-Pass SGD suffers from the limitations detailed in Equation (9) and requires $\Omega(d^2)$ samples to learn the second direction $\boldsymbol{w}_2^\star$. This contrasts with the behavior of Algorithm 0.1 that learns also $\boldsymbol{w}_2^\star$ in $T = O(d)$ steps.
- **SGD-hard symmetric targets ($\ell > 2, \ell^\star = 2$)** – The right section Figure 3 shows $h^\star(x_1, x_2, x_3) = x_1 x_2 x_3$. The Information Exponent is $\ell = 3$, hence vanilla SGD does not learn any direction in the target subspace in $T = O(d \log d)$ iterations. However, Algorithms 0.1 learns in $T = O(d \log d)$ steps all the three directions $\{\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star, \boldsymbol{w}_3^\star\}$ since the Generative Exponents $\ell^\star_{\boldsymbol{w}_r^\star}$ are all equal to two.

4.2 Learning hard functions through a hierarchical mechanism

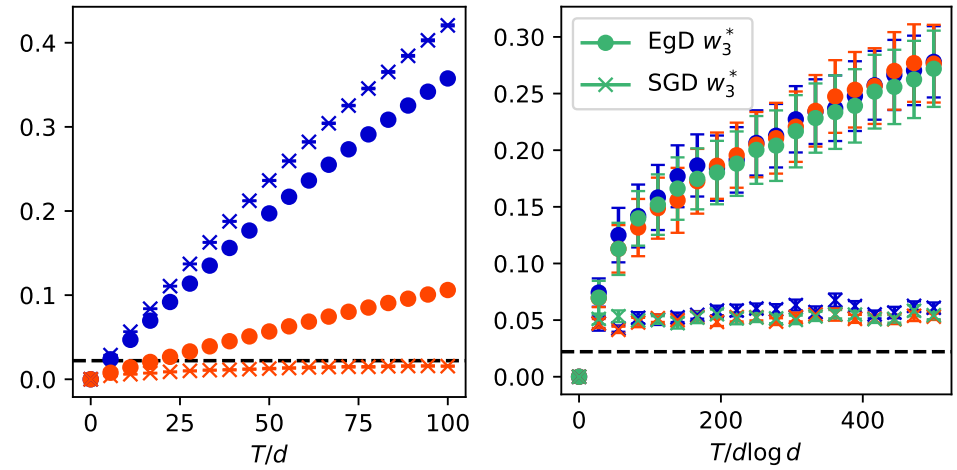
The investigation of the hierarchical nature of SGD learning has attracted noticeable attention [Abb+21; AAM22; AAM23; Dan+24a]. In this paper, we portray a completely different picture in terms of computational efficiencies when data repetition is considered in the algorithmic SGD routine. One may wonder if there is a generalization of such a hierarchical learning mechanism to the present novel setting. Indeed for optimal AMP algorithm, a different picture emerges, (called *Grand staircase* in [Tro+25]). Here we show that coupling between directions can hierarchically guide the learning process as well with SGD.

The Left panel in Figure 2 shows an example of the so-called staircase functions [Abb+21], i.e. $f^\star(\boldsymbol{z}) = z_1 + z_1 \mathrm{He}_3(z_2)$. While one-pass SGD needs to learn hierarchically first the direction $\boldsymbol{w}_1^\star$ in $T = O(d)$ and then $\boldsymbol{w}_2^\star$ in $T = O(d^2)$, Algorithm 0.1 easily learns both in linear time. This observation could lead to infer that hierarchical learning mechanisms are not present when more realistic training scenarios are considered.

To refute this, we illustrate in the right panel of Figure 2 a scenario that precisely exemplifies the presence of hierarchical mechanisms within Algorithm 0.1. We run this algorithm on two different target functions, $f^\star_{\text{sign}}(\boldsymbol{z}) = \mathrm{sign}(z_1 z_2 z_3)$ and $f^\star_{\text{stair}}(\boldsymbol{z}) = \mathrm{He}_2(z_1) + \mathrm{sign}(z_1 z_2 z_3)$. We show in Appendix B that 3-sparse parities of the form $\mathrm{sign}(z_1 z_2 z_3)$ are hard functions even for SGD with data repetition (Algorithm 0.1), and indeed they are not learned in (almost) linear time. On the other hand, for the function $f^\star_{\text{stair}}$, our simulations predict that the first direction $\boldsymbol{w}_1^\star$ is learned in $T = O(d \log d)$ iterations (see the rightmost panel of Figure 3). However, once the direction $\boldsymbol{w}_1^\star$ is learned, it can be used to obtain order-one correlation with $\{\boldsymbol{w}_2^\star, \boldsymbol{w}_3^\star\}$ again in



(a) $h^\star(\boldsymbol{x}) = x_1 + x_1x_2$ (b) $h^\star(\boldsymbol{x}) = \text{sign}(x_1x_2)$



(c) $h^\star(\boldsymbol{x}) = x_1 + \text{He}_3(x_2)$ (d) $h^\star(\boldsymbol{x}) = x_1x_2x_3$

Figure 3: **Learning multi-index targets** – Evolution of the maximum Cosine Similarities attained by EgD (crosses) and SGD (dots) as a function of the normalized iteration time. The Different directions $\{\boldsymbol{w}_r^\star\}_{r\in[k]}$ are identified by colors: $\boldsymbol{w}_1^\star$ (blue), $\boldsymbol{w}_2^\star$ (orange), $\boldsymbol{w}_3^\star$ (green). (dashed horizontal line $\frac{1}{\sqrt{d}}$ is a visual guide to place random performance). **(a)** $(\ell, \ell^\star) = (1, 1)$**:** both algorithms learn the first direction in linear time, as well as the second one in $O(d)$ steps using the staircase mechanism. **(b)** $(\ell, \ell^\star) = (2, 2)$**:** both algorithms learn the two directions simultaneously in $O(d \log d)$ steps. **(c)** $(\ell_{\boldsymbol{w}_2^\star}, \ell^\star_{\boldsymbol{w}_2^\star}) = (3, 1)$**:** both algorithms learn $\boldsymbol{w}_1^\star$ in linear time, but only EgD learns $\boldsymbol{w}_2^\star$ in $O(d)$; SGD instead requires $O(d^2)$. **(d)** $(\ell, \ell^\star) = (3, 2)$**:** EgD learns all 3 directions simultaneously in $O(d \log d)$ steps, while SGD needs $O(d^2)$ time (details in Appendix E).

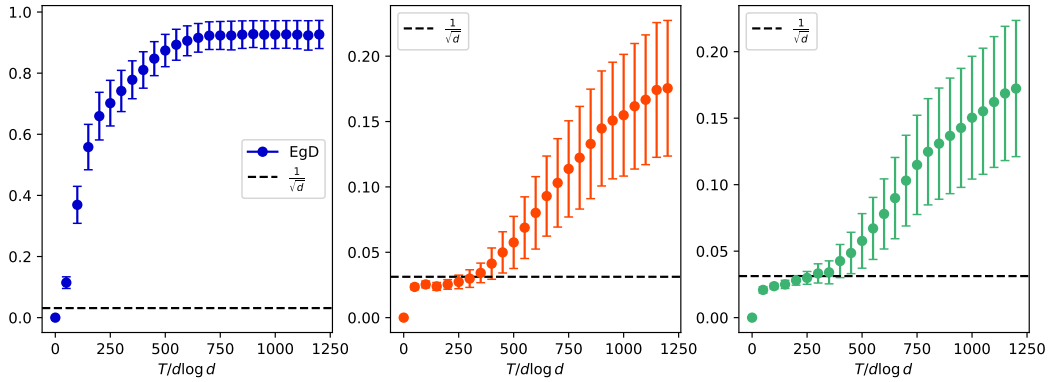


Figure 4: **Hierarchical learning** – Evolution of the maximum Cosine Similarities attained by EgD with different target directions $\{\boldsymbol{w}_r^\star\}_{r\in[k]}$ are identified by different colors: $\boldsymbol{w}_1^\star$ (blue), $\boldsymbol{w}_2^\star$ (orange), $\boldsymbol{w}_3^\star$ (green) as a function of the normalized iteration time. The target function is $f^\star_{\text{sq}-\text{stair}}(\boldsymbol{z}) = \text{He}_4(z_1) + \text{sign}(z_1z_2z_3)$, that is an SQ-staircase (or grand staircase), not a CSQ-staircase. EgD learn $\boldsymbol{w}_1^\star$ in $O(d\log d)$ steps and use the information to learn the other two directions in another $O(d\log d)$ steps. SGD (not showed in the plot) cannot take advantage of the staircase mechanism since $\text{He}_4(z_1)$ and $\text{sign}(z_1z_2z_3)$ have information exponent $\ell = 4$ and $\ell = 3$ respectively. (See Appendix E).

$T = O(d\log d)$ steps. While the latter example belongs to the class of "CSQ" staircase function depicted in [Abb+21], novel "SQ" hierarchical mechanisms arise in our framework, such as for instance the function $f^\star_{\text{sq}-\text{stair}}(\boldsymbol{z}) = \text{He}_4(z_1) + \text{sign}(z_1z_2z_3)$ (see Figure 4).

The staircase picture in terms of generative exponent was discussed for AMP-type algorithm in [Tro+25]. Our study indicate that this new class (they called grand staircase) is not restricted to AMP: We anticipate that grand staircase functions can be efficiently learned by two-layer networks when data reusing is permitted, whereas only standard staircase ones emerge under single-pass SGD. Additional recent results by [JMS24] align with this perspective.

5 Conclusions

We have pushed in this paper the boundaries of learning multi-index models in high-dimensional settings using gradient-type algorithms, and shown that, contrary to what was believed, they are very efficient in doing so. We argue in this manuscript with a simple analytically solvable model how to bypass the limitations of the CSQ framework: we process the same sample twice. Crucially, a similar phenomenon appears also in standard SGD when training over multiple epochs, with the only difference being that the same data is not seen consecutively. We believe that the model proposed in this manuscript paves the way for the analysis of more realistic training scenario where *correlation* in the data is taken into account, surpassing the limiting i.i.d. scenario.

Upon completion of this work, we became aware of the interesting paper from [Lee+24] showing strong recovery with similar sample complexity, also by exploiting the reuse of training data. Our

work was conducted independently and simultaneously

Acknowledgements

The authors would like to thank Nicolas Flammarion, Bruno Loureiro, Nati Srebro, Emanuele Troiani, and Lenka Zdeborová for interesting discussions. This work was supported by the Swiss National Science Foundation under grant SNSF OperaGOST (grant number 200390).

A Theoretical framework

A.1 Difference inequalities for vector-valued processes

We begin with a lemma that generalizes [BGJ21, Proposition 4.1]. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ be a random function, and consider the process $(\boldsymbol{u}_t)_{t\geq 0} \in \mathbb{S}^{d-1}$ given by

$$\boldsymbol{u}_{t+1} = \frac{\boldsymbol{u}_t + \gamma\Phi(\boldsymbol{u}_t)}{\left\|\boldsymbol{u}_t + \gamma\Phi(\boldsymbol{u}_t)\right\|}.$$

We make the following assumptions:

Assumption 0.6. *The initial value of $\boldsymbol{u_0} \in \mathbb{S}^{d-1}$ is approximately isotropic, in the sense that there exists a constant $c > 0$ such that*

$$\mathbb{E}\left[\boldsymbol{u_0}\boldsymbol{u_0}^\top\right] \succ \frac{c}{d} I$$

Assumption 0.7. *There exist constants $C, \iota > 0$ such that*

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \left\|\mathbb{E}\left[\Phi(\boldsymbol{u})\Phi(\boldsymbol{u})^\top\right]\right\| \leq C \tag{15}$$

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \mathbb{E}\left[\left\|\Phi(\boldsymbol{u})\right\|^{4+\iota}\right] \leq C d^{\frac{4+\iota}{2}} \tag{16}$$

Further, we have $\langle \boldsymbol{u}, \Phi(\boldsymbol{u})\rangle = 0$ a.s. for any $\boldsymbol{u} \in \mathbb{S}^{d-1}$.

