An Initial Alignment between Neural Network and Target is Needed for Gradient Descent to Learn

Emmanuel Abbe [1] **Elisabetta Cornacchia** [1] **Jan Hązła** [1 2] **Christopher Marquis** [1]

Abstract

This paper introduces the notion of "Initial Alignment" (INAL) between a neural network at initialization and a target function. It is proved that if a network and a Boolean target function do not have a noticeable INAL, then noisy gradient descent on a fully connected network with normalized i.i.d. initialization will not learn in polynomial time. Thus a certain amount of knowledge about the target (measured by the INAL) is needed in the architecture design. This also provides an answer to an open problem posed in (Abbe & Sandon, 2020a). The results are based on deriving lower-bounds for descent algorithms on symmetric neural networks without explicit knowledge of the target function beyond its INAL.

1. Introduction

Does one need an educated guess on the type of architecture needed in order for gradient descent to learn certain target functions? Convolutional neural networks (CNNs) have an architecture that is natural for learning functions having to do with image features: at initialization, a CNN is already well posed to pick up correlations with the image content due to its convolutional and pooling layers, and gradient descent (GD) allows to locate and amplify such correlations. However, a CNN may not be the right architecture for non-image based target functions, or even certain image-based functions that are non-classical (Liu et al., 2018). More generally, we raise the following question:

> Is a certain amount of 'initial alignment' needed between a neural network at initialization and a target function in order for GD to learn on a reasonable horizon? Or could a neural net that is not properly designed but large enough find its own path to correlate with the target?

In order to formalize the above question, one needs to define the notion of 'alignment' as well as to quantify the 'certain amount' and 'reasonable horizon' notions. This paper focuses on the 'polynomial-scaling' regime and on fully connected architectures, but we conjecture that a more general quantitative picture can be derived. Before defining the question formally, we stress a few connections to related problems.

A different type of 'gradual' question has recently been investigated for neural networks, namely, the 'depth gradual correlation' hypothesis. This postulates that if a neural network of low depth (e.g., depth 2) cannot learn to a non-trivial accuracy after GD has converged, then an augmentation of the depth to a larger constant will not help in learning (Malach & Shalev-Shwartz, 2019; Allen-Zhu & Li, 2020). In contrast, the question studied here is more of a 'time gradual correlation' hypothesis, saying that if at time zero GD cannot correlate non-trivially with a target function (i.e., if the neural net at time zero does not have an initial alignment), then a polynomial number of GD steps will not help.

From a lower-bound point of view, the question we ask is also slightly different than the traditional lower-bound questions posed in the learning literature that have to do with the difficulties of learning a class of functions irrespective of a specific architecture. For instance, it is known from (Blum et al., 1994; Kearns, 1998) that the larger the statistical dimension of a function class is, the more challenging it is for a statistical query (SQ) algorithm to learn, and similarly for GD-like algorithms (Abbe et al., 2021); these bounds hold irrespective of the type of neural network architectures used.

A more architecture-dependent lower-bound is derived in (Abbe & Sandon, 2020b), where the junk-flow is essentially used as replacement of the number of queries, and which depends on the type of architecture and initialization albeit being implicit. In (Shalev-Shwartz & Malach, 2021), a separation between fully connected and CNN architec-

[*]Equal contribution [1]Institute of Mathematics, EPFL, Lausanne, Switzerland [2]African Institute for Mathematical Sciences (AIMS), Kigali, Rwanda. Correspondence to: Elisabetta Cornacchia <elisabetta.cornacchia@epfl.ch>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

tures is obtained, showing that certain target functions have a locality property and are better learned by the latter architecture. In a different setting, (Tan et al., 2021) gives a generalization lower bound for decision trees on additive generative models, proving that decision trees are statistically inefficient at estimating additive regression functions. However, none of the bounds in these works give an explicit figure of merit to measure the suitability of a neural network architecture for a target.

One can interpret such bounds, especially the one in (Abbe & Sandon, 2020b), as follows. If the function class is such that for two functions F, F' sampled randomly from the class, the typical correlation is not noticeable, i.e., if the cross-predictability (CP) is given by

$$\mathrm{CP}(F, F') := \mathbb{E}_{F,F'} \langle F, F' \rangle^2 = n^{-\omega_n(1)}, \qquad (1)$$

(where we denoted by $\langle . \rangle$ the L^2-scalar product, namely, for some input distribution $P_{\mathcal{X}}$, $\langle f, g \rangle = \mathbb{E}_{x \sim P_{\mathcal{X}}}[f(x)g(x)]$ and by $\omega_n(1)$ any sequence that is diverging to ∞ as $n \to \infty$), then GD with polynomial precision and on a polynomial horizon will not be able to identify the target function with an inverse polynomial accuracy (weak learning), because at no time the algorithm will approach a good approximation of the target function; i.e. the gradients stay essentially agnostic to the target.

Instead, here we focus on a specific function — rather than a function class — and on a specific architecture and initialization. One can engineer a function class from a specific function if the initial architecture has some distribution symmetry. In such case, if the original function is learnable, then its orbit under the group of symmetry must also be learnable, and thus lower bounds based on the cross-predictability or statistical dimension of the orbit can be used. Such lower bounds are no longer applying to any architecture but exploit the symmetry of the architecture, however they still require knowledge of the target function in order to define the orbit.

In this paper, we would like to depart from the setting where we know the target function and thus can analyze the orbit directly. Instead, we would like to have a 'proxy' that depends on the underlying target function and the initialized neural net NN_{Θ^0} at hand, where the set of weights at time zero Θ^0 are drawn according to some distribution. In (Abbe & Sandon, 2020a), the following proposal is made (the precise statement will appear below): can we replace the correlation among a function class by the correlation between a target function and an initialized net in order to have a necessary requirement for learning, i.e., if

$$\mathbb{E}_{\Theta^0} \langle f, \mathrm{NN}_{\Theta^0} \rangle^2 = n^{-\omega_n(1)}, \qquad (2)$$

or in other words, if at initialization the neural net correlates negligibly with the target function, is it still possible for GD to learn[1] the function f if the number of epochs of GD is polynomial? We next formalize the question further and provide an answer to it.

Note the difference between (1) and (2): in (1) it is the class of functions that is too poorly correlated for *any* SQ algorithm to efficiently learn; in (2) it is the specific network initialization that is too poorly correlated with the specific target in order for GD to efficiently learn.

While previous works and our proof relies on creating the orbit of a target function using the network symmetries and then arguing from the complexity of the orbit (using cross-predictability (Abbe & Sandon, 2020a)), we believe that the INAL approach can be fruitful in additional contexts. In fact, the orbit approaches have two drawbacks: (1) they cannot give lower-bounds on functions like the full parity[2] that have no complex orbit (in fact the orbit of the full parity is itself under permutation symmetries), (2) to estimate the complexity measure of the orbit class (e.g., the cross-predictability) from a sample set without full access to the target function, one needs labels of data points under the group action that defines the orbit (e.g., permutations), and these may not always be available from an arbitrary sample set. In contrast, (i) the INAL can still be small for the full parity function on certain symmetric neural networks, suggesting that in such cases the full parity is not learnable (we do not prove this here due to our specific proof technique but conjecture that this result still holds), (ii) the INAL can always be estimated from a random i.i.d. sample set, using basic Monte Carlo simulations (as used in our experiments, see Section 5).

While the notion of INAL makes sense for any input distribution, our theoretical results are proved in a more limited setting of Boolean functions with uniform inputs. This follows the approach that has been taken in (Abbe & Sandon, 2020b) and we made that choice for similar reasons. Furthermore, any computer-encoded function is eventually Boolean and major part of the PAC learning theory has indeed focused on Boolean functions (we refer to (Shalev-Shwartz & Ben-David, 2014) for more on this subject). We nonetheless expect that the footprints of the proofs derived in this paper will apply to inputs that are iid Gaussians or spherical, using different basis than the Fourier-Walsh one.

Our general strategy in obtaining such a result is as follows: we first show that for the type of architecture considered, a low initial alignment (INAL) implies that the implicit target function is essentially high-degree in its Fourier basis; this part is specific to the architecture and the low INAL property. We next use the symmetry of the initialization to conclude

[1] Even with just an inverse polynomial accuracy, a.k.a., weak learning.

[2] we call full parity the function $f : \{\pm 1\}^n \to \{\pm\}$ s.t. $f(x) = \prod_{i=1}^n x_i$.

that learning under such high-degree Fourier requirement implies learning a low CP class, and thus conclude by leveraging the results from (Abbe & Sandon, 2020b). Finally, we do some experiments with the types of architecture used in our formal results, but also with convolutional neural nets to test the robustness of the original conjecture. We observe that generally the INAL gives a decent proxy for the difficulty to learn (lower INAL gives lower learning accuracy). While this goes beyond the scope of our paper — which is to obtain a first rigorous validation of the INAL conjecture for standard fully connected neural nets — we believe that the numerical simulations give some motivations to pursue the study of the INAL in a more general setting.

2. Definitions and Theoretical Contributions

For the purposes of our definition, a neural network NN consists of a set of neurons V_{NN}, a random variable $\Theta^0 \in \mathbb{R}^k$ which corresponds to the initialization and a collection of functions $\mathrm{NN}^{(v)}_{\Theta^0} : \mathbb{R}^n \to \mathbb{R}$ indexed with $v \in V_{\mathrm{NN}}$, representing the outputs of neurons in the network. The Initial Alignment (INAL) is defined as the average squared correlation between the target function and any of the neurons at initialization:

Definition 2.1 (Initial Alignment (INAL))**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function and $P_{\mathcal{X}}$ a distribution on $\mathbb{R}^n$. Let NN be a neural network with neuron set V_{NN} and random initialization Θ^0. Then, the INAL is defined as

$$\mathrm{INAL}(f, \mathrm{NN}) := \max_{v \in V_{\mathrm{NN}}} \mathbb{E}_{\Theta^0} \langle f, \mathrm{NN}^{(v)}_{\Theta^0} \rangle^2, \tag{3}$$

where we denoted by $\langle . \rangle$ the L^2-scalar product, namely $\langle f, g \rangle = \mathbb{E}_{x \sim P_{\mathcal{X}}}[f(x)g(x)]$.

While the above definition makes sense for any neural network architecture, in this paper we focus on fully connected networks. Thus, in the following NN will denote a fully connected neural network. Our main thesis is that in many settings a small INAL is bad news: If at initialization there is no noticeable correlation between any of the neurons and the target function, the GD-trained neural network will not be able to recover such correlation during training in polynomial time.

Of particular interest to us is the notion of INAL for a single neuron with activation σ and normalized Gaussian initialization.

Definition 2.2. Let $f : \mathbb{R}^n \to \mathbb{R}$, $\sigma : \mathbb{R} \to \mathbb{R}$ and let $P_{\mathcal{X}}$ be a distribution on $\mathbb{R}^n$. Then, we abuse the notation and write

$$\mathrm{INAL}(f, \sigma) := \mathbb{E}_{w^n, b^n} \left[\left(\mathbb{E}_{x \sim P_{\mathcal{X}}} f(x) \sigma((w^n)^T x + b^n) \right)^2 \right], \tag{4}$$

where w^n is a vector of iid $\mathcal{N}(0, 1/n)$ Gaussians and b^n is another independent $\mathcal{N}(0, 1/n)$ Gaussian. In the following, for readability, we will write $w = w^n$ and $b = b^n$, omitting the dependence on n.