We fix a vector $\boldsymbol{v}$ and consider the random process $m_t(\boldsymbol{v}) := \langle \boldsymbol{u}_t, \boldsymbol{v}\rangle$. For any $s \geq 0$, we define the stopping times

$$\tau_\theta^+(\boldsymbol{v}) = \inf\{t \geq 0 : m_t(\boldsymbol{v}) \geq \theta\} \quad \text{and} \quad \tau_\theta^-(\boldsymbol{v}) = \inf\{t \geq 0 : m_t(\boldsymbol{v}) \leq \theta\}$$

Proposition 0.1. *Fix an $\alpha > 0$. There exist constants $c(\delta, \alpha), \kappa(\delta) > 0$ such that if*

$$\gamma \leq c(\delta, \alpha) d^{-1} \quad \textit{and} \quad T \leq c(\delta, \alpha)\gamma^{-2} d^{-1}, \tag{17}$$

then, with probability $1 - \delta$, the following holds. The initialization $m_0(\boldsymbol{v})$ satisfies

$$|m_0(\boldsymbol{v})| \geq \frac{\kappa}{\sqrt{d}}, \tag{18}$$

and:

- *if* $m_0(v) > 0$, *for any* $t \le T \wedge \tau_0^-(\boldsymbol{v})$,

$$(1-\alpha)m_0(\boldsymbol{v}) + \gamma \sum_{s=0}^{t-1} \langle \Psi^+(\boldsymbol{u}_s), \boldsymbol{v}\rangle \le m_t(\boldsymbol{v}) \le (1+\alpha)m_0(\boldsymbol{v}) + \gamma \sum_{s=0}^{t-1} \langle \Psi(\boldsymbol{u}_s), \boldsymbol{v}\rangle, \tag{19}$$

- *if* $m_0(v) < 0$, *for any* $t \le T \wedge \tau_0^+(\boldsymbol{v})$,

$$(1+\alpha)m_0(\boldsymbol{v}) + \gamma \sum_{s=0}^{t-1} \langle \Psi(\boldsymbol{u}_s), \boldsymbol{v}\rangle \le m_t(\boldsymbol{v}) \le (1-\alpha)m_0(\boldsymbol{v}) + \gamma \sum_{s=0}^{t-1} \langle \Psi^-(\boldsymbol{u}_s), \boldsymbol{v}\rangle, \tag{20}$$

where $\Psi, \Psi^\pm$ *is a deterministic function equal to*

$$\Psi(\boldsymbol{u}) = \mathbb{E}\left[\Phi(\boldsymbol{u})\right] \quad \textit{and} \quad \Psi^\pm(\boldsymbol{u}) = \Psi(\boldsymbol{u}) \mp \gamma \mathbb{E}\left[\left\|\Phi(\boldsymbol{u})\right\|^2\right] \boldsymbol{u}.$$

Proof. We assume for simplicity that $m_0(\boldsymbol{v}) > 0$; the case where $m_0(\boldsymbol{v}) < 0$ is treated identically. From Assumption 0.6, by a simple Markov inequality, there exists a constant $\kappa(\delta) > 0$ such that

$$\mathbb{P}\left(|m_0(\boldsymbol{v})| \ge \frac{\kappa}{\sqrt{d}}\right) \ge 1 - \frac{\delta}{3}. \tag{21}$$

The update equation for $m_t(\boldsymbol{v})$ reads

$$m_{t+1}(\boldsymbol{v}) = \frac{m_t(\boldsymbol{v}) + \gamma\langle \Phi(\boldsymbol{u}_t), \boldsymbol{v}\rangle}{\sqrt{1+\gamma^2\left\|\Phi(\boldsymbol{u}_t)\right\|^2}}.$$

We then have two cases:

- either $m_t(\boldsymbol{v}) + \gamma\langle \Phi(\boldsymbol{u}_t), \boldsymbol{v}\rangle < 0$, which implies that $t+1 = \tau_0^-(\boldsymbol{v})$ and there is nothing to prove,
- or we can use the inequality $\frac{1}{\sqrt{1+x}} \ge 1 - x$, which implies that

$$\begin{aligned} m_{t+1}(\boldsymbol{v}) &\ge (m_t(\boldsymbol{v}) + \gamma\langle \Phi(\boldsymbol{u}_t), \boldsymbol{v}\rangle)(1 - \gamma^2\left\|\Phi(u_t)\right\|^2) \\ &= (1 - \gamma^2\left\|\Phi(u_t)\right\|^2)m_t + \gamma\langle \Phi(\boldsymbol{u}_t), \boldsymbol{v}\rangle - \gamma^3\langle \Phi(\boldsymbol{u}_t), \boldsymbol{v}\rangle\left\|\Phi(u_t)\right\|^2 \end{aligned}$$

By recursion, deterministically,

$$m_t(\boldsymbol{v}) \ge m_0(\boldsymbol{v}) + \gamma \sum_{s=0}^{t-1} \left(\langle \Phi(\boldsymbol{u}_s), \boldsymbol{v}\rangle - \gamma\left\|\Phi(\boldsymbol{u}_s)\right\|^2 \langle \boldsymbol{u}_s, \boldsymbol{v}\rangle\right) - \sum_{s=0}^{t-1} \gamma^3 \left|\langle \Phi(\boldsymbol{u}_s), \boldsymbol{v}\rangle\right| \left\|\Phi(\boldsymbol{u}_s)\right\|^2$$

Since Assumption 0.7 implies Assumption B in [BGJ21], for a small enough choice of $c(\delta, \alpha)$, one has

$$\mathbb{P}\left(\sum_{t=0}^{T} \gamma^3 \left|\langle \Phi(\boldsymbol{u}_t), \boldsymbol{v}\rangle\right| \left\|\Phi(\boldsymbol{u}_s)\right\|^2 \le \frac{\alpha\kappa}{2\sqrt{d}}\right) \ge 1 - \frac{\delta}{6}.$$

Additionnally, the same assumption implies that

$$\mathrm{Var}(\langle\Phi(\boldsymbol{u}_s),\boldsymbol{v}\rangle) \le \mathbb{E}\left[\langle\Phi(\boldsymbol{u}_s),\boldsymbol{v}\rangle^2\right] \le C_1 \quad \text{and} \quad \mathrm{Var}(\gamma\left\|\Phi(\boldsymbol{u}_s)\right\|^2) \le \gamma^2\mathbb{E}\left[\left\|\Phi(\boldsymbol{u}_s)\right\|^4\right] \le c(\delta)^2 C_1.$$

We can then apply Doob's martingale inequality: there exists a constant $C(\delta)$ such that with probability at least $1-\frac{\delta}{6}$

$$\sup_{t\le T}\left|\gamma\sum_{s=0}^{t}\left(\langle\Phi(\boldsymbol{u}_s),\boldsymbol{v}\rangle - \gamma\left\|\Phi(\boldsymbol{u}_s)\right\|^2\langle\boldsymbol{u}_s,\boldsymbol{v}\rangle - \mathbb{E}\left[\langle\Phi(\boldsymbol{u}_s),\boldsymbol{v}\rangle - \gamma\left\|\Phi(\boldsymbol{u}_s)\right\|^2\langle\boldsymbol{u}_s,\boldsymbol{v}\rangle\right]\right)\right| \le C(\delta)\gamma\sqrt{T}.$$

The bound (17) on T implies that we can choose $c(\delta,\alpha)$ small enough so that

$$C(\delta)\gamma\sqrt{T} \le \frac{C(\delta)\sqrt{c(\delta,\alpha)}}{\sqrt{d}} \le \frac{\alpha\kappa}{2\sqrt{d}}.$$

Combining all those inequalities with a union bound, under the event of Equation (21),

$$\mathbb{P}\left(m_t(\boldsymbol{v}) \ge (1-\alpha)m_0(\boldsymbol{v}) + \gamma\sum_{s=0}^{t-1}\langle\Psi^+(\boldsymbol{u}_s),\boldsymbol{v}\rangle \quad \forall t < T\wedge\tau_0^-(\boldsymbol{v})\right) \ge 1-\frac{\delta}{3}, \tag{22}$$

having used that $\langle\Psi^+(\boldsymbol{u}),\boldsymbol{v}\rangle = \mathbb{E}\left[\langle\Phi(\boldsymbol{u}),\boldsymbol{v}\rangle - \gamma\left\|\Phi(\boldsymbol{u})\right\|^2\langle\boldsymbol{u},\boldsymbol{v}\rangle\right]$.

The upper bound follows similarly: using this time the fact that for $x>0$ we have $\frac{1}{\sqrt{1+x}} \le 1$, we have under (21)

$$\mathbb{P}\left(m_t(\boldsymbol{v}) \le (1+\alpha)m_0(\boldsymbol{v}) + \gamma\sum_{s=0}^{t-1}\langle\Psi(\boldsymbol{u}_s),\boldsymbol{v}\rangle \quad \forall t < T\wedge\tau_0^-(\boldsymbol{v})\right) \ge 1-\frac{\delta}{3}. \tag{23}$$

A union bound between the events of Equations (21), (22), (23) concludes the proof. □

A.2 Low-dimensional sufficient statistics

In [BGJ21], the function $\Psi(\boldsymbol{u})$ can be written as $\Psi(\boldsymbol{u}) = \phi(\langle\boldsymbol{u},\boldsymbol{v}\rangle)\boldsymbol{v}$. As a result, the bounds of Proposition 0.1 are actually a closed function of $m_t(\boldsymbol{v})$. We generalize this phenomenon to a larger class:

Assumption 0.8. *There exists a $q>0$ such that the function Ψ is q-approximately k-dimensional, in the sense that there exist a function $\psi:\mathbb{R}^k\to\mathbb{R}^k$ and a matrix $W^\star\in\mathbb{R}^{k\times d}$ with orthonormal rows such that for any $\boldsymbol{u}\in\mathbb{R}^d$,*

$$\left\|W^\star\Psi(\boldsymbol{u}) - \psi(W^\star\boldsymbol{u})\right\| \le \left(\left\|W^\star\boldsymbol{u}\right\| \vee \frac{1}{\sqrt{d}}\right)\cdot q.$$

By an abuse of notation, we shall sometimes view ψ as a function from $V^\star$ to itself, where $V^\star$ is the span of the rows of $W^\star$. However, Assumption 0.8 ensures that all quantities related to ψ are only depending on the (oftentimes fixed) dimension k instead of d.

It turns out that in this case, we can control accurately the trajectory of $W^\star \boldsymbol{u}$, at least when f is not too flat around $\mathbf{0}$. Let $\Pi^\star = (W^\star)^\top W^\star$ be the projection operator on $V^\star$, and define the stopping time

$$\tau_\theta = \min\{t \geq 0 : \left\|\Pi^\star \boldsymbol{u}\right\| \geq \theta\}.$$

Proposition 0.2. *Assume that Condition 0.8 holds with q small enough, and that $\langle \boldsymbol{u}_0, \psi(\mathbf{0})\rangle > 0$ (in particular, $\psi(\mathbf{0}) \neq 0$). For any $\delta, \epsilon > 0$, there exist constants $c(\delta, \epsilon), \eta(\epsilon)$ such that if*

$$\gamma \leq cd^{-1} \quad \textit{and} \quad T \leq c\gamma^{-2}d^{-1}, \tag{24}$$

then

$$\mathbb{P}\left(\left\|\Pi^\star \boldsymbol{u}_t - \gamma t \psi(\mathbf{0})\right\| \leq \sqrt{k}\eta\epsilon \quad \forall t \leq T \wedge \tau_\eta\right) \geq 1 - \delta. \tag{25}$$

Proof. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ be an orthonormal basis of $V^\star$, with $\boldsymbol{v}_1$ aligned with $\psi(\mathbf{0})$. By applying Proposition 0.1 to $\delta' = \frac{\delta}{2k}$ and $\alpha = \frac{1}{2}$, with probability $1 - \frac{\delta}{2}$, for any $i \in [k]$ and $t \leq T \wedge \tau^{\pm}(\boldsymbol{v}_i)$, $m_t(\boldsymbol{v}_i)$ satisfies the inequality (19) or (20) (depending on the sign of $m_0(\boldsymbol{v_i})$). In particular, for $\boldsymbol{v}_1 = \psi(\mathbf{0})$, since we assumed that the initial sign is positive, we have

$$\frac{1}{2}\langle \boldsymbol{u}_0, \psi(\mathbf{0})\rangle + \gamma \sum_{s=0}^{t-1} \langle \Psi^+(\boldsymbol{u}_s), \psi(\mathbf{0})\rangle \leq \langle \boldsymbol{u}_t, \psi(\mathbf{0})\rangle \leq \frac{3}{2}\langle \boldsymbol{u}_0, \psi(\mathbf{0})\rangle + \gamma \sum_{s=0}^{t-1} \langle \Psi(\boldsymbol{u}_s), \psi(\mathbf{0})\rangle. \tag{26}$$

Define $C(\delta)$ such that $\langle \boldsymbol{u}_t, \psi(\mathbf{0})\rangle \leq C(\delta)d^{-1/2}$ with probability at least $1 - \frac{\delta}{2k}$. By a Taylor expansion, there exists an $\eta(\epsilon) > 0$ such that for any $\boldsymbol{v} \in \mathbb{R}^d$ with $\|\boldsymbol{v}\| \leq \eta, \left\|\psi(\boldsymbol{v}) - \psi(\mathbf{0})\right\| \leq \epsilon'$, where ϵ' will be fixed later. With Assumption 0.8, this implies that for any $\boldsymbol{u} \in \mathbb{R}^d$,

$$\left\|\psi(\mathbf{0})\right\|^2 - (\epsilon' + \eta q)\left\|\psi(\mathbf{0})\right\| \leq \langle \Psi(\boldsymbol{u}), \psi(\mathbf{0})\rangle \leq \left\|\psi(\mathbf{0})\right\|^2 + (\epsilon' + \eta q)\left\|\psi(\mathbf{0})\right\|.$$

Further, by Jensen's inequality, $\mathbb{E}\left[\left\|\Phi(\boldsymbol{u})\right\|^2\right] \leq C_1 d$ so by choosing $c(\delta, \epsilon)$ small enough, we can ensure that

$$\langle \Psi^+(\boldsymbol{u}), \psi(\mathbf{0})\rangle \geq \left\|\psi(\mathbf{0})\right\|^2 - (2\epsilon + \eta q)\left\|\psi(\mathbf{0})\right\|$$

Plugging those estimates in (26), we find that for $t \leq T \wedge \tau_0^-(\boldsymbol{v_1}) \wedge \tau_\eta$, and d large enough so that

$$C(\delta)d^{-1/2} \leq \frac{\eta\epsilon}{3}\left\|\psi(\mathbf{0})\right\|,$$

we have

$$\gamma t\left(\left\|\psi(\mathbf{0})\right\|^2 - (2\epsilon' + \eta q)\left\|\psi(\mathbf{0})\right\|\right) \leq \langle \boldsymbol{u}_t, \psi(\mathbf{0})\rangle \leq \frac{\eta\epsilon}{2}\left\|\psi(\mathbf{0})\right\| + \gamma t\left(\left\|\psi(\mathbf{0})\right\|^2 + (\epsilon' + \eta q)\left\|\psi(\mathbf{0})\right\|\right). \tag{27}$$

For small enough ϵ and q, the LHS of the above inequality is always at least $\gamma t\left\|\psi(0)\right\|^2/2$, which implies that

$$\tau_0^-(\boldsymbol{v}_1) \geq \tau_\eta \quad \text{and} \quad T \wedge \tau_\eta \leq \frac{2\eta}{\gamma\left\|\psi(\mathbf{0})\right\|} \tag{28}$$

Choosing ϵ' and η small enough so that

$$\frac{(2\epsilon' + \eta q)}{\|\psi(\mathbf{0})\|^2} \leq \frac{\epsilon}{2}$$

we ensure that for $t \leq T \wedge \tau_\eta$ we have

$$\left|\langle \boldsymbol{u_t} - \gamma t \psi(\mathbf{0}), \psi(\mathbf{0})\rangle\right| \leq \epsilon\eta \|\psi(\mathbf{0})\|. \tag{29}$$

We now move on to the other $\boldsymbol{v}_i$. For $i \geq 2$, $\boldsymbol{v}_i$ is orthogonal to $\psi(\mathbf{0})$, and this time Assumption 0.8 implies

$$\left|\langle \Psi(\boldsymbol{u}), \boldsymbol{v}_i\rangle\right| \leq \epsilon + \eta q.$$

Regardless of the sign of $m_0(\boldsymbol{v}_i)$, this implies that for $t \leq T \wedge \tau_\eta$,

$$|\langle \boldsymbol{u_t}, \boldsymbol{v}_i\rangle| \leq \frac{3}{2}|\langle \boldsymbol{u}_0, \boldsymbol{v}_i\rangle| + \gamma t(\epsilon' + \eta q) \leq \frac{\eta\epsilon}{2} + \frac{2\eta(2\epsilon' + \eta q)}{\|\psi(\mathbf{0})\|},$$

having used the bound (28) on τ_η and bounding the initial value as above with probability $1 - \frac{\delta}{2k}$. We can always choose η, ϵ' such that the above bound is at most ϵ; since

$$\|\Pi_V \boldsymbol{u}_t - \gamma t \psi(\mathbf{0})\|^2 = \frac{\langle \boldsymbol{u_t} - \gamma t\psi(\mathbf{0}), \psi(\mathbf{0})\rangle^2}{\|\psi(\mathbf{0})\|^2} + \sum_{i=2}^{k} \langle \boldsymbol{u_t}, \boldsymbol{v}_i\rangle^2 \leq k\epsilon^2,$$

this concludes the proof. □

Proposition 0.3. *Assume that $\psi(\mathbf{0}) = \mathbf{0}$, and $H_\psi := d\psi(\mathbf{0}) \neq 0$ is symmetric. Let $\bar{\lambda}$ be the largest eigenvalue of H_ψ, and assume that $\bar{\lambda} > 0$. For any $\delta, \epsilon > 0$, there exist constants $c(\delta, \epsilon), \eta(\epsilon), c'(\delta, \epsilon), C(\delta, \epsilon)$ such that if*

$$\gamma \leq cd^{-1} \quad \textit{and} \quad T \leq c\gamma^{-2}d^{-1}, \tag{30}$$

then

$$\mathbb{P}\left(\|\Pi^\star \boldsymbol{u}_t\| \geq c'\|\Pi^\star \boldsymbol{u}_0\| e^{\gamma(\bar{\lambda} - Ck(\epsilon+q))t} \quad \forall t \leq T \wedge \tau_\eta\right) \geq 1 - \delta. \tag{31}$$

Further, the mass of $\Pi^\star \boldsymbol{u}_t$ is concentrated on the top eigenspace of H_ψ: there exists a $\tilde{\lambda} < \bar{\lambda} - Ck(\epsilon + q)$ such that on the same event,

$$\|\Pi^\star \boldsymbol{u}_t - \bar{\Pi}^\star \boldsymbol{u}_t\| \leq \|\Pi^\star \boldsymbol{u}_0\| e^{\gamma\tilde{\lambda}t} \quad \forall t \leq T \wedge \tau_\eta,$$

where $\bar{\Pi}^\star$ is the projection on the eigenspace associated to $\bar{\lambda}$.

Proof. Fix $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ an orthonormal basis of $V^\star$, consisting of eigenvectors of H_ψ, and let λ_i be the eigenvalue associated with $\boldsymbol{v}_i$. The largest eigenvalue of H_ψ that differs from $\bar{\lambda}$ will be denoted $\underline{\lambda}$ (we can let $\underline{\lambda} = -\infty$ if $H_\psi = \bar{\lambda} I$).

We choose an $\eta(\epsilon)$ such that for $\boldsymbol{v} \in \mathbb{R}^k$ such that $\|\boldsymbol{v}\| \leq \eta$,

$$\left\|\psi(\boldsymbol{v}) - H_\psi \boldsymbol{v}\right\| \leq \epsilon \|\boldsymbol{v}\| \,.$$

In particular, for any $i \in [k]$ and any vector $\boldsymbol{u}$, we have for small enough choice of $c(\delta, \epsilon)$

$$\langle \Psi^+(\boldsymbol{u}), \boldsymbol{v}_i \rangle \geq \lambda_i \langle \boldsymbol{v}_i, \boldsymbol{u} \rangle - (2\epsilon + q)\left\|\Pi^\star \boldsymbol{u}\right\| \quad \langle \Psi(\boldsymbol{u}), \boldsymbol{v}_i \rangle \leq \lambda_i \langle \boldsymbol{v}_i, \boldsymbol{u} \rangle + (\epsilon + q)\left\|\Pi^\star \boldsymbol{u}\right\|$$

Applying Proposition 0.1 to $\alpha = \epsilon$, and assuming that $m_0(\boldsymbol{v}_i) > 0$, we find that on an event with probability $1 - \delta/(2k)$, for $t \leq T \wedge \tau_0^+(\boldsymbol{v_i}) \wedge \tau_\eta$,

$$\langle \boldsymbol{v}_i, \boldsymbol{u}_t \rangle \geq (1-\epsilon) m_0(\boldsymbol{v}_i) + \gamma \sum_{s=0}^{t-1} \left(\lambda_i \langle \boldsymbol{v}_i, \boldsymbol{u}_s \rangle - (2\epsilon + q)\left\|\Pi^\star \boldsymbol{u}_s\right\|\right)$$

$$\langle \boldsymbol{v}_i, \boldsymbol{u}_t \rangle \leq (1+\epsilon) m_0(\boldsymbol{v}_i) + \gamma \sum_{s=0}^{t-1} \left(\lambda_i \langle \boldsymbol{v}_i, \boldsymbol{u}_s \rangle + (\epsilon + q)\left\|\Pi^\star \boldsymbol{u}_s\right\|\right)$$

The same argument when $m_0(\boldsymbol{v}_i) < 0$, we obtain that with probability $1 - \delta/2$, for any $t \leq T \wedge \tau_\eta$ and $i \in [k]$:

$$|\langle \boldsymbol{v}_i, \boldsymbol{u}_t \rangle| \geq (1-\epsilon)|\langle \boldsymbol{v}_i, \boldsymbol{u}_0 \rangle| + \gamma \sum_{s=0}^{t-1} \left(\lambda_i |\langle \boldsymbol{v}_i, \boldsymbol{u}_s \rangle| - (2\epsilon + q)\left\|\Pi^\star \boldsymbol{u}_s\right\|\right) \tag{32}$$

$$|\langle \boldsymbol{v}_i, \boldsymbol{u}_t \rangle| \leq (1+\epsilon)|\langle \boldsymbol{v}_i, \boldsymbol{u}_0 \rangle| + \gamma \sum_{s=0}^{t-1} \left(\lambda_i |\langle \boldsymbol{v}_i, \boldsymbol{u}_s \rangle| + (\epsilon + q)\left\|\Pi^\star \boldsymbol{u}_s\right\|\right) \tag{33}$$

Define $\bar{\Pi}^\star$ the projection on the eigenspace associated to $\bar{\lambda}$, and $\underline{\Pi}^\star$ its orthogonal projection. Summing the upper bound of (33) on all v_i with $\lambda_i \leq \underline{\lambda}$,

$$\left\|\underline{\Pi}^\star \boldsymbol{u}_t\right\|_1 \leq (1+\epsilon)\left\|\underline{\Pi}^\star \boldsymbol{u}_0\right\|_1 + \gamma(\underline{\lambda} + k(\epsilon + q)) \sum_{s=0}^{t-1} \left\|\Pi^\star \boldsymbol{u}_s\right\|_1 \,,$$

where the ℓ^1 norm $\|\cdot\|_1$ in $V^\star$ is computed w.r.t the basis $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$, and we used the norm bound $\|\Pi^\star \boldsymbol{u}_s\| \leq \|\Pi^\star \boldsymbol{u}_s\|_1$.

Call $\mathcal{P}(t_0)$ the following condition:

$$\left\|\underline{\Pi}^\star \boldsymbol{u}_t\right\|_1 \leq C\left\|\bar{\Pi}^\star \boldsymbol{u}_t\right\|_1 \quad \text{for all } t < t_0, , \tag{34}$$

where C will be fixed later, and assume that $\mathcal{P}(t_0)$ holds for some $t_0 > 0$. Using the discrete Grönwall inequality, the following holds: for any $t \leq T \wedge \tau_\eta \wedge t_0$,

$$\left\|\underline{\Pi}^\star \boldsymbol{u}_t\right\|_1 \leq (1+\epsilon)\left\|\underline{\Pi}^\star \boldsymbol{u}_0\right\|_1 e^{\gamma(\underline{\lambda} + (1+C)k(\epsilon+q))t}. \tag{35}$$

We now sum the lower bound of (32) on all v_i with $\lambda_i = \bar{\lambda}$, to obtain

$$\left\|\bar{\Pi}^\star \boldsymbol{u}_t\right\|_1 \geq (1-\epsilon)\left\|\bar{\Pi}^\star \boldsymbol{u}_0\right\|_1 + \gamma \sum_{s=0}^{t-1} (\bar{\lambda} - k(2\epsilon + q))\left\|\bar{\Pi}^\star \boldsymbol{u}_s\right\|_1 - k(2\epsilon + q)\left\|\underline{\Pi}^\star \boldsymbol{u}_s\right\|_1 \,.$$

above bound implies that for all $t \le t_0$,

$$\left\|\bar{\Pi}^\star \boldsymbol{u}_t\right\|_1 \ge (1-\epsilon)\left\|\bar{\Pi}^\star \boldsymbol{u}_0\right\|_1 + \gamma \sum_{s=0}^{t-1} (\bar{\lambda} - (1+C)k(2\epsilon+q))\left\|\bar{\Pi}^\star \boldsymbol{u}_s\right\|_1 .$$

By the discrete Grönwall inequality,

$$\begin{aligned}
\left\|\bar{\Pi}^\star \boldsymbol{u}_t\right\|_1 &\ge (1-\epsilon)\left\|\bar{\Pi}^\star \boldsymbol{u}_0\right\|_1 \left(1+\gamma(\bar{\lambda} - (1+C)k(2\epsilon+q))\right)^t \\
&\ge (1-\epsilon)\left\|\bar{\Pi}^\star \boldsymbol{u}_0\right\|_1 \exp\left(\gamma(\bar{\lambda} - (1+C)k(2\epsilon+q))t - O(\gamma^2 T)\right) \\
&\ge (1-2\epsilon)\left\|\bar{\Pi}^\star \boldsymbol{u}_0\right\|_1 \exp\left(\gamma(\bar{\lambda} - (1+C)k(2\epsilon+q))t\right), \qquad (36)
\end{aligned}$$

where the last inequality holds for large enough d because of the bound (17). Now, fix a constant $r(\delta)$ such that with probability $1-\delta/2$, $\left\|\bar{\Pi}^\star \boldsymbol{u}_0\right\|_1 \le r(\delta)\|\underline{\Pi}^\star \boldsymbol{u}_0\|_1$. Then condition $\mathcal{P}(1)$ holds as long as $C > r(\delta)$, and $\mathcal{P}(t_0)$ implies $\mathcal{P}(t_0+1)$ whenever the lower bound of eq. (36) is higher than the upper bound of eq. (35) for $t \le t_0$. In particular, this holds whenever

$$C \ge \frac{1+\epsilon}{1-2\epsilon} r(\delta) \quad \text{and} \quad \bar{\lambda} - (1+C)k(2\epsilon+q) \ge \underline{\lambda} + (1+C)k(\epsilon+q).$$