In the following, we say that a function $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ is *noticeable* if there exists $c \in \mathbb{N}$ such that $f(n) = \Omega(n^{-c})$. On the other hand, we say that f is *negligible* if $\lim_{n \to \infty} n^c f(n) = 0$ for every $c \in \mathbb{N}$ (which we also write $f(n) = n^{-\omega_n(1)}$).

Definition 2.3 (Weak learning)**.** Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of functions such that $f_n : \mathbb{R}^n \to \mathbb{R}$ and (P_n) a sequence of probability distributions on $\mathbb{R}^n$. Let (A_n) be a family of randomized algorithms such that A_n outputs a function $\mathrm{NN}_n : \mathbb{R}^n \to \mathbb{R}$. Then, we say that A_n *weakly learns* f_n if the function

$$g(n) := \left| \mathbb{E}_{x \sim \mathcal{P}_n, A_n} [f_n(x) \cdot \mathrm{NN}_n(x)] \right| \tag{5}$$

is noticeable.

In this paper, we follow the example of (Abbe & Sandon, 2020b) and focus on *Boolean* functions with inputs and outputs in $\{\pm 1\}$. We consider sequences of Boolean functions $f_n : \{\pm 1\}^n \to \{\pm 1\}$, with the uniform input distribution $\mathcal{U}^n$, meaning that if $x \sim \mathcal{U}^n$, then for all $i \in [n]$, $x_i \overset{iid}{\sim} \mathrm{Rad}(1/2)$. We focus on fully connected neural networks with activation function σ, and trained by noisy GD — this means GD where the gradient's magnitude per the precision noise is polynomially bounded, as commonly considered in statistical query algorithms (Kearns, 1998; Blum et al., 1994) and GD learning (Abbe & Sandon, 2020b; Malach et al., 2021; Abbe et al., 2021); see Remark 3.5 for a remainder of the definition. We consider activation functions that satisfy the following conditions.

Definition 2.4 (Expressive activation)**.** We say that a function $\sigma : \mathbb{R} \to \mathbb{R}$ is *expressive* if it satisfies the following conditions:

a) σ is measurable and *polynomially bounded* i.e. there exists $C, c > 0$ such that $|\sigma(x)| \leq Cx^c + C$ for all $x \in \mathbb{R}$.

b) Let the Gaussian smoothing of σ be defined as $\Sigma(t) := \mathbb{E}_{Y \sim \mathcal{N}(0,1)}[\sigma(Y + t)]$. For each $m \in \mathbb{N}$ either $\Sigma^{(m)}(0) \neq 0$ or $\Sigma^{(m+1)}(0) \neq 0$ (where $\Sigma^{(m)}$ denotes the m-th derivative of Σ).

Remark 2.5. i) Note that we have the identities $d_m = \frac{\Sigma^{(m)}(0)}{m!}$, and $\sigma = \sum_{m=0}^{\infty} d_m H_m$, where H_m are the probabilist's Hermite polynomials. Therefore, an equivalent statement of the second condition in Definition 2.4 is that there are no two or more consecutive zeros in the Hermite expansion of σ.

ii) Many functions are expressive, including ReLU and sign (see Appendix B for the proofs of those two cases).

iii) On the other hand, it turns out that polynomials are *not* expressive, as they do not satisfy point b). This is necessary for our hardness results to hold, since for an activation function P which is a polynomial of degree k and M a monomial of degree $k+1$ it can be checked that $\mathrm{INAL}(M,P)=0$, but constant-degree monomials are learnable by GD.

Let us give one more definition before stating our main theorem.

Definition 2.6 (N-Extension)**.** For a function $f:\mathbb{R}^n\to\mathbb{R}$ and for $N>n$, we define its *N-extension* $\overline{f}:\mathbb{R}^N\to\mathbb{R}$ as

$$\overline{f}(x_1,x_2,...,x_n,x_{n+1},x_{n+2},...,x_N)=f(x_1,x_2,...,x_n). \tag{6}$$

We can now state our main result which connects INAL and weak learning.

Theorem 2.7 (Main theorem, informal)**.** *Let σ be an expressive activation function and (f_n) a sequence of Boolean functions with uniform distribution on $\{\pm1\}^n$. If $\mathrm{INAL}(f_n,\sigma)$ is negligible, then, for every $\epsilon>0$, the $n^{1+\epsilon}$-extension of f_n is* not *weakly learnable by $\mathrm{poly}(n)$-sized fully-connected neural networks with iid initialization and $\mathrm{poly}(n)$-number of steps of noisy gradient descent.*

Remark 2.8. Theorem 2.7 says that Boolean functions that have negligible correlation for *some* expressive activation and Gaussian iid initialization, cannot be learned by neural networks utilizing *any* activation on a fully-connected architecture and any iid initialization.

Remark 2.9. Consider a sequence of neural networks (NN_n) utilizing an expressive activation σ. We believe that the notion of $\mathrm{INAL}(f_n,\mathrm{NN}_n)$ is relevant to characterizing if a family of Boolean functions (f_n) is weakly learnable by noisy GD on those neural networks. On the one hand, if $\mathrm{INAL}(f_n,\mathrm{NN}_n)$ is noticeable, then at initialization there exists a neuron from which a weak correlation with f_n can be extracted. Therefore, in a sense weak learning is achieved at initialization.

On the other hand, assume additionally that the architecture is such that there exists a neuron computing $\sigma(w^Tx+b)$, where x is the input and (w,b) are initialized as iid $\mathcal{N}(0,1/n)$ Gaussians. (In other words, there exists a fully-connected neuron in the first hidden layer.) Then, by definition of INAL, if $\mathrm{INAL}(f_n,\mathrm{NN}_n)$ is negligible, then also $\mathrm{INAL}(f_n,\sigma)$ is negligible. Accordingly, by Theorem 2.7, an extension of (f_n) is not weakly learnable.

While we do not have a proof, we suspect that a similar property might hold also for some other architectures and initializations.

Note that we obtain hardness only for an extension of f_n, rather than for the original function. Interestingly, in some settings GD can learn the function, while the $2n$-extension of the same function is hard to learn[3]. However, we are not sure if such examples can be constructed for the continuous Gaussian initialization that we consider.

3. Formal Results

In this section, we write precise statements of our theorems. For this, we need a couple of more definitions.

Definition 3.1 (Cross-Predictability)**.** Let $P_{\mathcal{F}}$ be a distribution over functions from $\mathbb{R}^n$ to $\mathbb{R}$ and $P_{\mathcal{X}}$ a distribution over $\mathbb{R}^n$. Then,

$$\mathrm{CP}(P_{\mathcal{F}},P_{\mathcal{X}})=\mathbb{E}_{F,F'\sim P_{\mathcal{F}}}[\mathbb{E}_{X\sim P_{\mathcal{X}}}[F(X)F'(X)]^2]\ . \tag{7}$$

Definition 3.2 (Orbit)**.** For $f:\mathbb{R}^n\to\mathbb{R}$ and a permutation $\pi\in S_n$, we let $(f\circ\pi)(x)=f(x_{\pi(1)},\dots,x_{\pi(n)})$. Then, we define the *orbit* of f as

$$\mathrm{orb}(f):=\{f\circ\pi:\pi\in S_n\}\ . \tag{8}$$

Let us now give the full statement of our main theorem.

Theorem 3.3. *Let (f_n) be a sequence of Boolean functions with $f_n:\{\pm1\}^n\to\{\pm1\}$ and $x\sim\mathcal{U}^n$ and let σ be an expressive activation.*

If $\mathrm{INAL}(f_n,\sigma)$ is negligible, then, for every $\epsilon>0$, the cross predictability $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N)$ is negligible, where $N=n^{1+\epsilon}$ and $\mathrm{orb}(\overline{f_n})$ denotes (uniform distribution on) the orbit of the N-extension of f_n.

More precisely, if $\mathrm{INAL}(f_n,\sigma)=O(n^{-c})$, then $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N)=O(n^{-\frac{\epsilon}{1+\epsilon}(c-1)})$.

Applying (Abbe & Sandon, 2020b)[Theorem 3] to Theorem 3.3 implies the following corollary. We refer to Appendix E for additional clarifications on the notion of a fully connected neural net.

Corollary 3.4. *Let f_n and σ be as in Theorem 3.3 with negligible $\mathrm{INAL}(f_n,\sigma)$ and let $\epsilon>0$ with $N=n^{1+\epsilon}$ and $\overline{f_n}$ denote the N-extension of f_n.*

Let $\mathrm{NN}=(\mathrm{NN}_n)$ be any sequence of fully connected neural nets of polynomial size. Then, for any iid initializaton, and any polynomial bounds on the learning rate, learning time $T=(T_n)$, noise level and overflow range, the noisy GD algorithm after T steps of training outputs a neural net $\mathrm{NN}^{(T)}$ such that the correlation

$$g(n):=\left|\mathbb{E}_{\mathrm{NN}^{(T)}}\langle\mathrm{NN}^{(T)},\overline{f_n}\rangle\right| \tag{9}$$

is negligible.

[3]For example, for the Boolean parity function $M_n(x)=\prod_{i=1}^n x_i$ with both the input distribution and the weight initialization iid uniform in $\{\pm1\}$ and cosine activation (Boix-Adsera, 2021).

More precisely, if $\mathrm{INAL}(f_n, \sigma) = O(n^{-c})$*, then for a noisy GD run for* T *steps on a fully connected neural network with* E *edges, with learning rate* γ*, overflow range* A *and noise level* τ *it holds that*

$$g(n) = O\left(\frac{\gamma T\sqrt{E}A}{\tau} \cdot n^{-\frac{\epsilon}{4(1+\epsilon)}(c-1)}\right) . \qquad (10)$$

Remark 3.5. In the result above, the neural net can have any feed-forward architecture with layers of fully-connected neurons and any activation such that the gradients are almost surely well-defined. The initialization can be iid from any distribution (which can depend on n). We remark that the result of Corollary 3.4 can be strengthen to apply to any initialization such that the distribution of the weights in the first layer is invariant under permutations of input neurons. We refer to Appendix E for more details.

The algorithm considered is noisy gradient descent[4] using any differentiable loss function, meaning that at every step an iid $\mathcal{N}(0, \tau^2)$ noise vector is added to all components of the gradient, where τ is called the *noise level*. Furthermore, every component of the gradient during the execution of the algorithm whose evaluation exceeds the *overflow range* A in absolute value is clipped to A or $-A$, respectively. This covers in particular the bounded 'precision model' of (Abbe et al., 2021).

For the purposes of function $g(n)$, it is assumed that the neural network outputs a guess in $\{\pm 1\}$ using any form of thresholding (eg., the sign function) on the value of the output neuron. See (Abbe & Sandon, 2020b)[Section 2.3.1].

4. Proof of Main Theorem

In this section we sketch the proof of Theorem 3.3. We first state basic definitions from Boolean function analysis, then we give a short outline of the proof, and then we state main propositions used in the proof. Finally, we show how the propositions are combined to prove Theorem 3.3 and Corollary 3.4. Further proofs and details are in the appendices.

We introduce some notions of Boolean analysis, mainly taken from Chapters 1,2 of (O'Donnell, 2014). For every $f : \{\pm 1\}^n \to \mathbb{R}$ we denote its Fourier expansion as

$$f(x) = \sum_{S\subseteq[n]} \hat{f}(S) M_S(x), \qquad (11)$$

where $M_S(x) = \prod_{i\in S} x_i$ are the standard Fourier basis elements and $\hat{f}(S)$ are the Fourier coefficients of f, defined as $\hat{f}(S) = \langle f, M_S\rangle$. We denote by

$$W^k(f) = \sum_{S:|S|=k} \hat{f}(S)^2 \qquad (12)$$

$$W^{\leq k}(f) = \sum_{S:|S|\leq k} \hat{f}(S)^2 , \qquad (13)$$

the total weight of the Fourier coefficients of f at degree k (respectively up to degree k).