The above condition can be satisfied if and only if

$$\frac{1+\epsilon}{1-2\epsilon} r(\delta) < \frac{\bar{\lambda} - \underline{\lambda}}{3\epsilon + 2q} - 1,$$

which can always be satisfied when ϵ, q are low enough. The proposition ensues from the remark that $\left\|\bar{\Pi}^\star \boldsymbol{u}_t\right\| \le \|\Pi^\star \boldsymbol{u}\| \le (1+C)\left\|\bar{\Pi}^\star \boldsymbol{u}_t\right\|$ for any $t \le T \wedge \tau_\eta$. □

A.3 Bounding the hitting times

It finally remains to show how the bounds of Propositions 0.2 and 0.3 allow us to bound the hitting times τ_η. We summarize those in the following proposition:

Proposition 0.4. *Let $\boldsymbol{u}$ be a process that satisfies Assumptions 0.6-0.8 for low enough q. Then for any $\delta, \epsilon > 0$, there exist constants $\gamma_0(\delta,\epsilon), \eta(\epsilon), \kappa(\delta), K(\delta), C(\delta,\epsilon)$ such that, with probability $1-\delta$:*

(i) if $\psi(\mathbf{0}) \neq \mathbf{0}$ and $\langle \boldsymbol{u}_0, \psi(\mathbf{0})\rangle > 0$, letting $\gamma = \gamma_0 d^{-1}$,

$$\frac{\eta(1-\sqrt{k}\epsilon)}{\gamma\left\|\psi(\mathbf{0})\right\|} \le \tau_\eta \le \frac{\eta(1+\sqrt{k}\epsilon)}{\gamma\left\|\psi(\mathbf{0})\right\|}$$

(ii) if $\psi(\mathbf{0}) = 0$ and $d\psi(\mathbf{0})$ is symmetric and has a positive maximum eigenvalue $\bar{\lambda}$, then letting $\gamma = \gamma_0(d\log(d))^{-1}$

$$\frac{\log\left(\frac{\eta\sqrt{d}}{(1+\epsilon)K}\right)}{\gamma(\bar{\lambda} + k(\epsilon+q))} \le \tau_\eta \le \frac{\log\left(\frac{\eta\sqrt{d}}{\kappa c'}\right)}{\gamma(\bar{\lambda} - Ck(\epsilon+q))}$$

Further, the directions of $\Pi^\star \boldsymbol{u}_{\tau_\eta}$ *concentrate, in the following sense:*

(i) if $\psi(\mathbf{0}) \neq \mathbf{0}$,

$$\left\| \frac{\Pi^\star \boldsymbol{u}_{\tau_\eta}}{\eta} - \frac{\psi(\mathbf{0})}{\|\psi(\mathbf{0})\|} \right\| \leq 2\sqrt{k}\epsilon$$

(ii) if $\psi(\mathbf{0}) = \mathbf{0}$, *for some* $\beta < 1$,

$$\|\Pi^\star \boldsymbol{u}_{\tau_\eta} - \bar{\Pi}^\star \boldsymbol{u}_{\tau_\eta}\| \leq C(\delta, \epsilon) d^{-\frac{1-\beta}{2}},$$

where $\bar{\Pi}^\star$ *is the projection on the top eigenspace of* $d\psi(\mathbf{0})$.

Proof. The proof for the linear case is almost immediate. From Proposition 0.2, with probability $1 - \delta$, we have for all $t \leq T \wedge \tau_\eta$

$$\gamma t \|\psi(\mathbf{0})\| - \sqrt{k}\eta\epsilon \leq \|\Pi^\star \boldsymbol{u}_t\| \leq \gamma t \|\psi(\mathbf{0})\| + \sqrt{k}\eta\epsilon$$

Define

$$\bar{t}_\epsilon = \frac{\eta(1 + \sqrt{k}\epsilon)}{\gamma\|\psi(\mathbf{0})\|} \quad \text{and} \quad \underline{t}_\epsilon = \frac{\eta(1 - \sqrt{k}\epsilon)}{\gamma\|\psi(\mathbf{0})\|}.$$

For $t \leq \underline{t}_\epsilon$, the upper bound on $\|\Pi^\star \boldsymbol{u}_t\|$ is lower than η, which directly means that $\tau_\eta \geq \underline{t}_\epsilon$. The lower bound is higher than η for $t = \bar{t}_\epsilon$, which implies that $\tau_\eta \leq \bar{t}_\epsilon$ as long as $T \geq \bar{t}_\epsilon$. Such a T satisfies the bound (17) as long as

$$\frac{\eta(1 + \sqrt{k}\epsilon)}{\gamma\|\psi(\mathbf{0})\|} \leq c(\delta, \epsilon)\gamma^{-2}d^{-1},$$

which is equivalent to

$$\gamma \leq \frac{c(\delta, \epsilon)\|\psi(\mathbf{0})\|}{\eta(1 + \sqrt{k}\epsilon)} d^{-1},$$

which can always be ensured by decreasing $\gamma_0(\delta, \epsilon)$. Finally, we use the bound

$$\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \right\| \leq 2\frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\|\boldsymbol{x}\|}$$

to conclude that

$$\left\| \frac{\Pi^\star \boldsymbol{u}_{\tau_\eta}}{\eta} - \frac{\psi(\mathbf{0})}{\|\psi(\mathbf{0})\|} \right\| \leq 2\sqrt{k}\epsilon.$$

We now move on to the quadratic case. By Proposition 0.3 and the same reasoning as in the proof of (35), we have with probability $1 - \delta$, for all $t \leq T \wedge \tau_\eta$,

$$c'\|\Pi^\star \boldsymbol{u}_0\| e^{\gamma(\bar{\lambda} - Ck(\epsilon+q))t} \leq \|\Pi^\star \boldsymbol{u}_t\| \leq (1 + \epsilon)\|\Pi^\star \boldsymbol{u}_0\| e^{\gamma(\bar{\lambda} + k(\epsilon+q))t}$$

Upon redefining δ, we can assume that the "good" event implies that

$$\frac{\kappa(\delta)}{\sqrt{d}} \le \left\|\Pi^\star \boldsymbol{u}_0\right\| \le \frac{K(\delta)}{\sqrt{d}}$$

for constants k, K. As in the linear case, the upper bound above implies that

$$\tau_\eta \ge \frac{\log\left(\frac{\eta\sqrt{d}}{(1+\epsilon)K}\right)}{\gamma(\bar{\lambda} + k(\epsilon+q))},$$

while the lower bound implies that

$$T \wedge \tau_\eta \le \frac{\log\left(\frac{\eta\sqrt{d}}{\kappa c'}\right)}{\gamma(\bar{\lambda} - Ck(\epsilon+q))} =: \bar{t}'_\epsilon.$$

It therefore remains to show that $\bar{t}'_\epsilon \le T$; for large enough d and small enough ϵ, q, we can write $\bar{t}'_\epsilon \le C'(\delta,\epsilon)\gamma^{-1}\log(d)$, and

$$C'(\delta,\epsilon)\gamma^{-1}\log(d) \le c\gamma^{-2}d^{-1} \Leftrightarrow \gamma \le c(C'd\log(d))^{-1}.$$

This condition can again be ensured by decreasing $\gamma_0(\delta,\epsilon)$ as needed. Finally, let

$$\beta = \frac{\tilde{\lambda}}{\bar{\lambda} - Ck(\epsilon+q)},$$

where $\tilde{\lambda}$ is the one of Proposition 0.3. Then $\beta < 1$ and we have at $t = \tau_\eta$

$$\left\|\Pi^\star \boldsymbol{u}_{\tau_\eta} - \bar{\Pi}^\star \boldsymbol{u}_{\tau_\eta}\right\| \le \frac{K}{\sqrt{d}} \exp\left(\beta \log\left(\frac{\eta\sqrt{d}}{\kappa c'}\right)\right) \le C''(\delta,\epsilon) d^{-\frac{1-\beta}{2}}.$$

□

B Proof of the Main Results

B.1 Proof of Theorem 0.2

We formalize here the proof sketch of Theorem 0.2. We track the dynamics of $\boldsymbol{w}_t$ through the following sufficient statistic:

$$m_t := \langle \boldsymbol{w}_t, \boldsymbol{w}^\star \rangle. \tag{37}$$

The dynamics of m_t fall under the framework of Section A, by letting

$$\Phi(\boldsymbol{w}) = -\nabla^\perp_{\boldsymbol{w}} \mathcal{L}\left(f(\boldsymbol{z}; \tilde{\boldsymbol{w}}(\rho), a_0), y\right).$$

Indeed, for the spherical gradient,

$$\nabla^\perp_{\boldsymbol{w}} \mathcal{L}(f(\boldsymbol{z}; \boldsymbol{w}, a_0), y) = a_0 y \sigma'(\langle \boldsymbol{w}, z\rangle) \cdot \left(\boldsymbol{z} - \langle \boldsymbol{z}, \boldsymbol{w}\rangle \boldsymbol{w}\right).$$

, hence plugging in the expression for $\tilde{\boldsymbol{w}}(\rho)$

$$\Phi(\boldsymbol{w}) = a_0 h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)\sigma'\left(\langle \boldsymbol{w}, \boldsymbol{z}\rangle + a_0\rho h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\cdot\|\boldsymbol{z}\|^2\right)\left(\boldsymbol{z} - \langle \boldsymbol{z}, \boldsymbol{w}\rangle \boldsymbol{w}\right). \tag{38}$$

Assumptions 0.1 and 0.2 on σ and $h^\star$ then directly imply Assumptions 0.6 and 0.7 in the Appendix. As a result, we only need to study the quantity $\Psi(\boldsymbol{w}) = \mathbb{E}\left[\Phi(\boldsymbol{w})\right]$. We show the following:

Lemma 0.2. *The function Ψ is approximately one-dimensional: there exists a function ϕ such that for any $\boldsymbol{w} \in \mathbb{S}^{d-1}$, we have*

$$|\langle \boldsymbol{w}^\star, \Psi(\boldsymbol{w})\rangle - a_0\phi(\boldsymbol{w}, \boldsymbol{w}^\star)| \le \frac{C}{d}$$

Proof. Define the following quantities:

$$\Xi(x, x^\star, x^\perp) = h^\star(x^\star)\sigma'(x + a_0\rho_0\sigma'(x)h^\star(x^\star))x^\perp \tag{39}$$

$$\phi(m) = \mathbb{E}[\Xi(x, x^\star, x^\perp)] \text{ for } \begin{pmatrix} x \\ x^\star \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} 1 & m \\ m & 1 \end{pmatrix}\right) \text{ and } x^\perp = x^\star - mx. \tag{40}$$

For simplicity, we let $u = \langle \boldsymbol{w}, \boldsymbol{z}\rangle + a_0\rho_0\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)$. Doing a Taylor expansion of σ' in the second order, we have

$$\begin{aligned}\sigma'\left(\langle \boldsymbol{w}, \boldsymbol{z}\rangle + \rho h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\cdot\|\boldsymbol{z}\|^2\right) =& \sigma'(u) + \sigma''(u)\rho h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\cdot\left(\|\boldsymbol{z}\|^2 - d\right) \\ &+ R(u)\rho^2 h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)^2\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)^2\cdot\left(\|\boldsymbol{z}\|^2 - d\right)^2,\end{aligned}$$

where $R(u)$ is the Taylor remainder of σ' at second order around u.

The distribution of the triplet $(\langle \boldsymbol{w}, \boldsymbol{z}\rangle, \langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle, \langle \boldsymbol{w}^\star, \boldsymbol{z}^\perp\rangle)$ is the same as the one of $(x, x^\star, x^\perp)$, hence

$$\begin{aligned}\mathbb{E}\left[\langle \Phi(\boldsymbol{z}), \boldsymbol{w}^\star\rangle\right] - a_0\phi(m) =& \underbrace{\rho\mathbb{E}\left[\sigma''(u)h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)^2\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\cdot\left(\|\boldsymbol{z}\|^2 - d\right)\langle \boldsymbol{w}^\star, \boldsymbol{z}^\perp\rangle\right]}_{(1)} \\ &+ \underbrace{\rho^2\mathbb{E}\left[R(u)h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)^3\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)^2\cdot\left(\|\boldsymbol{z}\|^2 - d\right)^2\langle \boldsymbol{w}^\star, \boldsymbol{z}^\perp\rangle\right]}_{(2)}.\end{aligned}$$

We first bound (2). From the Hölder inequality (since $\frac{1}{2} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = 1$),

$$(2) \le \rho^2\mathbb{E}\left[R(u)^8\right]^{\frac{1}{8}}\mathbb{E}\left[h^\star(x^\star)^8\right]^{\frac{1}{8}}\mathbb{E}\left[\sigma'(x)^8\right]^{\frac{1}{8}}\mathbb{E}\left[\langle \boldsymbol{w}^\star, \boldsymbol{z}^\perp\rangle^8\right]^{\frac{1}{8}}\mathbb{E}\left[\left(\|\boldsymbol{z}\|^2 - d\right)^4\right]^{\frac{1}{2}}$$

Since σ is three times differentiable almost everywhere, and from the polynomial bounds of Assumption 0.4, the first four expectations are bounded by an absolute constant. The last term is

the centered fourth moment of a $\chi^2(d)$ distribution, which is equal to $12d(d+4)$. As a result, using that $\rho^2 = O(d^{-2})$,

$$(2) \leq Cd^{-1}.$$

For the first term, we let $\boldsymbol{z} = \boldsymbol{z}_\parallel + \boldsymbol{z}'$, where $\boldsymbol{z}_\parallel$ is the projection of $\boldsymbol{z}$ on the span of $\boldsymbol{w}$ and $\boldsymbol{w}^\star$. We can then write

$$\begin{aligned}(1) = &\underbrace{\mathbb{E}\left[\sigma''(u)\rho h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)^2\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\cdot\left\|\boldsymbol{z}_\parallel\right\|^2\cdot\langle \boldsymbol{w}^\star, \boldsymbol{z}^\perp\rangle\right]}_{(1')}\\ &+\underbrace{\mathbb{E}\left[\sigma''(u)\rho h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)^2\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\cdot\left(\left\|\boldsymbol{z}'\right\|^2 - d\right)\langle \boldsymbol{w}^\star, \boldsymbol{z}^\perp\rangle\right]}_{(1'')}.\end{aligned}$$

Up to the factor of ρ, the term $(1')$ is the expectation of a polynomially bounded function of a 2-dimensional Gaussian variable, hence

$$(1') \leq C\rho \leq C'd^{-1}.$$

In $(1'')$, the vector $\boldsymbol{z}'$ is independent from $\langle \boldsymbol{w}, \boldsymbol{z}\rangle$ and $\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle$, so the expectation factorizes as

$$\begin{aligned}|(1'')| &= \left|\mathbb{E}\left[\sigma''(u)\rho h^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}\rangle)^2\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\langle \boldsymbol{w}^\star, \boldsymbol{z}^\perp\rangle\right]\cdot\mathbb{E}\left[\left(\left\|\boldsymbol{z}'\right\|^2 - d\right)\right]\right|\\ &\leq C\rho|d-2-d| \leq C'd^{-1},\end{aligned}$$

where we used that $\boldsymbol{z}'$ is a $(d-2)$-dimensional Gaussian variable. This concludes the proof. □

Lemma 0.2 implies that Assumption 0.8 holds for Ψ and any choice of q, as long as d is large enough. As a result, the proof of Theorem 0.2 hinges on the following facts about ϕ:

Proposition 0.5. *For almost any choice of ρ_0 (w.r.t to the Lebesgue measure),*

- *if $\ell_p^\star = 1$, then $\phi(0) \neq 0$;*
- *if $\ell_p^\star = 2$, then $\phi'(0) \neq 0$.*

of Theorem 0.2. Our goal is to apply Proposition 0.4 to the function ϕ. If $\ell_p^\star = 1$, then $\phi(0) \neq 0$ by the previous lemma, and with a uniform initialization of $\boldsymbol{w}^\star$ we have $m_0\phi(0) > 0$ with probability $1/2$. Combining this with the $(1-\delta)$ probability event of Proposition 0.4 for small enough ϵ, with probability $1/2 - \delta$, we have $\tau_\eta \leq C(\delta)d$. We now move to the proof of the case where $\ell_p^\star = 2$. In this case, Proposition 0.4 implies the desired result as soon as $\phi'(0) > 0$ (which corresponds in the $k = 1$ setting to the only eigenvalue of $d\phi(0)$ being positive). Since the Lebesgue measure is invariant w.r.t multiplication by ± 1, the factor $a_0\rho_0$ in the definition of $\phi'(0)$ has the same distribution as ρ_0, hence we can write $\phi'(0) = a_0'|\phi'(0)|$ where a_0' is also uniform in $\{-1, 1\}$. The event $\phi'(0) > 0$ therefore also has probability $1/2$, and we conclude as before. □

Analysis of ϕ: proof of Proposition 0.5

The main ingredient here is to write $\phi(m) = \phi(m; \rho_0)$ and differentiate w.r.t ρ_0. For simplicity, we take $a_0 = 1$ in the definition of ϕ. We show the following:

Lemma 0.3. *Under Assumption 0.4, the function $\phi(m; \rho_0)$ is analytic in ρ_0. Further, we have*

$$\left.\frac{\partial^k \phi(m;\rho_0)}{\partial \rho_0^k}\right|_{\rho_0=0} = \mathbb{E}\left[h^\star(x^\star)^{k+1} x^\perp \sigma^{(k+1)}(x)\sigma'(x)^k\right] \tag{41}$$

where $x, x^\star$ follow the same distribution as in eq. (40).

Proof. The first statement stems from the analyticity of σ, and the fact that it is a property conserved through integration. For the second, we differentiate inside the expectation, to find

$$\frac{\partial \phi}{\partial \rho_0} = \mathbb{E}\left[\sigma'(x)h^\star(x^\star)^2\sigma''(x+\rho_0\sigma'(x)h^\star(x^\star))x^\perp\right];$$

the result follows by induction and taking $\rho_0 = 0$ at the end. □

We define the following quantities related to Assumption 0.2:

$$u_0^{(k)} = \mathbb{E}\left[\sigma^{(k)}(x)\sigma'(x)^{k-1}\right] \qquad u_1^{(k)} = \mathbb{E}\left[x\sigma^{(k)}(x)\sigma'(x)^{k-1}\right]$$
$$v_0^{(k)} = \mathbb{E}\left[x^\star h^\star(x^\star)^k\right] \qquad v_1^{(k)} = \mathbb{E}\left[[(x^\star)^2-1]h^\star(x^\star)^k\right]$$
$$\Delta^{(k)} = \mathbb{E}\left[h^\star(x^\star)^k\right]$$

Then, by Proposition 11.37 from [ODo14], and using $x^\perp = x^\star - mx$:

$$\begin{aligned}\left.\frac{\partial^k \phi(m;\rho_0)}{\partial \rho_0^k}\right|_{\rho_0=0} &= \mathbb{E}\left[h^\star(x^\star)^{k+1}x^\star\sigma^{(k+1)}(x)\sigma'(x)^k\right] - m\cdot\mathbb{E}\left[h^\star(x^\star)^{k+1}x\sigma^{(k+1)}(x)\sigma'(x)^k\right]\\ &= \left(u_0^{(k)}v_0^{(k)} + u_1^{(k)}(v_1^{(k)}+\Delta^{(k)})\cdot m + o(m)\right) - m\left(u_1^{(k)}\Delta^{(k)} + o(1)\right)\\ &= u_0^{(k)}v_0^{(k)} + u_1^{(k)}v_1^{(k)}\cdot m + o(m)\end{aligned}$$

This allows us to show Proposition 0.5:

Proof. Assume first that $\ell_p^\star = 1$. Then $\phi(0;\rho_0)$ is an analytic function of ρ_0 with

$$\left.\frac{\partial^k \phi(0;\rho_0)}{\partial \rho_0^k}\right|_{\rho_0=0} = u_0^{(k)}v_0^{(k)}.$$

Since $\ell_p^\star = 1$, there exists a $k \in \mathbb{N}$ such that $v_0^{(k)} \neq 0$, and by Assumption 0.4 the coefficient $u_0^{(k)}$ is nonzero for every k. As a result, $\phi(0;\rho_0)$ is an analytic and non-identically zero function

of ρ_0, so it is non-zero for almost every choice of ρ_0, as requested. The case $\ell_p^\star = 2$ is done in a similar way, noting that this time

$$\left.\frac{\partial^k \phi'(0;\rho_0)}{\partial \rho_0^k}\right|_{\rho_0=0} = u_1^{(k)} v_1^{(k)}.$$

□

B.2 Proof of Theorem 0.4

The proof of Theorem 0.4 follows in the same way as the one of Theorem 0.2. Recall the definition of Φ in eq. (38), and define

$$\Xi(z, \boldsymbol{z}^\star, \boldsymbol{z}^\perp) = h^\star(\boldsymbol{z}^\star)\sigma'(z + a_0\rho_0\sigma'(z)h^\star(\boldsymbol{z}^\star))\boldsymbol{z}^\perp \tag{42}$$

$$\psi(\boldsymbol{p}) = \mathbb{E}\left[\Xi(z, \boldsymbol{z}^\star, \boldsymbol{z}^\perp)\right] \text{ for } \begin{pmatrix} z \\ \boldsymbol{z}^\star \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} 1 & \boldsymbol{p} \\ \boldsymbol{p} & I_k \end{pmatrix}\right) \text{ and } \boldsymbol{z}^\perp = \boldsymbol{z}^\star - z\boldsymbol{p}. \tag{43}$$

Again, the variables $(z, \boldsymbol{z}^\star)$ correspond to $\langle \boldsymbol{w}, \boldsymbol{z}\rangle$, $W^\star \boldsymbol{z}$, and the vector $\boldsymbol{p}$ to $W^\star \boldsymbol{w}$. As before, the only difference between $\mathbb{E}\left[\Phi(\boldsymbol{w})\right]$ and $\psi(W^\star \boldsymbol{w})$ lies in replacing $\|\boldsymbol{z}\|$ with d, and the proof of Lemma 0.2 implies the following: for any $\boldsymbol{w} \in \mathbb{S}^{d-1}$,

$$\left\|W^\star \Psi(\boldsymbol{w}) - \psi(W^\star \boldsymbol{w})\right\| \le \frac{C}{d}.$$

As a result, the function Ψ is approximately k-dimensional for any choice of q as $d \to \infty$.

We can view the vector $\psi(\boldsymbol{p})$ as a random variable depending on the choice of a_0, ρ_0. Since a_0 is a sign, we will only consider the dependency of ψ in ρ_0. As before, let $u_i^{(k)}$ be the i-th Hermite coefficient of $z \mapsto \sigma^{(k+1)}(z)\sigma'(z)^k$, and $C_i^{(k)}$ the i-th Hermite tensor of $h^\star$ [Gra49] – in particular, $C_i^{(k)}$ is a tensor of order i. We have the analogous result to Lemma 0.3:

Lemma 0.4. *Under Assumption 0.4, the function $\psi(\boldsymbol{p};\rho_0)$ is analytic in ρ_0, with*

$$\left.\frac{\partial^k \psi(\boldsymbol{p};\rho_0)}{\partial \rho_0^k}\right|_{\rho_0=0} = u_0^{(k)} C_1^{(k)} + u_1^{(k)} C_2^{(k)} \boldsymbol{p} + o(\|\boldsymbol{p}\|).$$

Proof. As in Lemma 0.3, we can write $\psi(\boldsymbol{0}) = \psi(\boldsymbol{0}, \rho_0)$, with

$$\left.\frac{\partial^k \psi(\boldsymbol{p};\rho_0)}{\partial \rho_0^k}\right|_{\rho_0=0} = \mathbb{E}\left[h^\star(\boldsymbol{z}^\star)^{k+1} \boldsymbol{z}^\perp \sigma^{(k+1)}(z)\sigma'(z)^k\right].$$

From [Dan+24a, Lemma 19], we have

$$\mathbb{E}\left[h^\star(\boldsymbol{z}^\star)^{k+1} \boldsymbol{z}^\star \sigma^{(k+1)}(z)\sigma'(z)^k\right] = \sum_{i=0}^{\infty} u_{i+1}^{(k)} \langle \boldsymbol{p}^{\otimes i}, C_i^{(k)}\rangle \boldsymbol{p} + \sum_{i=0}^{\infty} u_i^{(k)} C_{i+1}^{(k)} \times_{1,\dots,i} \boldsymbol{p}^{\otimes i}.$$

Since $\boldsymbol{z}^\perp = \boldsymbol{z} - z\boldsymbol{p}$, we find

$$\left.\frac{\partial^k \psi(\boldsymbol{p};\rho_0)}{\partial \rho_0^k}\right|_{\rho_0=0} = \sum_{i=0}^{\infty} u_i^{(k)} \left(C_{i+1}^{(k)} \times_{1,\dots,i} \boldsymbol{p}^{\otimes i} - \langle C_{i+1}^{(k)}, \boldsymbol{p}^{\otimes(i+1)} \rangle \boldsymbol{p} \right). \tag{44}$$

For $i \geq 0$, we have $\left\| C_{i+1}^{(k)} \times_{1,\dots,i} \boldsymbol{p}^{\otimes i} \right\| = O\left(\|\boldsymbol{p}\|^i \right)$ and $\langle C_i^{(k)}, \boldsymbol{p}^{\otimes(i)} \rangle = O\left(\|\boldsymbol{p}\|^i \right)$. The only terms in the sum which are not of order $\|\boldsymbol{p}\|^2$ are the first two terms in the left sum, which yields the statement of the Lemma. □

We are now ready to show the first part of Theorem 0.4. We begin with a characterization of $U^\star$ whenever $\ell_p^\star = 1$:

Lemma 0.5. *If $\ell_p^\star = 1$, then*

$$U^\star = \operatorname{span}\left(\left\{ C_1^{(k)} \right\}_{k\geq 1} \right).$$

If $\ell_p^\star = 2$, then

$$U^\star = \bigoplus_{k\geq 1} \operatorname{Im}\left(C_2^{(k)} \right).$$

Proof. By the properties of the Hermite tensors, we have for any $v \in V^\star$ and any polynomial p

$$\mathbb{E}\left[y^k H_1(\langle \boldsymbol{v}, \boldsymbol{z} \rangle) \right] = \langle \boldsymbol{v}, C_1^{(k)} \rangle$$

As a result, if $\boldsymbol{v} = \sum_{k\geq 0} v_k C_1^{(k)}$, then by letting $p(y) = \sum_{k\geq 0} a_k y^k$ we have

$$\mathbb{E}\left[p(y) H_1(\langle \boldsymbol{v}, \boldsymbol{z} \rangle) \right] = \|\boldsymbol{v}\|^2 \neq 0.$$

Conversely, if $\boldsymbol{v}$ is orthogonal to every vector $C_1^{(k)}$, then $\mathbb{E}\left[y^k H_1(\langle \boldsymbol{v}, \boldsymbol{z} \rangle) \right] = 0$ for all $k \geq 0$, which implies that $\boldsymbol{v} \notin U^\star$.