Definition 4.1 (High-Degree)**.** We say that a family of functions $f_n : \{\pm 1\}^n \to \mathbb{R}$ is "high-degree" if for any fixed k, $W^{\leq k}(f_n)$ is negligible.

Proof Outline of Theorem 3.3.

1. We initially restrict our attention to the basis Fourier elements, i.e. the monomials $M_S(x) := \prod_{i\in S} x_i$ for $S \in [n]$. We consider the single-neuron alignments $\mathrm{INAL}(M_S, \sigma)$ for expressive activations. We prove that these INALs are noticeable for constant degree monomials (Proposition 4.3).

2. For a general $f : \{\pm 1\}^n \to \mathbb{R}$ we show that the initial alignment between f and a single-neuron architecture can be computed from its Fourier expansion (Proposition 4.4). As a consequence, for any expressive σ, if $\mathrm{INAL}(f, \sigma)$ is negligible, then f is high-degree (Corollary 4.5).

3. We construct the extension of f and take its orbit $\mathrm{orb}(\overline{f})$. Since the extension has a sparse structure of its Fourier coefficients, that guarantees that the cross-predictability of $\mathrm{orb}(\overline{f})$ is negligible (Proposition 4.6).

4. In order to prove Corollary 3.4, we invoke the lower bound of (Abbe & Sandon, 2020b) (Theorem 4.7) applied to the class $\mathrm{orb}(\overline{f})$.

A crucial property of the expressive activations is that they correlate with constant-degree monomials. To emphasize this, we introduce another definition.

Definition 4.2. An activation σ is *correlating* if for every k, the sequence $\mathrm{INAL}(M_k, \sigma)$ is noticeable, where we think of $M_k(x) = \prod_{i=1}^k x_i$ as a sequence of Boolean functions for every input dimension $n \geq k$.

Furthermore, if there exists c such that for every k it holds $\mathrm{INAL}(M_k, \sigma) = \Omega(n^{-(k+c)})$, then we say that σ is c-*strongly correlating*.

Proposition 4.3. *If* σ *is expressive (according to Definition 2.4), then it is 1-strongly correlating.*

The proof of Proposition 4.3 is our main technical contribution. Since the magnitude of the correlations is quite

[4] In fact, it can be SGD with batch size m for large enough m.

small (in general, of the order n^{-k} for monomials of degree k), careful calculations are required to establish our lower bounds.

In fact, we conjecture that any polynomially bounded function that is not a polynomial (almost everywhere) is correlating.

Then, we show that $\mathrm{INAL}(f,\sigma)$ decomposes into monomial INALs according to its Fourier coefficients:

Proposition 4.4. *For any $f:\{\pm1\}^n\to\mathbb{R}$ and any activation σ,*

$$\mathrm{INAL}(f,\sigma) := \sum_{T\in[n]} \hat{f}(T)^2\,\mathrm{INAL}(M_T,\sigma)\,. \tag{14}$$

As a corollary, functions with negligible INAL on correlating activations are high-degree:

Corollary 4.5. *Let σ be an activation with $\mathrm{INAL}(M_{k'},\sigma)=\Omega(n^{-k_0})$ for $k'=0,1,\dots,k$. Then, $W^{\le k}(f_n)\le \mathrm{INAL}(f_n,\sigma)O(n^{k_0})$.*

In particular, if σ is correlating and $\mathrm{INAL}(f_n,\sigma)$ is negligible, then (f_n) is high degree.

Finally, the cross-predictability of $\mathrm{orb}(\overline{f_n})$ is negligible for high degree functions.

Proposition 4.6. *Let $\epsilon>0$ and (f_n) a family of Boolean functions. Let $(\overline{f_n})$ denote the family of N-extensions of f_n for $N=n^{1+\epsilon}$, and consider the uniform distribution on its orbit.*

If (f_n) is high degree, then $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N)$ is negligible. Furthermore, if for some universal c and every fixed k it holds $W^{\le k}(f_n)=O(n^{k-c})$, then $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N)=O(n^{-\frac{\epsilon}{1+\epsilon}\cdot c})$.

Theorem 4.7 ((Abbe & Sandon, 2020b), informal)**.** *If the cross-predictability of a class of functions is negligible, then noisy GD cannot learn it in poly-time.*

We provide here an outline of the proof of Proposition 4.3, and refer to Appendix A for a detailed proof. We further prove Proposition 4.6 and Theorem 3.3. The proofs of the remaining results are in appendices.

Proof of Proposition 4.3 (outline). The main goal of the proof is to estimate the dominant term (as n approaches infinity) of $\mathrm{INAL}(M_k,\sigma)$, and show that it is indeed noticeable, for any fixed k. We initially use Jensen inequality to lower bound the INAL with the following

$$\mathrm{INAL}(M_k,\sigma)\ge \mathbb{E}\left[\mathbb{E}_{|\theta|,x}\left[M_k(x)\sigma(w^Tx+b)\mid \mathrm{sgn}(\theta)\right]^2\right], \tag{15}$$

where for brevity we denoted $\theta=(w,b)$, $|\theta|$ and $\mathrm{sgn}(\theta)$ are $(n+1)$-dimensional vectors such that $|\theta|_i=|\theta_i|$ and $\mathrm{sgn}(\theta)_i=\mathrm{sgn}(\theta_i)$, for all $i\le n+1$. By denoting $|w|_{>k},x_{>k}$ the coordinates of $|w|$ and x respectively that do not appear in M_k, and by $G:=\sum_{i=1}^k w_ix_i+b$ we observe that

$$\mathbb{E}_{|w|_{>k},x_{>k}}[\sigma(w^Tx+b)]=\mathbb{E}_{Y\sim\mathcal{N}(0,1-\frac{k}{n})}\left[\sigma(G+Y)\right], \tag{16}$$

since $\sum_{i=k+1}^n w_ix_i$ is indeed distributed as $\mathcal{N}(0,1-\frac{k}{n})$. We call the RHS the "$n$-Gaussian smoothing" of σ and we denote it by $\Sigma_n(z):=\mathbb{E}_{Y\sim\mathcal{N}(0,1-\frac{k}{n})}[\sigma(z+Y)]$. We will compare it to the "ideal" Gaussian smoothing denoted by $\Sigma(z):=\mathbb{E}_{Y\sim\mathcal{N}(0,1)}[\sigma(z+Y)]$.

For polynomially bounded σ, we can prove that Σ_n has some nice properties (see Lemma A.1), specifically it is C^∞ and polynomially bounded and it uniformly converges to Σ as $n\to\infty$. These properties crucially allow to write Σ_n in terms of its Taylor expansion around 0, and bound the coefficients of the series for large n. In fact, we show that there exists a constant $P>k$, such that if we split the Taylor series of Σ_n at P as

$$\Sigma_n(G)=\sum_{\nu=0}^{P} a_{\nu,n}G^\nu+R_{P,n}(G), \tag{17}$$

(where $a_{\nu,n}$ are the Taylor coefficients and $R_{P,n}$ is the remainder in Lagrange form), and take the expectation over $|\theta|_{\le k}$ as:

$$\begin{aligned}
&\mathbb{E}_{|\theta|_{\le k},x_{\le k}}\left[M_k(x)\Sigma_n(G)\right] && (18)\\
&\quad=\sum_{\nu=0}^{P} a_{\nu,n}\mathbb{E}_{|\theta|_{\le k},x_{\le k}}\left[M_k(x)G^\nu\right] && (19)\\
&\qquad+\mathbb{E}_{|\theta|_{\le k},x_{\le k}}\left[R_{P,n}(G)\right] && (20)\\
&\quad=:A+B, && (21)
\end{aligned}$$

then A is $\Omega(n^{-P/2})$ (Proposition A.3), and B is $O(n^{-P/2-1/2})$ (Proposition A.4), uniformly for all values of $\mathrm{sgn}(\theta)$. For A we use the observation that $\mathbb{E}_{|\theta|_{\le k},x_{\le k}}[M_k(x)G^\nu]=0$ for all $\nu<k$ (Lemma A.5), and the fact that $|a_{P,n}|>0$ for n large enough (due to hypothesis b in Definition 2.4 and the continuity of Σ_n in the limit of $n\to\infty$, given by Lemma A.1). For B, we combine the concentration of Gaussian moments and the polynomial boundness of all derivatives of Σ_n.

Taking the square of (21) and going back to (15), one can immediately conclude that $\mathrm{INAL}(M_k,\sigma)$ is indeed noticeable.

4.1. Proof of Proposition 4.6

Let $f=f_n$ and let $\hat{f}$ be the Fourier coefficients of the original function f, and let $\hat{h}$ be the coefficients of the augmented function $\bar{f}$. Recall that $\bar{f}:\{\pm1\}^N\to\{\pm1\}$ is such

that $\bar{f}(x_1,...,x_n,x_{n+1},...,x_N) = f(x_1,...,x_n)$. Thus, the Fourier coefficients of $\bar{f}$ are

$$\hat{h}(T) = \begin{cases} \hat{f}(T) & \text{if } T \subseteq [n], \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Let us proceed to bounding the cross-predictability. Below we denote by π a random permutation of N elements:

$$\begin{aligned}
&\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N) && (23)\\
&= \mathbb{E}_\pi\left[\mathbb{E}_x\left[\bar{f}(x)\bar{f}(\pi(x))\right]^2\right] && (24)\\
&= \mathbb{E}_\pi\left[\left(\sum_{T\subseteq[N]} \hat{h}(T)\hat{h}(\pi(T))\right)^2\right] && (25)\\
&= \mathbb{E}_\pi\left[\left(\sum_{T\subseteq[n]} \hat{f}(T)\hat{h}(\pi(T))\cdot \mathbb{1}\left(\pi(T)\subseteq[n]\right)\right)^2\right] && (26)\\
&\overset{C.S}{\leq} \mathbb{E}_\pi\left[\left(\sum_{S\subseteq[n]} \hat{h}(\pi(S))^2\right)\right. && (27)\\
&\qquad \left.\cdot\left(\sum_{T\subseteq[n]} \hat{f}(T)^2\mathbb{1}\left(\pi(T)\subseteq[n]\right)\right)\right] && (28)\\
&\leq \sum_{T\subseteq[n]} \hat{f}(T)^2\cdot\mathbb{P}_\pi\left(\pi(T)\subseteq[n]\right). && (29)
\end{aligned}$$

Now, for any k we have

$$\begin{aligned}
&\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N) && (30)\\
&\leq \sum_{T:|T|<k} \hat{f}(T)^2\cdot\mathbb{P}_\pi\left(\pi(T)\subseteq[n]\right) && (31)\\
&\quad + \sum_{T:|T|\geq k} \hat{f}(T)^2\cdot\mathbb{P}_\pi\left(\pi(T)\subseteq[n]\right) && (32)\\
&\leq W^{<k}(f) + \mathbb{P}_\pi\left(\pi(T)\subseteq[n] \mid |T|=k\right), && (33)
\end{aligned}$$

where the second term in (33) is further bounded by (recall that $N = n^{1+\epsilon}$):

$$\begin{aligned}
\mathbb{P}_\pi\left(\pi(T)\subseteq[n] \mid |T|=k\right) &= \frac{\binom{n}{k}}{\binom{N}{k}} && (34)\\
&\leq \frac{\left(\frac{ne}{k}\right)^k}{\left(\frac{N}{k}\right)^k} && (35)\\
&= e^k\frac{n^k}{N^k} = e^k n^{-\epsilon\cdot k}. && (36)
\end{aligned}$$

Accordingly, for any $k\in\mathbb{N}_{>0}$ it holds that

$$\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N) \leq W^{<k}(f) + e^k n^{-\epsilon k}. \tag{37}$$

Now, if (f_n) is a high degree sequence of Boolean functions, then $W^{<k}(f)$ is negligible for every k, and therefore the cross-predictability in (34) is $O(n^{-k})$ for every k, that is the cross-predictability is negligible as we claimed.

On the other hand, if for some c and every k it holds that $W^{\leq k}(f_n) = O(n^{k-c})$, then we can choose $k_0 := \frac{c}{1+\epsilon}$ and apply (37) to get $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N) = O(n^{-\frac{\epsilon}{1+\epsilon}\cdot c})$.

4.2. Proof of Theorem 3.3

Let σ be an expressive activation and let (f_n) be a sequence of Boolean functions with negligible $\mathrm{INAL}(f_n,\sigma)$. By Proposition 4.3, σ is correlating, and by Corollary 4.5 (f_n) is high-degree. Therefore, by Proposition 4.6, the cross-predictability $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N)$ is negligible.

For the more precise statement, let (f_n) be a sequence of Boolean functions with $\mathrm{INAL}(f_n,\sigma) = O(n^{-c})$. By Proposition 4.3, σ is 1-strongly correlating. That means that for every k we have $\mathrm{INAL}(M_k,\sigma) = \Omega(n^{-(k+1)})$. By Corollary 4.5, for every k it holds $W^{\leq k}(f_n) = O(n^{k+1-c})$. Finally, applying Proposition 4.6, we have that $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N) = O(n^{-\frac{\epsilon}{1+\epsilon}(c-1)})$.

5. Experiments

In this section we present a few experiments to show how the INAL can be estimated in practice. Our theoretical results connect the performance of GD to the Fourier spectrum of the target function. However, in applications we are usually given a dataset with data points and labels, rather than an explicit target function, and it may not be trivial to infer the Fourier properties of the function associated to the data. Conveniently, the INAL can be estimated with sufficient datapoints and labels, and do not need an explicit target.

Experiments on Boolean functions. In our first experiment, we consider three Boolean functions, namely the majority-vote over the whole input space ($\mathrm{Maj}_n(x) := \mathrm{sgn}(\sum_{i=1}^n x_i)$), a 9-staircase ($\mathrm{S}_9(\mathrm{x}) := \mathrm{sgn}(\mathrm{x}_1 + \mathrm{x}_1\mathrm{x}_2 + \mathrm{x}_1\mathrm{x}_2\mathrm{x}_3 + ... + \mathrm{x}_1\mathrm{x}_2\mathrm{x}_3...\mathrm{x}_9)$ and a 3-parity ($M_3(x) = \prod_{i=1}^3 x_i$), on an input space of dimension 100. We take a 2-layer fully connected neural network with ReLU activations and normalised Gaussian iid initialization (according to the setting of our theoretical results), and we train it with SGD with batch-size 1000 for 100 epochs, to learn each of the three functions. On the other hand, we estimate the INAL between each of the three targets and the neural network, through Monte-Carlo. Our observations confirm our theoretical claim, i.e. that low INAL is bad news. In fact, for the 3-parity and the 9-staircase, that have very low INAL ($\sim$1/20 of the majority-vote case), GD does not achieve good generalization accuracy after training (Figure 1).

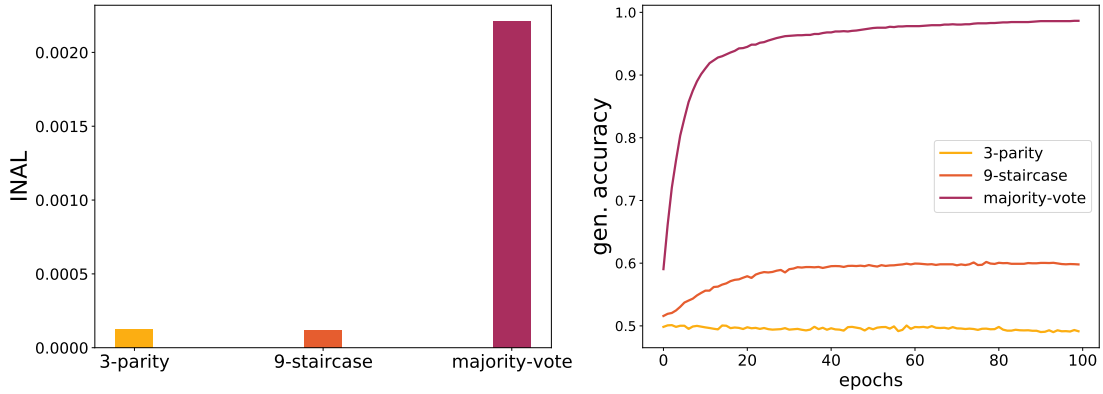

Figure 1: Comparison of INAL and generalization accuracy for three Boolean functions. On the left, we estimate the INAL between each target function and a 2-layers ReLU fully connected neural network with normalized gaussian initialization. On the right, we train the network to learn each target function with SGD with batch 1000 for 100 epochs. We observe that low INAL is bad news.

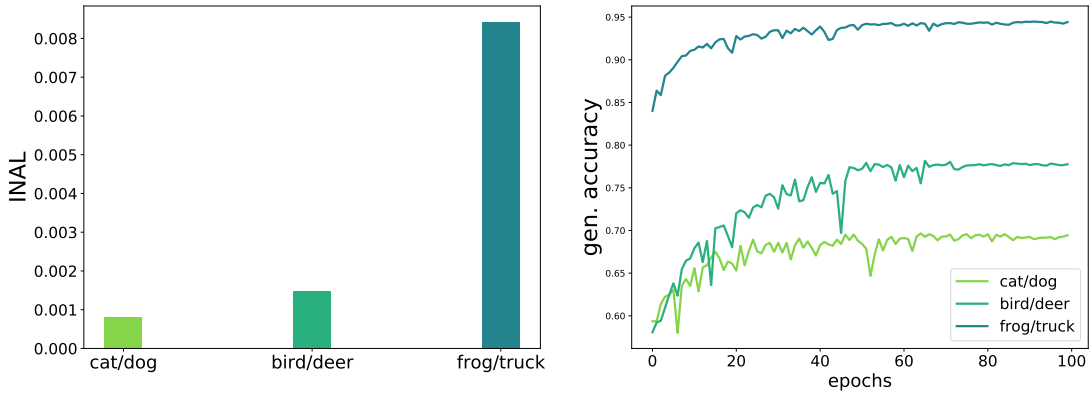

Figure 2: Comparison of INAL and generalization accuracy for binary classification in the CIFAR dataset. On the left, we estimate the INAL between the neural network and the target function associated to each task. On the right we train a CNN with 1 VGG block with SGD with batch size 64 for 100 epochs. We observe that a significant difference in the INAL corresponds to a significant difference in the generalization accuracy achieved.

Experiments on real data. Given a dataset $D = (x_m, y_m)_{m\in[M]}$, where $x_m \in \mathbb{R}^n$, and $y_m \in \mathbb{R}$, and given a randomly initialized neural network NN_{Θ^0} with Θ^0 drawn from some distribution, we can estimate the initial alignment between the network and the target function associated to the dataset as

$$\max_{v\in V_{NN}} \mathbb{E}_{\Theta^0}\left[\left(\frac{1}{M}\sum_{m=1}^{M} y_m \cdot \mathrm{NN}^{(v)}_{\Theta^0}(x_m)\right)^2\right], \qquad (38)$$

where the outer expectation can be performed through Monte-Carlo approximation.

We ran experiments on the CIFAR dataset. We split the dataset into 3 different pairs of classes, corresponding to 3 different binary classification tasks (specifically cat/dog, bird/deer, frog/truck). We take a CNN with 1 VGG block and ReLU activation, and for each task, we train the network with SGD with batch-size 64, and we estimate the INAL according to (38). We notice that also in this setting (not covered by our theoretical results), the INAL and the generalization accuracy present some correlation, and a significant difference in the INAL corresponds to a significant difference in the accuracy achieved after training. This may give some motivation to study the INAL beyond the fully connected setting.

6. Conclusion and Future Work

There are several directions that can follow from this work. The most relevant would be to extend the result beyond

fully connected architectures. As mentioned before, we suspect that our result can be generalized to all architectures that contain a fully connected layer anywhere in the network. Another direction would be to extend the present work to other continuous distributions of intitial weights (beyond gaussian). As a matter of fact, in the setting of iid gaussian inputs (instead of Boolean inputs), our proof technique extends to all weight initialization distributions with zero mean and variance $O(n^{-1})$. However, in the case of Boolean inputs that we consider in this paper, this may not be a trivial extension. Another extension on which we do not touch here are non-uniform input distributions.

Acknowledgements We thank Peter Bartlett for a helpful discussion.

References

Abbe, E. and Sandon, C. On the universality of deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20061–20072, 2020a.

Abbe, E. and Sandon, C. Poly-time universality and limitations of deep learning. arXiv:2001.02992, 2020b.

Abbe, E., Kamath, P., Malach, E., Sandon, C., and Srebro, N. On the power of differentiable learning versus PAC and SQ learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Allen-Zhu, Z. and Li, Y. Backward feature correction: How deep learning performs deep learning. arXiv:2001.04413, 2020.

Blum, A., Furst, M., Jackson, J., Kearns, M., Mansour, Y., and Rudich, S. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Symposium on Theory of Computing (STOC)*, pp. 253–262, 1994.

Boix-Adsera, E. Personal communication, 2021.

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

Liu, R., Lehman, J., Molino, P., Such, F. P., Frank, E., Sergeev, A., and Yosinski, J. An intriguing failing of convolutional neural networks and the CoordConv solution. In *NeurIPS*, pp. 9628–9639, 2018. URL http://dblp.uni-trier.de/db/conf/nips/nips2018.html#LiuLMSFSY18.

Malach, E. and Shalev-Shwartz, S. Is deeper better only when shallow is good? In *Advances in Neural Information Processing Systems*, volume 32, pp. 6429–6438, 2019.

Malach, E., Kamath, P., Abbe, E., and Srebro, N. Quantifying the benefit of using differentiable learning over tangent kernels. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7379–7389. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/malach21a.html.

O'Donnell, R. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

Shalev-Shwartz, S. and Malach, E. Computational separation between convolutional and fully-connected networks. In *International Conference on Learning Representations (ICLR)*, 2021.

Tan, Y. S., Agarwal, A., and Yu, B. A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds, 2021. URL https://arxiv.org/abs/2110.09626.

Winkelbauer, A. Moments and absolute moments of the normal distribution. *ArXiv*, abs/1209.4340, 2012.

A. Proof of Proposition 4.3

For an activation $\sigma : \mathbb{R} \to \mathbb{R}$, we denote its v-Gaussian smoothing as

$$\Sigma_v(t) := \mathbb{E}_{Y\sim\mathcal{N}(0,v)}[\sigma(Y+t)]. \tag{39}$$

We also write $\Sigma := \Sigma_1$ for brevity. As mentioned, we will be working with functions that are *polynomially bounded*, ie., such that there exists a polynomial P with $|\sigma(x)| < P(x)$ holding for all $x \in \mathbb{R}$. We will use the fact that such polynomial can be assumed wlog to be of the form $|\sigma(x)| < Cx^\ell + C$ for some $C > 0$ and $\ell \in \mathbb{N}_{\geq 0}$ (since any polynomial can be upper bounded by a polynomial of such form). Note that if σ is a measurable, polynomially bounded function, then Σ_v is well defined for every $v > 0$.