The case $\ell_p^\star = 2$ is similar, since this time

$$\mathbb{E}\left[y^k H_2(\langle \boldsymbol{v}, \boldsymbol{z} \rangle) \right] = \boldsymbol{v}^\top C_2^{(k)} \boldsymbol{v}.$$

As a result, for fixed k, the largest vector space U_k such that $\mathbb{E}\left[y^k H_2(\langle \boldsymbol{v}, \boldsymbol{z} \rangle) \right]$ for all $\boldsymbol{v} \in U_k$ is $\ker(C_2^{(k)}) = \operatorname{Im}(C_2^{(k)})^\perp$. The result ensues from $U^\star = \bigcap_{k\geq 1} U_k^\perp$. □

Using Assumption 0.4, the following corollary ensues:

Corollary 0.1. *If $\ell_p^\star = 1$, then for any $\boldsymbol{v} \in U^\star$ and almost every ρ_0,*

$$\langle \boldsymbol{v}, \psi(\mathbf{0}) \rangle \neq 0.$$

As a result, with probability one, the set $\{\psi(\mathbf{0}; \rho_{0,i})\}_{i\in[p]}$ spans $U^\star$ as soon as $p \geq k$.

If $\ell_p^\star = 2$, then with probability at least $1/2$ on the choice of ρ_0, a_0, $d\psi(\mathbf{0})$ is symmetric and has a strictly positive eigenvalue.

Proof. Let $\boldsymbol{v} \in U^\star$, and $k \geq 1$ such that $\langle \boldsymbol{v}, C_1^{(k)} \rangle \neq 0$. Then

$$\left.\frac{\partial^k \langle \boldsymbol{v}, \psi(\mathbf{0}; \rho_0)\rangle}{\partial \rho_0^k}\right|_{\rho_0=0} = u_0^{(k)} \langle \boldsymbol{v}, C_1^{(k)} \rangle \neq 0,$$

so the function $\rho_0 \mapsto \langle \boldsymbol{v}, \psi(\mathbf{0}; \rho_0)\rangle$ is not the zero function. Since it is also analytic, it is nonzero for almost every choice of ρ.

We now prove that if $p \leq \dim(U^\star)$, then the family $\{\psi(\mathbf{0}; \rho_{0,i})\}_{i\in[p]}$ is free, which implies our second result. Assume by recursion that $\{\psi(\mathbf{0}; \rho_{0,i})\}_{i\in[p-1]}$ is free, and let $\boldsymbol{v}_p \in U^\star$ orthogonal to this family. Then with probability one, $\langle \boldsymbol{v}_p, \psi(\mathbf{0}; \rho_{0,p})\rangle \neq 0$, so $\psi(\mathbf{0}; \rho_{0,p})$ is independent from $\{\psi(\mathbf{0}; \rho_{0,i})\}_{i\in[p-1]}$.

Finally, if $\ell_p^\star = 2$, then by symmetry of the distribution of ρ_0 we can write $\psi(\boldsymbol{p}; a_0, \rho_0) = a_0\psi(\boldsymbol{p}; 1, \rho')$ with $\rho' = a_0\rho_0$ chosen independently from a_0. Writing

$$d\psi(\mathbf{0}; \rho') = \sum_{k\geq 1} (\rho')^k u_1^{(k)} C_2^{(k)},$$

we can see that $d\psi(\mathbf{0})$ is a symmetric matrix, which is non-zero with probability one on ρ'. Since either $d\psi(\mathbf{0})$ or $-d\psi(\mathbf{0})$ has a positive eigenvalue, $a_0 d\psi(\mathbf{0})$ does with probability $1/2$. □

We are now ready to prove Theorem 0.4.

of Theorem 0.4. We begin with the case $\ell_p^\star = 1$. Fix a $p = p(\delta)$ to be chosen later, and let $\psi_i(\boldsymbol{p}) = \psi(\boldsymbol{p}; a_{0,i}, \rho_{0,i})$ for $i \in [p]$. With probability $1/2$ on $\boldsymbol{w}_{0,i}$, we have $\langle \boldsymbol{w}_{0,i}, \psi_i(\mathbf{0})\rangle$; let p_+ be the number of such neurons. For $p = O(\log(1/\delta))$, we have $p_+ > \max(p/4, k)$ with probability at least $1 - \frac{\delta}{2}$.

In particular, with probability one, the set $\{\psi_i(\mathbf{0})\}_{i\in[p_+]}$ spans $U^\star$. Let $\epsilon > 0$ to be chosen later, $\eta_i(\epsilon)$ adapted to ψ_i in Proposition 0.4, τ_i the hitting time for η_i in the i-th neuron. Consider the time $\tau = \min_i \tau_i$; applying Proposition 0.4 to $\delta' = \delta/p$, we have

$$\tau \leq \max_{i\in[p_+]} \frac{\eta_i(1+\sqrt{k}\epsilon)}{\gamma_0 \|\psi_i(\mathbf{0})\|} d \leq C(\delta) d.$$

However, we also have $\tau > c(\delta)\gamma^{-1}$, and hence from Proposition 0.2, for any $i \in [p_+]$

$$\|\boldsymbol{w}_{t,i}\| \geq \gamma\tau \|\psi_i(\mathbf{0})\| - \sqrt{k}\eta_i\epsilon \geq c'(\delta)$$

for small enough ϵ. This proves the first part of the theorem. Now, consider the vectors

$$\boldsymbol{u}_i = \frac{\Pi^\star \boldsymbol{w}_{\tau,i}}{\left\|\Pi^\star \boldsymbol{w}_{\tau,i}\right\|};$$

by the same argument as in Proposition 0.4, for any $i \in [p_+]$,

$$\left\| \boldsymbol{u}_i - \frac{\psi_i(\mathbf{0})}{\left\|\psi_i(\mathbf{0})\right\|} \right\| \leq \sqrt{k}\epsilon.$$

Since the $\{\psi_i(\mathbf{0})\}_{i\in[p_+]}$ span $U^\star$ with probability one, so do the $\{\boldsymbol{u}_i\}_{i\in[p_+]}$ for a small enough choice of ϵ.

For the case $\ell_p^\star = 2$, we define the p_+ "good" neurons as those with $d\psi_i(\mathbf{0})$ having a positive eigenvalue. Since this happens with probability at least $1/2$ independently for each neuron, the bounds on $p(\delta)$ from earlier still apply. For those neurons, and small enough ϵ, we define the $\eta_i(\epsilon)$ as those in Proposition 0.4, and $\tau = \max_i \tau_i$. Then by Proposition 0.4, we have $\tau \leq C(\delta) d \log(d)^2$ for some constant $C(\delta)$.

Now, fix a vector $i \in [p_+]$. We define the j-th "excursion" to be the j-th sequence of times $\tau_i^{(j)}, \dots, \tau_i^{(j)} + e_i^{(j)}$ such that $\left\|\Pi^\star \boldsymbol{w}_{t,i}\right\| \geq \eta_i$. Then:

- With probability at least $1 - e^{-c\gamma}$, we have $\left\|\boldsymbol{w}_{t+1,i} - \boldsymbol{w}_{t,i}\right\| \leq C\sqrt{\gamma}$, and thus for any excursion j we have $\left\|\Pi^\star \boldsymbol{w}_{\tau_i^{(j)}+e_i^{(j)}+1,i}\right\| \geq \eta_i/2$.
- between two excursions, the results of Proposition 0.3 apply, and thus the norm of $\left\|\Pi^\star \boldsymbol{w}_{t,i}\right\|$ increases exponentially.

As a result, for $\tau_i \leq t \leq \tau$, we have $\left\|\Pi^\star \boldsymbol{w}_{t,i}\right\| > \eta_i/2$, which finishes the proof of Theorem 0.4. □

C Additional proofs

C.1 Proof of Lemma 0.1

Let σ be an analytic function which is not a polynomial. Then for any $k \geq 0$, the function $\sigma^{(k)}$ is also analytic and non-identically zero, and hence so is the function $f_n = \sigma^{(k)} \cdot (\sigma')^k$. Lemma 0.1 thus ensues from the following result:

Lemma 0.6. *Let f be an analytic and non-identically zero function. Then*

$$\mathbb{E}\left[f(x+b)\right] \neq 0 \quad \textit{and} \quad \mathbb{E}\left[xf(x+b)\right] \neq 0$$

for almost every b under the Lebesgue measure.

Proof. The function $\psi(b) = \mathbb{E}\left[f(x+b)\right]$ is analytic in b, and we have

$$\psi^{(k)}(0) = \mathbb{E}\left[f^{(k)}(x)\right] = c_k u_k(f),$$

where $u_k(f)$ is the k-th Hermite coefficient of f and c_k is a nonzero absolute constant. Since f is nonzero, at least one of its Hermite coefficients is nonzero, and ψ is a non-identically zero analytic function. As a result, $\psi(b) \neq 0$ for almost every choice of b.

The other case is handled as the first one, noting that by Stein's lemma

$$\mathbb{E}\left[xf(x+b)\right] = \mathbb{E}\left[f'(x+b)\right].$$

□

The above shows that for fixed $n \in \mathbb{N}$, the set

$$\mathcal{B}_n = \left\{ b \in \mathbb{R} \,:\, \mathbb{E}\left[\sigma^{(n)}(x+b)\sigma'(x+b)^{n-1}\right] = 0 \quad \text{or} \quad \mathbb{E}\left[x\sigma^{(n)}(x+b)\sigma'(x+b)^{n-1}\right] = 0 \right\}$$

has measure 0. Since there is a countable number of such subsets, the set

$$\mathcal{B} = \bigcup_{n\in\mathbb{N}} \mathcal{B}_n$$

also has measure zero, which is equivalent to the statement of Lemma 0.1.

C.2 Proof of Theorem 0.3

We decompose the polynomial $h^\star$ as an even and odd part $e(x)$ and $o(x)$, with degrees d_o and d_e. We first assume that o is non-zero, and that the leading coefficient of o is positive. Then, for odd $m \geq 0$, we have

$$\mathbb{E}\left[xh^\star(x)^m\right] = \sum_{k=0}^{m} \binom{m}{k} \mathbb{E}\left[xe(x)^k o(x)^{m-k}\right]$$

Each term in the above sum is zero if k is odd, we focus on the terms where k is even. There exist constants $A, \varepsilon > 0$ such that if $|x| \geq A$,

$$|e(x)| \geq \varepsilon |x|^{d_e} \quad \text{and} \quad |o(x)| \geq \varepsilon |x|^{d_o},$$

and we let

$$B = \sup_{x\in[-A,A]} |e(x)| \vee \sup_{x\in[-A,A]} |o(x)|$$

As a result, when k is even, both $e(x)^k$ and $xo(x)^{m-k}$ are even polynomials with positive leading coefficients, so

$$\begin{aligned}
\mathbb{E}\left[xe(x)^k o(x)^{m-k}\right] &= \mathbb{E}\left[xe(x)^k o(x)^{m-k}\mathbf{1}_{x\in[-A,A]}\right] + \mathbb{E}\left[xe(x)^k o(x)^{m-k}\mathbf{1}_{x\notin[-A,A]}\right] \\
&\geq \varepsilon^2 \mathbb{E}\left[x^{kd_e+(m-k)d_o+1}\mathbf{1}_{x\notin[-A,A]}\right] - AB^m
\end{aligned}$$

Let $d(m,k) = kd_e + (m-k)d_o + 1$. Then,

$$\mathbb{E}\left[xe(x)^k o(x)^{m-k}\right] \geq \varepsilon^2 \mathbb{E}\left[x^{d(m,k)}\right] - A^{d(m,k)} - AB^m.$$

Going back to the sum, and with the crude bound $\binom{m}{k} \leq 2^m$,

$$\begin{aligned}\mathbb{E}\left[xh^\star(x)^m\right] &\geq \varepsilon^2\mathbb{E}\left[x^{d(m,0)}\right] - \sum_{k=1}^{m} 2^m\left(A^{d(m,k)} + AB^m\right)\\ &\geq \mathbb{E}\left[x^{m+1}\right] - C^m\end{aligned}$$

Since the Gaussian moments grow faster than any power of m, this last expression is non-zero for a large enough choice of m.