We now state the intermediate step in the proof of Proposition 4.3:

Lemma A.1 (Conditions on Σ and Σ_v)**.** *If σ is a measurable, polynomially bounded function, then it satisfies the following conditions:*

i) $\Sigma_v \in C^\infty(\mathbb{R})$ *for every* $v > 0$*;*

ii) For every $k \in \mathbb{N}_{\geq 0}$ *and* $v > 0$, $\Sigma_v^{(k)}(t) := \frac{\partial^k}{\partial t^k}\Sigma_v(t)$ *is polynomially bounded. Furthermore, this bound is uniform, that is,* $|\Sigma_v^{(k)}(t)| < Ct^\ell + C$ *holds for every* $t \in \mathbb{R}$ *and every* $1/2 \leq v \leq 1$*, for some* C, ℓ *that do not depend on* v*.*

iii) For all $k \in \mathbb{N}_{\geq 0}$*, it holds* $|\Sigma_{1-\epsilon}^{(k)}(0) - \Sigma^{(k)}(0)| = O(\epsilon)$*.*

Lemma A.1 is then used in the proof of

Lemma A.2. *Let σ be expressive (according to Definition 2.4). Then, for every $k \geq 0$ and $P \geq k$ such that $\Sigma^{(P)}(0) \neq 0$, it holds that* $\mathrm{INAL}(M_k, \sigma) = \Omega(n^{-P})$.

In particular, from Lemma A.2 it follows that if σ is expressive, then it is correlating. Furthermore, since by condition b) in Definition 2.4 for every k we have $\Sigma^{(k)}(0) \neq 0$ or $\Sigma^{(k+1)}(0) \neq 0$, by Lemma A.2 it holds $\mathrm{INAL}(M_k, \sigma) = \Omega(n^{-(k+1)})$ and σ is 1-strongly correlating.

In the following subsections we prove Lemma A.1 and Lemma A.2.

A.1. Proof of Lemma A.1

In the following let ϕ_v denote the density function of $\mathcal{N}(0, v)$, ie., $\phi_v(t) = \frac{1}{\sqrt{2v\pi}} \exp\left(-\frac{t^2}{2v}\right)$. Note the relation to the standard Gaussian density $\phi = \phi_1$ where $\phi_v(t) = \frac{1}{\sqrt{v}}\phi(t/\sqrt{v})$.

We recall some useful facts about the derivatives of ϕ_v. First, it is well known that for ϕ it holds $\phi^{(k)}(t) = P_k(t)\phi(t)$ for some polynomial P_k of degree k. This formula extends to ϕ_v according to

$$\phi_v^{(k)}(t) = \frac{1}{\sqrt{v}}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\phi(t/\sqrt{v}) = v^{-k/2-1/2}\phi^{(k)}(t/\sqrt{v}) = v^{-k/2-1/2}P_k(t/\sqrt{v})\phi(t/\sqrt{v}) \tag{40}$$

$$= v^{-k/2}P_k(t/\sqrt{v})\phi_v(t)\ . \tag{41}$$

i) Let us write $\phi_v(t) = (\phi_{v/2} * \phi_{v/2})(t)$ where $*$ denotes the convolution in $\mathbb{R}$, i.e. $(g * h)(y) = \int_{\mathbb{R}} g(x)h(y-x)dx$. Thus,

$$\Sigma_v = \sigma * \phi_v = (\sigma * \phi_{v/2}) * \phi_{v/2}. \tag{42}$$

Now, $\sigma * \phi_{v/2}$ is in $L_1(\mathbb{R})$, since σ is measurable and polynomially bounded. Furthermore, $\phi_{v/2}$ is in $L_1(\mathbb{R})$ and $C^\infty(\mathbb{R})$. Therefore, by formulas for derivatives of convolution, $\Sigma_v \in C^\infty(\mathbb{R})$.

ii) Let us start with the claim that $\Sigma_v^{(k)}$ is polynomially bounded for every v and k. For that, we recall some facts. First, it is easy to establish by direct computation that if σ is polynomially bounded, then $\Sigma_v = \sigma * \phi_v$ is also polynomially bounded. Furthermore, if P is any polynomial, then also $\sigma * (P\phi_v)$ is polynomially bounded (this can be seen, eg., by observing that for every P and every $v' > v$ there exists C such that $|P\phi_v| \leq C\phi_{v'}$).

Accordingly, using (40) and (42) we have that

$$\Sigma_v^{(k)} = (\sigma * \phi_{v/2}) * \phi_{v/2}^{(k)} = (\sigma * \phi_{v/2}) * (P_{k,v}\phi_{v/2}) \tag{43}$$

is polynomially bounded.

Let us move to the second claim with uniform bound. For that let $k \geq 0$ and $1/2 \leq v \leq 1$. Let $v' := v - 1/4$ and note that $1/4 \leq v' \leq 3/4$. Then, we have the sequence of bounds on functions which hold pointwise:

$$|\Sigma_v^{(k)}| = \left|(\sigma * \phi_{1/4}) * \phi_{v'}^{(k)}\right| \leq C_1 \left(|\sigma * \phi_{1/4}| * |P_k(x/\sqrt{v'})|\phi_{v'}\right) \tag{44}$$

$$\leq C_1 \left(|\sigma * \phi_{1/4}| * (C_2 + C_2(x/\sqrt{v})^{2\ell})\phi_{v'}\right) \tag{45}$$

$$\leq C_3 \left(|\sigma * \phi_{1/4}| * (C_4 + C_4 x^{2\ell})\phi\right) \ , \tag{46}$$

which is now bounded by a polynomial which does not depend on v.

iii) Recall,

$$\Sigma_v^{(k)}(0) = \int_{-\infty}^{\infty} (\phi_{v/2} * \sigma)(x) \cdot \frac{\partial^k}{\partial t^k}\phi_{v/2}(x+t)\Big|_{t=0} dx, \tag{47}$$

where we denoted by $\phi_{v/2}^{(k)}$ the k-th derivative of $\phi_{v/2}$. Firstly, note that

$$\frac{\partial^k}{\partial t^k}\phi_{v/2}(x+t)\Big|_{t=0} = \frac{\partial^k}{\partial (x+t)^k}\phi_{v/2}(x+t)\Big|_{t=0} = \phi_{v/2}^{(k)}(x). \tag{48}$$

Let us give a formula for the k-th derivative of the Gaussian density:

$$\phi_v^{(k)}(x) = \phi_v(x) \cdot (-1)^k v^{-2k} \cdot \sum_{l=0}^{k} D_{l,k} \left(\frac{x}{\sqrt{v}}\right)^{k-l}, \tag{49}$$

where $D_{l,k}$ is a constant that does not depend on v, specifically

$$D_{l,k} := B_{(2k+l)\frac{1-(-1)^l}{2}} \cdot 2^{\frac{l}{2}} \cdot \frac{\Gamma(\frac{l+1}{2})}{\Gamma(\frac{1}{2})} \cdot \cos\left(\frac{l\pi}{2}\right) \tag{50}$$

where $\Gamma(.)$ denotes the Gamma function and B_n are the Bernoulli numbers. The exact values of the $D_{l,k}$ will not be relevant for this proof. Thus,

$$\Sigma_v^{(k)}(0) := \int_{-\infty}^{\infty} (\phi_{v/2} * \sigma)(x) P_{v/2,k}(x)\phi_{v/2}(x)dx, \tag{51}$$

where we denoted $P_{v/2,k}(x) = (-1)^k v^{-2k} \cdot \sum_{l=0}^{k} D_{l,k} \left(\frac{x}{\sqrt{v}}\right)^{k-l}$. On the other hand,

$$\Sigma_1^{(k)}(0) := \int_{-\infty}^{\infty} (\phi_{v/2} * \sigma)(x) P_{1-v/2,k}(x)\phi_{1-v/2}(x)dx, \tag{52}$$

and

$$|\Sigma_v^{(k)}(0) - \Sigma_1^{(k)}(0)| = \Big| \int_{-\infty}^{\infty} (\phi_{v/2} * \sigma)(x) \cdot \left(P_{v/2,k}(x)\phi_{v/2}(x) - P_{1-v/2,k}(x)\phi_{1-v/2}(x)\right) \Big|. \tag{53}$$

We note that

$$P_{1-v/2,k}(x) = \frac{(1-\frac{v}{2})^{-2k}}{(\frac{v}{2})^{-2k}} \left(\frac{v}{2}\right)^{-2k} (-1)^k \tag{54}$$

$$\cdot \left(\sum_{l=0}^{k} D_{l,k} \left(\frac{x}{\sqrt{v/2}} \right)^{k-l} + D_{l,k} \left[\left(\frac{x}{\sqrt{1-v/2}} \right)^{k-l} - \left(\frac{x}{\sqrt{v/2}} \right)^{k-l} \right] \right) \tag{55}$$

$$= \frac{(1-\frac{v}{2})^{-2k}}{(\frac{v}{2})^{-2k}} P_{v/2,k}(x) \tag{56}$$

$$+ \left(1-\frac{v}{2}\right)^{-2k} (-1)^k \sum_{l=0}^{k} D_{l,k} \left(\left(\frac{x}{\sqrt{1-v/2}} \right)^{k-l} - \left(\frac{x}{\sqrt{v/2}} \right)^{k-l} \right). \tag{57}$$

Recalling $\epsilon = 1 - v$, and expanding for such ϵ we get

$$\left(1+\frac{2\epsilon}{1-\epsilon}\right)^{-2k} P_{v/2,k}(x) + (1+\epsilon)^{-2k} \frac{(-1)^k}{2^{-2k}} \sum_{l=0}^{k} D_{l,k} x^{k-l} \frac{(1-\epsilon)^{k-l} - (1+\epsilon)^{k-l}}{(1+\epsilon)^{\frac{k-l}{2}} (1-\epsilon)^{\frac{k-l}{2}}} \tag{58}$$

$$= \left(1 - 4k\frac{\epsilon}{1-\epsilon} + o(\epsilon)\right) P_{v/2,k}(x) \tag{59}$$

$$+ (1 - 2k\epsilon + o(\epsilon)) \frac{(-1)^k}{2^{-2k}} \sum_{l=0}^{k} D_{l,k} x^{k-l} \frac{-2(k-l)\epsilon + (\epsilon)}{(1+\frac{k-l}{2}\epsilon + o(\epsilon))(1-\frac{k-l}{2}\epsilon + o(\epsilon))} \tag{60}$$

$$= \left(1 - 4k\frac{\epsilon}{1-\epsilon}\right) P_{v/2,k}(x) + O(\epsilon)\mathcal{P}_k(x), \tag{61}$$

where $\mathcal{P}_k(x)$ is a polynomial in x of degree $\leq k$. Moreover,

$$\phi_{1-v/2}(x) = \frac{e^{-\frac{x^2}{v}}}{\sqrt{2\pi v/2}} \cdot \sqrt{\frac{v/2}{1-v/2}} \cdot e^{-\frac{x^2}{2}\left(\frac{1}{1-\frac{v}{2}} - \frac{2}{v}\right)} \tag{62}$$

$$= \phi_{v/2}(x) \cdot \left(1 - \frac{2\epsilon}{1+\epsilon}\right)^{1/2} \cdot e^{x^2 \frac{2\epsilon}{(1+\epsilon)(1-\epsilon)}} \tag{63}$$

$$= \phi_{v/2}(x) \cdot \left(1 - \frac{\epsilon}{1+\epsilon} + o(\epsilon)\right) \cdot \left(1 + x^2 \frac{2\epsilon}{(1+\epsilon)(1-\epsilon)} + o(\epsilon)x^4\right) \tag{64}$$

$$= \phi_{v/2}(x) \cdot \left(1 + (x^2-1)O(\epsilon)\right). \tag{65}$$

Plugging these bounds in the previous expression, we get

$$|\Sigma_v^{(k)}(0) - \Sigma_1^{(k)}(0)| \tag{66}$$

$$= \left| \int_{-\infty}^{\infty} (\phi_{v/2} * \sigma)(x) \cdot \left(P_{v/2,k}(x)\phi_{v/2}(x) - P_{1-v/2,k}(x)\phi_{v/2}(x) \left(1 + (x^2-1)O(\epsilon)\right)\right) \right| \tag{67}$$

$$= \left| \int_{-\infty}^{\infty} (\phi_{v/2} * \sigma)(x)\phi_{v/2}(x) \cdot \left(P_{v/2,k}(x) - P_{1-v/2,k}(x) \left(1 + (x^2-1)O(\epsilon)\right)\right) \right| \tag{68}$$

$$= \left| \int_{-\infty}^{\infty} (\phi_{v/2} * \sigma)(x) P_{v/2,k}(x)\phi_{v/2}(x) \cdot \left(1 - (1 - O(\epsilon) + (x^2-1)O(\epsilon)) + O(\epsilon)\mathcal{P}_k(x)\right) \right| \tag{69}$$

$$= O(\epsilon)\,. \qquad \square \tag{70}$$

A.2. Proof of Lemma A.2

Note that we only need to show that $\mathrm{INAL}(M_k, \sigma) = \Omega(n^{-P})$ for the first index P such that $P \geq k$ and $\Sigma^{(P)}(0) \neq 0$. By Definition 2.4, we only need with two cases $P = k$ and $P = k+1$. From now on, let us consider a fixed pair of k and P.

We denote by $x \in \{\pm 1\}^n$ the vector of all inputs, by $w \in \mathbb{R}^n$ the vector of all weights and by $b \in \mathbb{R}$ the bias. Additionally, we denote $\tau_i := \operatorname{sgn}(w_i)$, and by $\tau \in \{\pm 1\}^n$ the vector of all weight signs. Recall that we consider $w_i, b \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{n})$ and that for $g, h : \{\pm 1\}^n \to \{\pm 1\}$ and $\mathcal{U}^n$ being the uniform distribution over the hypercube, we denote $\langle g, h \rangle = \mathbb{E}_{x \sim \mathcal{U}^n}[g(x)h(x)]$. We have

$$\begin{aligned} \mathrm{INAL}(M_k, \sigma) &= \mathbb{E}_{w,b}\left[\langle M_k, \sigma \rangle^2\right] && (71)\\ &= \mathbb{E}_{|w|,\tau,|b|,\operatorname{sgn}(b)}\left[\langle M_k, \sigma \rangle^2\right] && (72)\\ &\overset{(C.S.)}{\geq} \mathbb{E}_{\tau,\operatorname{sgn}(b)}\left[\mathbb{E}_{|w|,|b|}\left[\langle M_k, \sigma \rangle \mid \tau, \operatorname{sgn}(b)\right]^2\right], && (73) \end{aligned}$$

where (73) follows by Cauchy-Schwartz inequality. We will prove a lower bound on the inner expectation $\left(\mathbb{E}_{|w|,|b|}\langle M_k, \sigma \rangle\right)^2$ which is independent of τ and $\operatorname{sgn}(b)$. Accordingly, from now on consider τ and $\operatorname{sgn}(b)$ to be fixed at arbitrary values.

Let $T := \{1, \ldots, k\}$ and denote by x_T the coordinates of x contained in T, and by $x_{\sim T} := x_{T^C}$ the coordinates of x that are not contained in T and hence do not appear in the monomial M_T. Similarly, we denote by $|w|_T, |w|_{\sim T}$ the coordinates of $|w|$ that appear (respectively do not appear) in set T. We proceed,

$$\begin{aligned} \mathbb{E}_{|w|,|b|}\langle M_T, \sigma \rangle &= \mathbb{E}_{x,|w|,|b|}\left[M_T(x) \cdot \sigma\left(\sum_{i\in[n]} w_i x_i + b\right)\right] && (74)\\ &= \mathbb{E}_{|w|_T, x_T, |b|}\left[M_T(x) \cdot \mathbb{E}_{|w|_{\sim T}, x_{\sim T}} \sigma\left(\sum_{i\in[n]} w_i x_i + b\right)\right] && (75) \end{aligned}$$

Observe that $\sum_{i \notin T} w_i x_i \sim \mathcal{N}(0, \frac{n-k}{n})$, and denote $\Sigma_n(z) := \Sigma_{1-\frac{k}{n}}(z) = \mathbb{E}_{Y \sim \mathcal{N}(0, \frac{n-k}{n})}[\sigma(z + Y)]$. Moreover, let $G := \sum_{i \in T} w_i x_i + b$. Then,

$$\mathbb{E}_{|w|,|b|}\langle M_T, \sigma \rangle = \mathbb{E}_{|w|_T, |b|, x_T}\left[M_T(x)\Sigma_n\left(G\right)\right]. \tag{76}$$

Since, by condition i) in Lemma A.1, function Σ_n is C^∞ and therefore C^P, we apply Taylor's theorem with Lagrange remainder and write

$$\Sigma_n(z) = \sum_{\nu=0}^{P} a_{\nu,n} z^\nu + R_{P,n}(z), \tag{77}$$

where $a_{\nu,n} = \frac{\Sigma_n^{(\nu)}(0)}{\nu!}$ and

$$R_{P,n}(z) = \frac{\Sigma_n^{(P+1)}(\xi_z)}{(P+1)!} z^{P+1} \qquad \text{for some } |\xi_z| \leq |z|. \tag{78}$$

Plugging this in (76), we get

$$\mathbb{E}_{|w|,|b|}\langle M_T, \sigma \rangle = \sum_{\nu=0}^{P} a_{\nu,n} \mathbb{E}_{|w|_T,|b|,x_T}\left[M_T(x)G^\nu\right] + \mathbb{E}_{|w|_T,|b|,x_T}\left[M_T(x)R_{P,n}\left(G\right)\right]. \tag{79}$$

The following two propositions give the asymptotic characterization of the first and second term in (79).

Proposition A.3.

$$\sum_{\nu=0}^{P} a_{\nu,n} \mathbb{E}_{|w|_T,|b|,x_T}\left[M_T(x)G^\nu\right] = C(P)(-1)^{C'(\tau_T, \operatorname{sgn}(b))} n^{-P/2} + O(n^{-P/2-1/2}). \tag{80}$$

where $C(P) \neq 0$ and $C'(\tau_T, \operatorname{sgn}(b)) \in \mathbb{Z}$ are constants that do not depend on n.

Proposition A.4.

$$\mathbb{E}_{|w|_T,|b|,x_T}\left[M_T(x)R_{P,n}\left(G\right)\right] = O(n^{-P/2-1/2}). \tag{81}$$

Before proving Propositions A.3 and A.4, let us see how Lemma A.2 follows from them. But this is clear: substituting into (79), we have

$$\left(\mathbb{E}_{|w|,|b|}\langle M_T,\sigma\rangle\right)^2 = C(P)^2 n^{-P} + O(n^{-P-1/2}) = \Omega(n^{-P}) \ , \tag{82}$$

where the claimed bound does not depend on τ nor on $\mathrm{sgn}(b)$.

A.2.1. Proof of Proposition A.3

The main step for proving Proposition A.3 is the computation of $\langle M_T, G^\nu\rangle$, for $\nu \le P$. This is summarized in the following formula.

Lemma A.5. *We have:*

$$\mathbb{E}_{|w|_T,|b|,x_T} M_T(x)G^\nu = \begin{cases} 0 & \text{if } \nu < k \\ C(\nu)(-1)^{C'(\tau_T,\mathrm{sgn}(b))}n^{-\nu/2} & \text{if } \nu \ge k \ , \end{cases} \tag{83}$$

where $C(\nu) > 0$.

Let us first see how to finish the proof once Lemma A.5 is established. Recall that $a_{\nu,n} = \frac{\Sigma^{(\nu)}_{1-k/n}(0)}{\nu!}$ and let $a_\nu := \frac{\Sigma^{(\nu)}(0)}{\nu!}$. We are considering a sum with $P+1$ terms, so let $s_\nu := a_{\nu,n}\mathbb{E}_{|w|_T,|b|,x_T}\left[M_T(x)G^\nu\right]$. Accordingly, our objective is to show that

$$\sum_{\nu=0}^{P} s_\nu = C(P)(-1)^{C'(\tau_T,\mathrm{sgn}(b))}n^{-P/2} + O(n^{-P/2-1/2}) \ . \tag{84}$$

We do that by considering the terms s_ν one by one. For $\nu < k$, from (83) we immediately have $s_\nu = 0$.

For $k \le \nu < P$, by Definition 2.4 recall that the only possible case is $P = k+1$ and $\Sigma^{(k)}(0) = 0$. Then, applying condition iii) from Lemma A.1,

$$|a_{\nu,n}| = \left|\frac{\Sigma^{(\nu)}_{1-k/n}(0) - \Sigma^{(\nu)}(0)}{\nu!}\right| = O(n^{-1}) \ , \tag{85}$$

which together with (83) gives $|s_v| = O(n^{-P/2-1/2})$.

Finally, for $\nu = P$, by assumption we have $a_P \neq 0$. Then, by condition iii), we have $|a_{P,n} - a_P| = O(1/n)$ and (83) gives us the correct form for s_P and the whole expression.

All that is left is the proof of Lemma A.5.