Finally, if o is the zero polynomial, we can do the same reasoning with $e(x)^m$ to get

$$\mathbb{E}\left[(x^2-1)e(x)^m\right] = \varepsilon^2(\mathbb{E}\left[x^{md_e+2}\right] - \mathbb{E}\left[x^{md_e}\right]) - C^m = \varepsilon^2 m d_e \mathbb{E}\left[x^{md_e}\right] - C^m,$$

and we conclude as before.

C.3 Hardness of sign functions

We show in the appendix the following statement:

Proposition 0.6. *Let $m \in \mathbb{N}$, and*

$$h^\star(\boldsymbol{x}) = \text{sign}(x_1 \dots x_m)$$

Then the function $h^\star$ has Information and Generative Exponents $\ell = \ell^\star = m$.

We first show the lower bound. For any vector $\boldsymbol{v} \in \mathbb{R}^m$, and $k < m$, $H_k(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)$ is a polynomial in z of degree k, and therefore each monomial term is missing at least one variable. We show that this implies the lower bound of Proposition 0.6 through the following lemma:

Lemma 0.7. *Let $f : \{-1, 1\} \to \mathbb{R}$ be an arbitrary function, and $g : \mathbb{R}^m \to \mathbb{R}$ a function which is independent from x_m. Then*

$$\mathbb{E}\left[f(h^\star(\boldsymbol{x}))g(\boldsymbol{x})\right] = \frac{f(1)+f(-1)}{2}\mathbb{E}\left[g(\boldsymbol{x})\right]$$

Proof. We simply write

$$\begin{aligned}\mathbb{E}\left[f(h^\star(\boldsymbol{x}))g(\boldsymbol{x})\right] &= \mathbb{E}\left[\mathbb{E}\left[f(h^\star(\boldsymbol{x}))g(\boldsymbol{x}) \mid x_1, \dots, x_{m-1}\right]\right]\\ &= \mathbb{E}\left[g(\boldsymbol{x})\mathbb{E}\left[f(h^\star(\boldsymbol{x})) \mid x_1, \dots, x_{m-1}\right]\right]\\ &= \mathbb{E}\left[g(\boldsymbol{x})\frac{f(1)+f(-1)}{2}\right]\\ &= \frac{f(1)+f(-1)}{2}\mathbb{E}\left[g(\boldsymbol{x})\right],\end{aligned}$$

where at the second line we used that g only depends on $x_1, \dots, x_{m-1}$ and at the third line the addition of x_m randomly flips the sign of $h^\star$. □

Using this property on every monomial in $H_k(\langle \boldsymbol{v}, \boldsymbol{x} \rangle)$ and summing back, we have for any function f

$$\mathbb{E}\left[f(h^\star(x))H_k(\langle \boldsymbol{v}, \boldsymbol{x} \rangle)\right] = \frac{f(1)+f(-1)}{2}\mathbb{E}\left[H_k(\langle \boldsymbol{v}, \boldsymbol{x} \rangle)\right] \tag{45}$$

But the latter is zero when $k \geq 1$ by orthogonality of the Hermite polynomials. This shows that $\ell^\star_{\boldsymbol{v}} \geq m$ for all $\boldsymbol{v} \in \mathbb{R}^m$, and therefore $m \leq \ell^\star \leq \ell$.

For the upper bound, let $\boldsymbol{v} = \boldsymbol{e}_1 + \cdots + \boldsymbol{e_m}$. Then

$$H_m(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) = m! x_1 \dots x_m + Q(\boldsymbol{x}),$$

where Q is a polynomial where each monomial has a missing variable. As a result,

$$\begin{aligned}
\mathbb{E}\left[h^\star(\boldsymbol{x})H_m(\langle \boldsymbol{v}, \boldsymbol{x} \rangle)\right] &= m!\mathbb{E}\left[h^\star(\boldsymbol{x})x_1 \dots x_m\right] \\
&= m!\mathbb{E}\left[|x_1 \dots x_m|\right] \\
&= m!\left(\sqrt{\frac{2}{\pi}}\right)^m
\end{aligned}$$

which is a non-zero value, hence $\ell^\star \leq \ell \leq m$.

D Additional Numerical Investigation

In this section, we present further numerical investigation to support our theory and extend it beyond the mathematical hypotheses used in the formal proof.

D.1 Testing wider range of Algorithm 0.1

In the main we presented all the numerical simulation using $\rho > 0$, i.e. *ExtraGradient method-/descent*. In Figure 5 Left, we repeat one of the experiment of Figure 1 in the case $\rho < 0$, the *Sharpness-Aware Minimization*. The two algorithms are based on very different principle, but in our context they behave in the same way: the reusing of the data makes high information exponent function easily learnable. The global picture for $\rho \neq 0$ is the same regardless of its sign.

Although Algorithm 0.1 includes a broad variety of algorithms, it is not the most general domain of validity of our theory. Simple algorithms such as repeating the gradient step twice on the same data before discarding it (also known as *2-lookahead optimizer* [Zha+19]) are not part of Algorithm 0.1, yet our considerations are still valid. In fact, the data repetition is sufficient to introduce correlations across two consecutive steps, whose effect builds up along the dynamics causing the network to learn hard targets. A simple illustration of this claim is available in Figure 5 Right.

D.2 Extensive batch size

In this section, we want to present some evidence that our result does not change when the batch size is $1 < n_b \leq O(d^{\frac{\ell^\star}{2}})$. The formal proof we presented in the paper is valid in the case where the batch size is $n_b = 1$, but we strongly believe it is the case when training with larger batches.

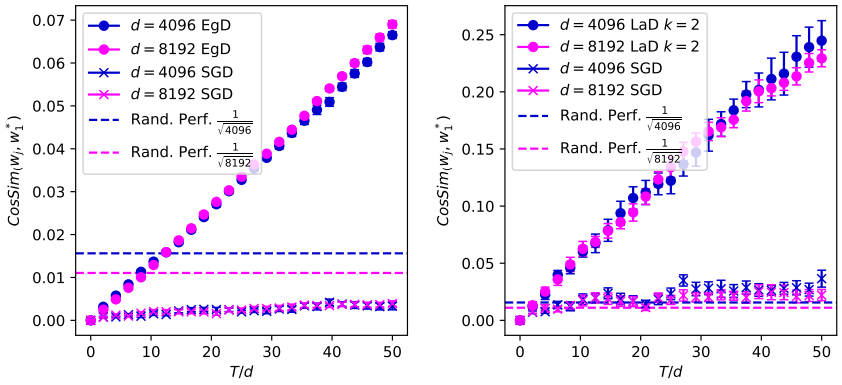


Figure 5: example of two different algorithm with data repetitions learning an hard single-index target $h^\star(\boldsymbol{z}) = \mathrm{He}_3(z_1)$, $\ell = 3, \ell^\star = 1$. **Left**: SAM, with $\rho_0 = 0.1, \gamma_0 = 0.01$. **Right**: 2-Lookahed with $\gamma_0 = 0.1$. See details in Appendix E.

In the context of the information exponent, it has been shown in [Arn+24] that the total sample complexity does not change when using batch sizes $n_b \leq O\left(d^{\frac{\ell}{2}}\right)$. The heuristic argument for this behavior is that using a larger batch size increases the learning rate since the noise of the gradient is "more" averaged out, ultimately reducing the number of steps. The same argument should pass from information exponent to generative exponent, allowing the claim (when $n_b \leq O\left(d^{\frac{\ell^\star}{2}}\right)$)

$$T \cdot n_b \sim \begin{cases} O(d^{\ell^\star - 1}) & \text{if } \ell^\star > 2 \\ O(d \log d) & \text{if } \ell^\star = 2 \\ O(d) & \text{if } \ell^\star = 1 \end{cases} . \tag{46}$$

Note that this drastically reduces the time steps needed to learn the teacher's directions, although the sample complexity does not change. [Dan+24b] first shows that some high-information exponent functions, such as He_3, can be learned in just a few steps with full-batch gradient descent $n_b = O(d)$. In the same work, they introduce a larger class of hard functions that cannot be recovered in $T = O(1)$, not providing any information on possible bounds on the number of time steps, nor any claim suggesting that not all the function in the class have the same sample complexity. Our theory can provide a full understanding of the phenomena:

- He_3 has generative exponent $\ell^\star = 1$, that implies we need $O(d)$ samples to weakly recover the target; when training with batch size $n_b = O(d)$, we have $T = O(1)$ as in [Dan+24b]; Figure 6 (Left) is a numerical proof of this fact.
- He_4 has generative exponent $\ell^\star = 2$, that implies we need $O(d \log d)$ samples to weakly recover the target; when training with batch size $n_b = O(d)$, we have $T = O(\log d)$; Figure 6 (Right) is a numerical proof of this fact.

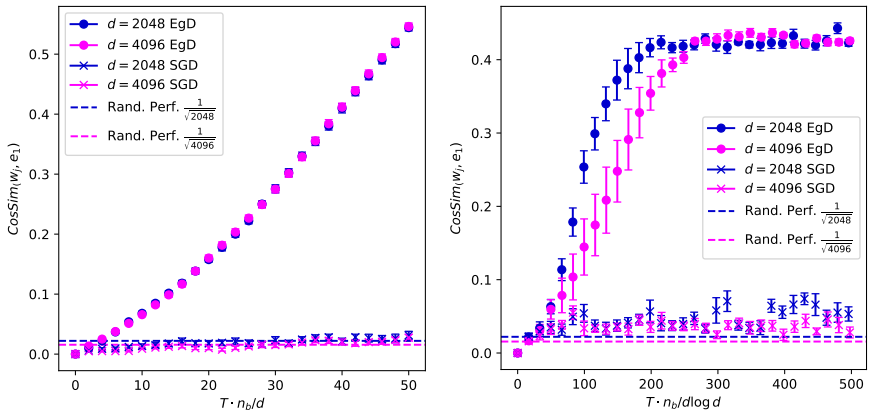


Figure 6: Single-index recovery for large-batch training. The dashed horizontal line $\frac{1}{\sqrt{d}}$ is a visual guide to place random performance. The plots show $\sigma = \text{relu}$, $\gamma = 0.01$, $\rho = 0.1$, and the performance is computed averaging across 10 different runs. **Left** $(\ell, \ell^\star) = (3, 1)$**:** $h^\star = \text{He}_3$. **Right** $(\ell, \ell^\star) = (4, 2)$**:** $h^\star = \text{He}_4$. See details in Appendix E.

We can push our claims beyond single-index targets and illustrate the same phenomena for multi-index functions. In Figure 7 we test Extragradient Descent against two hard multi-index functions: once again the sample complexity is the same as the case $n_b = 1$, while the number of steps is reduced by a factor $n_b = d$.

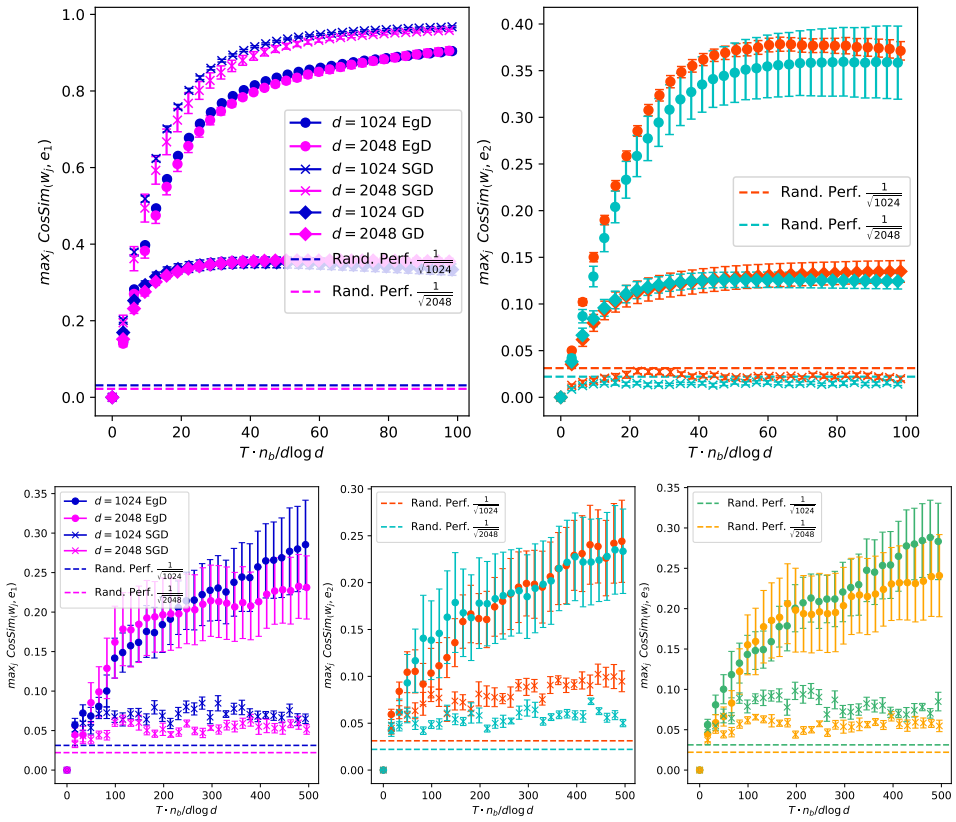


Figure 7: Multi-index recovery for large-batch training. The dashed horizontal line $\frac{1}{\sqrt{d}}$ is a visual guide to place random performance. The plots show $\sigma = \mathrm{relu}$. **Upper** $(\ell_{\boldsymbol{v}}, \ell^\star_{\boldsymbol{v}}) = (3, 1)$**:** $h^\star(\boldsymbol{z}) = \mathrm{sign}(z_1 z_2)$. **Lower** $(\ell_{\boldsymbol{v}}, \ell^\star_{\boldsymbol{v}}) = (4, 2)$**:** $h^\star(\boldsymbol{z}) = z_1 z_2 z_3$. See Appendix E for details.