Proof of Lemma A.5. The proof proceeds by using the linearity of expectation and independence and expanding the formula for G^ν. Recall that we assumed wlog that $T = \{1,\dots,k\}$ and let $z_i := w_i x_i$ for $i \le k$ and $z_{k+1} := b$:

$$\mathbb{E}_{|w|_T,|b|,x_T} M_T(x)G^\nu = \mathbb{E}_{|w|_T,|b|,x_T}\left(\prod_{i=1}^{k} x_i\right)\left(\sum_{i=1}^{k} w_i x_i + b\right)^\nu \tag{86}$$

$$= \sum_{I=(i_1,\dots,i_\nu)\in[k+1]^\nu} \mathbb{E}_{|w|_T,|b|,x_T}\left(\prod_{i=1}^{k} x_i\right)\left(\prod_{i\in I} z_i\right) \ . \tag{87}$$

Let us focus on a single term of the sum in (87) for $I = (i_1, \ldots, i_\nu) \in [k+1]^\nu$. For $j = 1, \ldots, k+1$, let $\alpha_j = \alpha_j(I) := |\{m : i_m = j\}|$. Accordingly, we can rewrite a term from (87) as

$$\mathbb{E}_{|w|_T,|b|,x_T} \left(\prod_{i=1}^{k} x_i\right)\left(\prod_{i\in I} z_i\right) \tag{88}$$

$$= \mathbb{E}_{|w|_T,|b|,x_T} \left(\prod_{i=1}^{k} w_i^{\alpha_i} x_i^{\alpha_i+1}\right) b^{\alpha_{k+1}} \tag{89}$$

$$= \left(\prod_{i=1}^{k} \tau_i^{\alpha_i}\right) \operatorname{sgn}(b)^{\alpha_{k+1}} \left(\prod_{i=1}^{k} \mathbb{E}_{|w|_i}\left[|w|_i^{\alpha_i}\right] \cdot \mathbb{E}_{x_i}\left[x_i^{\alpha_i+1}\right]\right) \mathbb{E}_{|b|}\left[|b|^{\alpha_{k+1}}\right]. \tag{90}$$

Since $\mathbb{E}[x_i^{\alpha_i+1}] = 0$ if α_i is even, for a term in (90) to be non-zero it is necessary that α_i is odd for every $1 \le i \le k$. Consequently, since $\sum_{i=1}^{k+1} \alpha_i = \nu$, in any non-zero term the parity of α_{k+1} is equal to the parity of $\nu - k$. Therefore, every non-zero term is of the form

$$\mathbb{E}_{|w|_T,|b|,x_T} \left(\prod_{i=1}^{k} x_i\right)\left(\prod_{i\in I} z_i\right) = \left(\prod_{i=1}^{k} \tau_i\right) \operatorname{sgn}(b)^{\mathbb{1}[\nu-k \text{ odd}]} \cdot \left(\prod_{i=1}^{k} \mathbb{E}_{|w|_i}\left[|w_i|^{\alpha_i}\right]\right) \mathbb{E}_{|b|}\left[|b|^{\alpha_{k+1}}\right] \tag{91}$$

$$= (-1)^{C'(\tau_T, \operatorname{sgn}(b))} \cdot \left(\prod_{i=1}^{k} \mathbb{E}_{|w|_i}\left[|w_i|^{\alpha_i}\right]\right) \mathbb{E}_{|b|}\left[|b|^{\alpha_{k+1}}\right]. \tag{92}$$

We now establish the first case from (83). If $\nu < k$, then since $\nu = \sum_{i=1}^{k+1} \alpha_i$ at least one of α_i, $1 \le i \le k$ must be zero, and therefore even. Consequently, each term in (87) is zero and it follows that $\mathbb{E}_{|w|_T,|b|,x_T} M_T(x) G^\nu = 0$.

On the other hand, for $\nu \ge k$, there exists a non-zero term, for example taking $\alpha_1 = \ldots = \alpha_k = 1$ and $\alpha_{k+1} = \nu - k$. Take any such term arising from $I \in [k+1]^\nu$. Since $w_i, b \sim \mathcal{N}(0, 1/n)$, we have $\mathbb{E}_{|w|_i}\left[|w_i|^j\right], E_{|b|}\left[|b|^j\right] = C_j \cdot n^{-j/2}$ for some $C_j > 0$ for every fixed j. Substituting in (92) and using $\nu = \sum_{i=1}^{k+1} \alpha_i$, we get

$$\mathbb{E}_{|w|_T,|b|,x_T} \left(\prod_{i=1}^{k} x_i\right)\left(\prod_{i\in I} z_i\right) = (-1)^{C'(\tau_T, \operatorname{sgn}(b))} C_I n^{-\nu/2} \tag{93}$$

for some $C_I > 0$. Therefore, $C(\nu)(-1)^{C'(\tau_T, \operatorname{sgn}(b))} n^{-\nu/2}$ with $C(\nu) > 0$ follows since it is a sum of at most $(k+1)^\nu$ positive terms. □

A.2.2. Proof of Proposition A.4

Let D be a positive constant. We apply the decomposition

$$\left|\mathbb{E}_{|w|_T,|b|,x_T}\left[M_T(x) R_{P,n}(G)\right]\right| \le \mathbb{E}_{|w|_T,|b|,x_T}\left[\left|R_{P,n}(G)\right| \cdot \mathbb{1}(|G| \le D)\right] \tag{94}$$

$$+ \mathbb{E}_{|w|_T,|b|,x_T}\left[\left|R_{P,n}(G)\right| \cdot \mathbb{1}(|G| > D)\right] \tag{95}$$

The proposition follows from Lemmas A.6 and A.7 applied to an arbitrary value of D, eg., $D = 1$.

Lemma A.6. *For any $D > 0$,*

$$\mathbb{E}_{|w|_T,|b|,x_T}\left[\left|R_{P,n}(G)\right| \cdot \mathbb{1}(|G| \le D)\right] = O\left(n^{-\frac{P+1}{2}}\right). \tag{96}$$

Proof. Let us observe that for a fixed b, $G \sim \mathcal{N}(b, \frac{k}{n})$, thus

$$\mathbb{E}_{|w|_T,x_T}\left[|R_{P,n}(G)| \mathbb{1}(|G| \le D)\right] = \mathbb{E}_{y\sim\mathcal{N}(b,\frac{k}{n})}\left[|R_{P,n}(y)| \mathbb{1}(|y| \le D)\right]. \tag{97}$$

Recall that $R_{P,n}(x) = \frac{\Sigma_n^{(P+1)}(\xi_x)}{(P+1)!}x^{P+1}$ for some $|\xi_x| \leq |x|$. Thus,

$$\mathbb{E}_{y\sim\mathcal{N}(b,\frac{k}{n})}\left[|R_{P,n}(y)|\mathbb{1}(|y|\leq D)\right] \leq \sup_{|y|\leq D}\frac{|\Sigma_n^{(P+1)}(y)|}{(P+1)!}\cdot\mathbb{E}_{y\sim\mathcal{N}(b,\frac{k}{n})}|y|^{P+1}. \tag{98}$$

On the one hand, assuming that $n \geq 2k$, we have $\Sigma_n = \Sigma_v$ for some $1/2 \leq v \leq 1$, and thus using the common polynomial bound in property ii) $\sup_{|y|\leq D}|\Sigma_n^{(P+1)}(y)| \leq M_D$, where the constant M_D does not depend on n. On the other hand,

$$\begin{aligned}
\mathbb{E}_{y\sim\mathcal{N}(b,\frac{k}{n})}|y|^{P+1} &= n^{-\frac{P+1}{2}}\cdot\mathbb{E}_y\left|\sqrt{n}\cdot y\right|^{P+1} && (99)\\
&\leq n^{-\frac{P+1}{2}}\cdot 2^{P+1}\cdot\left(|\sqrt{n}b|^{P+1}+\mathbb{E}_{z\sim\mathcal{N}(0,k)}|z|^{P+1}\right) && (100)\\
&= n^{-\frac{P+1}{2}}\cdot 2^{P+1}\cdot\left(|\sqrt{n}b|^{P+1}+\frac{(2k)^{\frac{P+1}{2}}\Gamma(\frac{P+2}{2})}{\sqrt{\pi}}\right), && (101)
\end{aligned}$$

where in the last equation we plugged the (P+1)-th central moment of the Gaussian distribution (see, eg., (Winkelbauer, 2012)). Since $|\sqrt{n}b|$ is also distributed like an absolute value of $\mathcal{N}(0,1)$, taking the expectation over $|b|$, we get that for fixed P, k,

$$\mathbb{E}_{|w|_T,|b|,x_T}\left[\left|R_{P,n}(G)\right|\cdot\mathbb{1}(|G|\leq D)\right] = O\left(n^{-\frac{P+1}{2}}\right). \tag{102}$$

□

Lemma A.7. *For any constant $D > 0$, there exist $C_1, C_2 > 0$ such that*

$$\mathbb{E}_{|w|_T,|b|,x_T}\left[\left|R_{P,n}(G)\right|\cdot\mathbb{1}(|G|> D)\right] \leq C_1\exp(-C_2 n). \tag{103}$$

Proof. By Cauchy-Schwartz inequality,

$$\mathbb{E}_{|w|_T,|b|,x_T}\left[|R_{P,n}(G)|\mathbb{1}(|G|>D)\right] \overset{(C.S.)}{\leq} \mathbb{E}_{|w|_T,|b|,x_T}[R_{P,n}(G)^2]^{1/2}\cdot\Pr_{|w|_T,|b|,x_T}[|G|>D]^{1/2}. \tag{104}$$

For the first term, we use the universal polynomial bound from property ii):

$$\begin{aligned}
\left|\mathbb{E}_{|w|_T,|b|,x_T}[R_{P,n}(G)^2]\right| &= \mathbb{E}_{|w|_T,|b|,x_T}\left[\left(\frac{\sup_{|y|\leq|G|}\Sigma_n^{(P+1)}(y)}{(P+1)!}|G|^{P+1}\right)^2\right] && (105)\\
&\leq \mathbb{E}_{|w|_T,|b|,x_T}\left[\left(\frac{\sup_{|y|\leq|G|}Cy^{2\ell}+C}{(P+1)!}|G|^{P+1}\right)^2\right] && (106)\\
&= \mathbb{E}_{|w|_T,|b|,x_T}\left[\left(\frac{CG^{2\ell}+C}{(P+1)!}|G|^{P+1}\right)^2\right] = O_n(1), && (107)
\end{aligned}$$

using a similar reasoning as in Lemma A.6.

On the other hand, writing $G = G' + |b|$, we have

$$\begin{aligned}
\Pr_{|w|_T,|b|,x_T}[|G|>D] &\leq \Pr_{|b|}[|b|>D/2] + \Pr_{|w|_T,x_T}[|G'|>D/2] && (108)\\
&\leq 2\Pr_{y\sim\mathcal{N}(0,1/n)}[|y|>D/2] \leq 4\exp(-D^2 n/8). && (109)
\end{aligned}$$

We get desired bound putting together (105) and (109). □

B. Expressivity of Common Activation Functions

In this section we show that ReLU and sign are expressive. It is clear that both of these functions are polynomially bounded, so we only need to analyze their Hermite expansions for condition b) in Definition 2.4. In both cases we do it by writing a closed form for $\Sigma^{(k)}(0)$.

Proposition B.1. $\mathrm{ReLU}(x) := \max\{0, x\}$ *is expressive.*

Proof. We will see that in the case $\sigma = \mathrm{ReLU}$ we have $\Sigma(z) = \frac{z}{2} + \frac{z}{2}\operatorname{erf}(z) + \frac{1}{2\sqrt{\pi}}\exp(-z^2)$. Indeed,

$$\Sigma(z) = \int_{-\infty}^{\infty} \mathbb{1}(z+y \geq 0)(z+y)\phi(y)\,\mathrm{d}y = \int_{-z}^{\infty} (z+y)\phi(y)\,\mathrm{d}y = z\Phi(z) + \phi(z) \tag{110}$$

$$= \frac{z}{2} + \frac{z\operatorname{erf}(z/\sqrt{2})}{2} + \phi(z)\ . \tag{111}$$

Using well-known Taylor expansions of erf and ϕ, this results in

$$\Sigma^{(k)}(0) = \begin{cases} \frac{1}{2} & \text{if } k = 1\ , \\ \frac{(-1)^{k/2+1}}{\sqrt{2\pi}2^{k/2}(k-1)(k/2)!} & \text{if } k \text{ is even}, \\ 0 & \text{otherwise.} \end{cases} \tag{112}$$

In particular, $\Sigma^{(k)}(0) \neq 0$ for every even k and ReLU is expressive. □

Proposition B.2. *The sign function* $\mathrm{sgn}(x)$ *is expressive.*

Proof. In this case, similarly, we have

$$\Sigma(z) = -\int_{-\infty}^{-z} \phi(z)\,\mathrm{d}z + \int_{-z}^{\infty} \phi(z)\,\mathrm{d}z = 2\Phi(z) - 1 = \operatorname{erf}(z/\sqrt{2})\ , \tag{113}$$

which can be seen to have the expansion

$$\Sigma^{(k)}(0) = \begin{cases} \frac{2}{\sqrt{\pi}} \cdot \frac{(-1)^{(k-1)/2}}{2^{k/2}\left(\frac{k-1}{2}\right)!k} & \text{if } k \text{ is odd}, \\ 0 & \text{otherwise.} \end{cases} \tag{114}$$