E Implementation Details

In this section, we provide all the implementation details needed to reproduce the Figures of the paper. The full implementation can be found in the repository linked in the main paper.

Unless explicitly stated, we always run the plain version of Algorithm 0.1 with squared loss, as opposed to Assumption 0.3 which requires *correlation loss* and spherical gradient. We use the reLU function as an activation. Finally, we always run simulations with $W^0 \perp W^\star$, which does not fall under Assumption 0.2. Indeed, our numerical simulations deviate from the theoretical assumptions needed for the formal proofs, but the aim here is to show that our claims are valid beyond the technical limitations of the theory and that they can allow us to understand settings closer to what is used in practice.

In the plots with a single-index target we plot the *absolute value* of the cosine similarity between the student weight $\boldsymbol{w}$ and the teacher weight $\boldsymbol{w}^\star$, averaged across N different runs (where we vary both seen samples and initial conditions). If q in the index for different runs:

$$\text{yaxis}(t) = \frac{1}{N}\sum_{q=1}^{N}\left|\text{CosSim}(\boldsymbol{w}_t^{(q)}, \boldsymbol{w}^{\star(q)})\right|,$$

where

$$\text{CosSim}(\boldsymbol{w}, \boldsymbol{w}^\star) = \frac{\boldsymbol{w}\cdot\boldsymbol{w}^\star}{\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2}.$$

In the plots with a multi-index target, we take the maximum across all the network weights

$$\text{yaxis}(t) = \frac{1}{N}\sum_{q=1}^{N}\max_{j\in[p]}\left|\text{CosSim}(\boldsymbol{w}_{j,t}^{(q)}, \boldsymbol{w}^{\star(q)})\right|.$$

E.1 Hyperparameters of the Figures

Here are the hyperparameters used in the figures:

- **Figure 1 (all 4)**: $d = 8192, \gamma_0 = 0.01, \rho_0 = 0.1, N = 40$.
- **Figure 3 (left)**: $p = 8, d = 2048, \gamma_0 = 0.01, \rho_0 = 0.1, N = 40, a_j^0 \sim \text{Rad}(1/2)$.
- **Figure 3 (center-left)**: $p = 8, d = 2048, \gamma_0 = 0.01, \rho_0 = 0.1, N = 40, a_j^0 \sim \mathcal{N}(0,1)$.
- **Figure 3 (center-right)**: $p = 8, d = 2048, \gamma_0 = 0.1, \rho_0 = 0.1, N = 40, a_j^0 \sim \mathcal{N}(0,1)$.
- **Figure 3 (right)**: $p = 8, d = 2048, \gamma_0 = 0.1, \rho_0 = 0.1, N = 20, a_j^0 \sim \mathcal{N}(0,1)$.
- **Figure 2 (left)**: $p = 8, d = 2048, \gamma_0 = 0.1, \rho_0 = 0.1, N = 40, a_j^0 \sim \mathcal{N}(0,1)$.
- **Figure 2 (right)**: $p = 4, d = 1024, \gamma_0 = 0.01, \rho_0 = 0.1, N = 10, a_j^0 \sim \mathcal{N}(0,1)$.
- **Figure 4**: $p = 4, d = 1024, \gamma_0 = 0.01, \rho_0 = 0.1, N = 10, a_j^0 \sim \mathcal{N}(0,1)$.

- **Figure 5 (left)**: SAM $d = 4096, 8192, \gamma_0 = 0.01, \rho_0 = 0.1, N = 40$.
- **Figure 5 (right)**: 2-Lookahead $d = 4096, 8192, \gamma_0 = 0.01, \rho_0 = 0.1, N = 40$.
- **Figure 6 (all 2)**: $d = 8192, n_b = 2d, \gamma_0 = 0.1, N = 40$.
- **Figure 7 (upper)**: $p = 8, d = 2048, n_b = 2d, \gamma_0 = 0.1, \rho_0 = 0.1, N = 40, a_j^0 \sim \mathcal{N}(0, 1)$.
- **Figure 7 (lower)**: $p = 8, d = 2048, n_b = 2d, \gamma_0 = 0.1, \rho_0 = 0.1, N = 20, a_j^0 \sim \mathcal{N}(0, 1)$.

References

[AAM22] 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: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 4782–4887.

[AAM23] Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. "SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 2552–2623. URL: https://proceedings.mlr.press/v195/abbe23a.html.

[Abb+21] Emmanuel Abbe et al. "The staircase property: How hierarchical structure can guide deep learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 26989–27002.

[ABP22] Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. "Neural Networks as Kernel Learners: The Silent Alignment Effect". In: *International Conference on Learning Representations*. 2022.

[Arn+23] Luca Arnaboldi et al. "From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 1199–1227. URL: https://proceedings.mlr.press/v195/arnaboldi23a.html.

[Arn+24] Luca Arnaboldi et al. "Online Learning and Information Exponents: The Importance of Batch size & Time/Complexity Tradeoffs". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, July 2024, pp. 1730–1762.

[Ba+22] Jimmy Ba et al. "High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 37932–37946.

[Ba+23] Jimmy Ba et al. “Learning in the Presence of Low-dimensional Structure: A Spiked Random Matrix Perspective”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 17420–17449. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/38a1671ab0747b6ffe4d1c6ef117a3a9-Paper-Conference.pdf.

[Bar+19] Jean Barbier et al. “Optimal errors and phase transitions in high-dimensional generalized linear models”. In: *Proceedings of the National Academy of Sciences* 116.12 (2019), pp. 5451–5460. DOI: 10.1073/pnas.1802705116. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1802705116. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1802705116.

[BBP23] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. “On Learning Gaussian Multi-index Models with Gradient Flow”. In: *arXiv preprint arXiv:2310.19793* (2023).

[BD20] Björn Barz and Joachim Denzler. “Do we train on test data? Purging cifar of near-duplicates”. In: *Journal of Imaging* 6.6 (2020), p. 41.

[BGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. “Online stochastic gradient descent on non-convex losses from high-dimensional inference”. In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 4788–4838.

[Bie+22] Alberto Bietti et al. “Learning single-index models with shallow neural networks”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 9768–9783.

[CB18] Lénaic Chizat and Francis Bach. “On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

[CCM21] Michael Celentano, Chen Cheng, and Andrea Montanari. “The high-dimensional asymptotics of first order methods with random data”. In: *arXiv preprint arXiv:2112.07572* (2021).

[CM20] Sitan Chen and Raghu Meka. “Learning Polynomials in Few Relevant Dimensions”. In: *Proceedings of Thirty Third Conference on Learning Theory*. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 1161–1227.

[CS00] A. C. C. Coolen and D. Saad. “Dynamics of learning with restricted training sets”. In: *Phys. Rev. E* 62 (4 Oct. 2000), pp. 5444–5487. DOI: 10.1103/PhysRevE.62.5444. URL: https://link.aps.org/doi/10.1103/PhysRevE.62.5444.

[CSX00] A. C. C. Coolen, D. Saad, and Yuan-Sheng Xiong. “On-line learning from restricted training sets in multilayer neural networks”. In: *Europhysics Letters* 51.6 (Sept. 2000), p. 691. DOI: 10.1209/epl/i2000-00394-5. URL: https://dx.doi.org/10.1209/epl/i2000-00394-5.

[Dam+23] Alex Damian et al. “Smoothing the Landscape Boosts the Signal for SGD: Optimal Sample Complexity for Learning Single Index Models”. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh et al. 2023. URL: http://papers.nips.cc/paper%

5C_files/paper/2023/hash/02763667a5761ff92bb15d8751bcd223-Abstract-Conference.html.

[Dam+24] Alex Damian et al. "Computational-Statistical Gaps in Gaussian Single-Index Models". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 1262–1262.

[Dan+24a] Yatin Dandi et al. "How two-layer neural networks learn, one (giant) step at a time". In: *Journal of Machine Learning Research* 25.349 (2024), pp. 1–65.

[Dan+24b] Yatin Dandi et al. "The Benefits of Reusing Batches for Gradient Descent in Two-Layer Networks: Breaking the Curse of Information and Leap Exponents". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, July 2024, pp. 9991–10016. URL: https://proceedings.mlr.press/v235/dandi24a.html.

[DH18] Rishabh Dudeja and Daniel Hsu. "Learning Single-Index Models in Gaussian Space". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 1887–1930. URL: https://proceedings.mlr.press/v75/dudeja18a.html.

[DLS22] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. "Neural Networks can Learn Representations with Gradient Descent". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 5413–5452.

[For+21] Pierre Foret et al. "Sharpness-aware Minimization for Efficiently Improving Generalization". In: *International Conference on Learning Representations*. 2021.

[Gho+20] Behrooz Ghorbani et al. "When Do Neural Networks Outperform Kernel Methods?" In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 14820–14830.

[Gra49] Harold Grad. "Note on N-dimensional hermite polynomials". en. In: *Communications on Pure and Applied Mathematics* 2.4 (1949), pp. 325–330. ISSN: 1097-0312. DOI: 10.1002/cpa.3160020402.

[JGH18] Arthur Jacot, Franck Gabriel, and Clement Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

[JMS24] Nirmit Joshi, Theodor Misiakiewicz, and Nathan Srebro. "On the Complexity of Learning Sparse Functions with Statistical and Gradient Queries". In: *Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[Kor76] G. M. Korpelevič. "An extragradient method for finding saddle points and for other problems". In: *Èkonom. i Mat. Metody* 12.4 (1976), pp. 747–756. ISSN: 0424-7388.

[Kou+24] Yiwen Kou et al. "Matching the Statistical Query Lower Bound for k-Sparse Parity Problems with Sign Stochastic Gradient Descent". In: *Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[LAL19] Wangyu Luo, Wael Alghamdi, and Yue M Lu. "Optimal spectral initialization for signal recovery with applications to phase retrieval". In: *IEEE Transactions on Signal Processing* 67.9 (2019), pp. 2347–2356.

[Lee+24] Jason D. Lee et al. "Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit". In: *Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[Mai+20] Antoine Maillard et al. "Phase retrieval in high dimensions: Statistical and computational phase transitions". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 11071–11082.

[MM18] Marco Mondelli and Andrea Montanari. "Fundamental Limits of Weak Recovery with Applications to Phase Retrieval". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 1445–1450.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–E7671.

[Mon+24] Behrad Moniri et al. "A theory of non-linear feature learning with one gradient step in two-layer neural networks". In: *Proceedings of the 41st International Conference on Machine Learning*. Vienna, Austria, 2024.

[Mou+23] Alireza Mousavi-Hosseini et al. "Gradient-Based Feature Learning under Structured Data". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 71449–71485. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/e21955c93dede886af1d0d362c756Paper-Conference.pdf.

[ODo14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge: Cambridge University Press, 2014. ISBN: 9781107038325. DOI: 10.1017/CBO9781139814782.

[Pet+22] Leonardo Petrini et al. "Learning sparse features can lead to overfitting in neural networks". In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Ed. by Sanmi Koyejo et al. 2022. URL: http://papers.nips.cc/paper%5C_files/paper/2022/hash/3d3a9e085540c65dd3e5731361f9320e-Abstract-Conference.html.

[RV22] Grant Rotskoff and Eric Vanden-Eijnden. "Trainability and Accuracy of Artificial Neural Networks: An Interacting Particle System Approach". In: *Communications on Pure and Applied Mathematics* 75.9 (2022), pp. 1889–1935. DOI: https://doi.org/10.1002/cpa.22074. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22074. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22074.

[SS20] Justin Sirignano and Konstantinos Spiliopoulos. "Mean field analysis of neural networks: A central limit theorem". In: *Stochastic Processes and their Applications* 130.3 (2020), pp. 1820–1852.

[SS95] David Saad and Sara A. Solla. “On-line learning in soft committee machines”. In: *Phys. Rev. E* 52 (4 Oct. 1995), pp. 4225–4243. DOI: 10.1103/PhysRevE.52.4225.

[Tro+25] Emanuele Troiani et al. “Fundamental computational limits of weak learnability in high-dimensional multi-index models”. In: *The 28th International Conference on Artificial Intelligence and Statistics*. 2025.

[ZB24] Aaron Zweig and Joan Bruna. “Symmetric Single Index Learning”. In: *The Twelfth International Conference on Learning Representations*. 2024.

[Zha+19] Michael Zhang et al. “Lookahead Optimizer: k steps forward, 1 step back”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/90fd4f88f588ae64038134f1eeaa023f-Paper.pdf.