Again, the sign function is expressive since $\Sigma^{(k)} \neq 0$ for every odd k. □

C. Proof of Proposition 4.4

Using the definition of INAL and the Fourier expansion of f, we get

$$\mathrm{INAL}(f, \sigma) = \mathbb{E}_{w,b}\left[\langle f, \sigma\rangle^2\right] \tag{115}$$

$$= \mathbb{E}_{w,b}\left[\left(\sum_{T\in[n]} \hat{f}(T)\langle M_T, \sigma\rangle\right)^2\right] \tag{116}$$

$$= \mathbb{E}_{w,b}\left[\sum_{T} \hat{f}(T)^2\langle M_T, \sigma\rangle^2 + \sum_{S\neq T} \hat{f}(S)\hat{f}(T)\langle M_T, \sigma\rangle\langle M_S, \sigma\rangle\right]. \tag{117}$$

We show that the second term of (117) is zero. Let S, T be two distinct sets. Without loss of generality, assume that $|S| \geq |T|$, and let i be such that $i \in S$ but $i \notin T$ (such i must exist since $S \neq T$). Fix w and b and decompose w into

$w_{\sim i}, |w_i|, \operatorname{sgn}(w_i)$ where $w_{\sim i}$ denotes the vector of weights, excluding coordinate i. By applying the change of variable $\operatorname{sgn}(w_i)x_i \mapsto y_i$ and noticing that x_i has the same distribution of y_i, we then get

$$\langle M_S, \sigma\rangle = \mathbb{E}_x[M_S(x) \cdot \sigma(x_i \operatorname{sgn}(w_i)|w_i| + \sum_{j\neq i} x_j w_j + b)] \tag{118}$$

$$= \operatorname{sgn}(w_i) \cdot \mathbb{E}_{x_{\sim i}, y_i}[M_{S_{\sim i}}(x) \cdot y_i \cdot \sigma(y_i|w_i| + \sum_{j\neq i} x_j w_j + b)] \tag{119}$$

$$:= \operatorname{sgn}(w_i) \cdot E_S, \tag{120}$$

where E_S does not depend on $\operatorname{sgn}(w_i)$. On the other hand,

$$\langle M_T, \sigma\rangle = \mathbb{E}_x[M_T(x) \cdot \sigma(x_i \operatorname{sgn}(w_i)|w_i| + \sum_{j\neq i} x_j w_j + b)] \tag{121}$$

$$= \mathbb{E}_{x_{\sim i}, y_i}[M_T(x) \cdot \sigma(y_i|w_i| + \sum_{j\neq i} x_j w_j + b)], \tag{122}$$

which means that $\langle M_T, \sigma\rangle$ does not depend on $\operatorname{sgn}(w_i)$. Thus, we get

$$\mathbb{E}_{w,b}\left[\langle M_T, \sigma\rangle \langle M_S, \sigma\rangle\right] = \mathbb{E}_{w,b}\left[\operatorname{sgn}(w_i) \cdot \langle M_T, \sigma\rangle \cdot E_S\right] = 0. \tag{123}$$

Hence,

$$\operatorname{INAL}(f, \sigma) = \sum_T \hat{f}(T)^2 \mathbb{E}_{w,b}\left[\langle M_T, \sigma\rangle^2\right] \tag{124}$$

$$= \sum_T \hat{f}(T)^2 \operatorname{INAL}(M_T, \sigma). \tag{125}$$

D. Proof of Corollary 4.5

Indeed, by Proposition 4.4 for any $f : \{\pm 1\}^n \to \mathbb{R}$ and k it holds

$$\operatorname{INAL}(f, \sigma) = \sum_T \hat{f}(T)^2 \operatorname{INAL}(M_T, \sigma) \geq W^k(f) \operatorname{INAL}(M_k, \sigma)\ . \tag{126}$$

Accordingly, if $\operatorname{INAL}(M_k, \sigma) = \Omega(n^{-k_0})$, we have

$$W^k(f_n) \leq \operatorname{INAL}(f_n, \sigma) \cdot O(n^{k_0})\ , \tag{127}$$

and then, under our assumptions, also

$$W^{\leq k}(f_n) \leq \operatorname{INAL}(f_n, \sigma) \cdot O(n^{k_0})\ . \tag{128}$$

For the "in particular" statement, let (f_n) be a function family with negligible $\operatorname{INAL}(f_n, \sigma)$ for a correlating σ. Let $k \in \mathbb{N}$. Since σ is correlating, the assumption $\operatorname{INAL}(M_{k'}, \sigma) = \Omega(n^{-k_0})$ for $k' = 0, \ldots, k$ holds. Therefore, (128) also holds and $W^{\leq k}(f)$ is negligible. Since k was arbitrary, the function family (f_n) is high-degree.

E. Details and Proof of Corollary 3.4

Corollary 3.4 states a hardness results for learning on *fully connected neural networks with iid initialization*. This is a more specific definition than the one we gave for a neural network in Section 2. Let us state it precisely, following the treatment in (Abbe & Sandon, 2020b).

Definition E.1. For the purposes of Corollary 3.4, a neural network on n inputs consists of a differentiable activation function $\sigma : \mathbb{R} \to \mathbb{R}$, a threshold function $f : \mathbb{R} \to \{\pm 1\}$ and a weighted, directed graph with a vertex set labeled with $\{1, x_1, \ldots, x_n, v_1, \ldots, v_m, v_{\text{out}}\}$. The vertices labeled with $x_1, \ldots, x_n$ are called the input vertices, the vertex labeled with 1 is the constant vertex and v_{out} is the output vertex.

We assume that the graph does not contain loops, the constant and input vertices do not have any incoming edges, the output vertex does not have outgoing edges and for the remaining vertices there are no edges (v_i, v_j) for $i > j$. Each vertex (a neuron) has an associated function (the output of the neuron) from $\mathbb{R}^n$ to $\mathbb{R}$ which is defined recursively as follows: The output of the constant vertex is $y_1 = 1$ and the output of the input vertex is (abusing notation) $y_{x_i} = x_i$. The output of any other vertex v_i is given by $y_{v_i} = \sigma(\sum_{v:(v,v_i)\in E(G)} w_{v,v_i} y_v)$. Finally, the output of the whole network is given by $f(y_{v_{\mathrm{out}}})$.

We say that the neural network is *fully connected* if every vertex that has an incoming edge from an input vertex has incoming edges from all input vertices.

Note that our definition of "fully connected network" covers any feed-forward architecture that consists of a number of fully connected hidden layers stacked on top of each other.

Let us restate Theorem 3 from (Abbe & Sandon, 2020b) with the bound[5] from their Corrolary 1 applied to the junk flow term JF_T:

Theorem E.2 ((Abbe & Sandon, 2020b)). *Let $P_{\mathcal{F}}$ be a distribution on Boolean functions $f : \{\pm 1\}^n \to \{\pm 1\}$. Consider any neural network as defined in Definition E.1 with E edges. Assume that a function f is chosen from $P_{\mathcal{F}}$ and then T steps of noisy GD with learning rate γ, overflow range A and noise level τ are run on the initial network using function f and uniform input distribution $\mathcal{U}^n$.*

Then, in expectation over the initial choice of f, the training noise, and a fresh sample $x \sim \mathcal{U}^n$, the trained neural network $\mathrm{NN}^{(T)}$ satisfies

$$\Pr\left[\mathrm{NN}^{(T)}(x) = f(x)\right] \leq \frac{1}{2} + \frac{\gamma T\sqrt{E}A}{\tau} \cdot \mathrm{CP}(P_{\mathcal{F}}, \mathcal{U}^n)^{1/4} . \tag{129}$$

Finally, we need to discuss the fact that Corollary 3.4 applies for any fully connected neural network *with iid initialization*. What we mean by this is that the initial neural network has a fixed activation σ, threshold function f and graph (vertices and edges), but the weights on edges are not fixed. Instead, they are chosen randomly iid from any fixed probability distribution. More precisely, we can make a weaker assumption that the weights on all edges that are outgoing from the input vertices are chosen[6] iid from a fixed distribution and all the other weights have arbitrary fixed values.

We can now proceed to prove Corollary 3.4.

E.1. Proof of Corollary 3.4

Let randomly initialized, fully connected neural network NN be trained in the following way. First, a function $\overline{f_n} \circ \pi$ is chosen uniformly at random from the orbit of $\overline{f_n}$. Then, a noisy GD algorithm is run with the parameters stated: T steps, learning rate γ, overflow range A and noise level τ. Finally, a fresh sample $x \sim \mathcal{U}^N$ is presented to the trained neural network. Then, Theorem E.2 says that

$$\Pr\left[\mathrm{NN}^{(T)}(x) = (\overline{f_n} \circ \pi)(x)\right] \leq \frac{1}{2} + \frac{\gamma T\sqrt{E}A}{\tau} \cdot \mathrm{CP}(\mathrm{orb}(\overline{f_n}), \mathcal{U}^N)^{1/4} . \tag{130}$$

Since we can apply Theorem E.2 to the class of all orbits of $-\overline{f_n}$, which has the same cross-predictability, the same upper bound also holds for $\Pr[\mathrm{NN}^{(T)}(x) \neq (\overline{f_n} \circ \pi)(x)]$. Consequently, we have the expectation bound

$$\left|\mathbb{E}\langle \mathrm{NN}^{(T)}, \overline{f_n} \circ \pi\rangle\right| \leq \frac{2\gamma T\sqrt{E}A}{\tau} \cdot \mathrm{CP}(\mathrm{orb}(\overline{f_n}), \mathcal{U}^N)^{1/4} . \tag{131}$$

Recall that the neural network is fully connected and the weights on the edges outgoing from the input vertices are iid. The expectation in (131) is an average of conditional expectations for different initial choices of permutation π. Consider the action induced by π on the weights outgoing from the input vertices. By properties of GD, it follows that each conditional expectation over π contributes equally to the left-hand side of (131). It follows that the same bound holds also for the single

[5] Since we are discussing GD, we are applying their bound with infinite sample size $m = \infty$.

[6] Even more precisely, we can assume only that the distribution of these weights is symmetric under permutations of input vertices $x_1, \ldots, x_n$.

function $\overline{f_n}$:

$$\left|\mathbb{E}_{\mathrm{NN}^{(T)}}\langle \mathrm{NN}^{(T)}, \overline{f_n}\rangle\right| \leq \frac{2\gamma T\sqrt{E}A}{\tau}\,\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N)^{1/4}\,. \tag{132}$$

Accordingly, if $\mathrm{INAL}(f_n,\sigma)$ is negligible, then, by Theorem 3.3, $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N)$ is negligible and the right-hand side of (132) remains negligible for any polynomial bounds on γ, T, E, A and τ, as claimed.

For the more precise statement, if $\mathrm{INAL}(f_n,\sigma) = O(n^{-c})$, then again by Theorem 3.3 it holds $\mathrm{CP}(\mathrm{orb}(\overline{f_n}),\mathcal{U}^N) = O(n^{-\frac{\epsilon}{1+\epsilon}(c-1)})$ and we get the bound of $O\left(\frac{\gamma T\sqrt{E}A}{\tau}\cdot n^{-\frac{\epsilon}{4(1+\epsilon)}(c-1)}\right)$ on the right-hand side of (132). □