Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit

Jason D. Lee[1], **Kazusato Oko**[2,4], **Taiji Suzuki**[3,4], **Denny Wu**[5,6]

[1]Princeton University, [2]University of California, Berkeley, [3]University of Tokyo
[4]RIKEN AIP, [5]New York University, [6]Flatiron Institute

jasonlee@princeton.edu, oko@berkeley.edu,
taiji@mist.i.u-tokyo.ac.jp, dennywu@nyu.edu

Abstract

We study the problem of gradient descent learning of a single-index target function $f_*(\boldsymbol{x}) = \sigma_*(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)$ under isotropic Gaussian data in $\mathbb{R}^d$, where the unknown link function $\sigma_* : \mathbb{R} \to \mathbb{R}$ has information exponent p (defined as the lowest degree in the Hermite expansion). Prior works showed that gradient-based training of neural networks can learn this target with $n \gtrsim d^{\Theta(p)}$ samples, and such complexity is predicted to be necessary by the correlational statistical query lower bound. Surprisingly, we prove that a two-layer neural network optimized by an SGD-based algorithm (on the squared loss) learns f_* with a complexity that is not governed by the information exponent. Specifically, for arbitrary polynomial single-index models, we establish a sample and runtime complexity of $n \simeq T = \Theta(d \cdot \mathrm{polylog}\, d)$, where $\Theta(\cdot)$ hides a constant only depending on the degree of σ_*; this dimension dependence matches the information theoretic limit up to polylogarithmic factors. More generally, we show that $n \gtrsim d^{(p_*-1)\vee 1}$ samples are sufficient to achieve low generalization error, where $p_* \leq p$ is the *generative exponent* of the link function. Core to our analysis is the reuse of minibatch in the gradient computation, which gives rise to higher-order information beyond correlational queries.

1 Introduction

Single-index models are a classical class of functions that capture low-dimensional structure in the learning problem. To efficiently estimate such functions, the learning algorithm should extract the relevant (one-dimensional) subspace from high-dimensional observations; hence this problem setting has been extensively studied in deep learning theory [BL20, BES+22, BBSS22, MHPG+23, MZD+23, WWF24], to examine the adaptivity to low-dimensional targets and benefit of representation learning in neural networks (NNs) optimized by gradient descent (GD). In this work we study the learning of a single-index target function under isotropic Gaussian data:

$$y_i = f_*(\boldsymbol{x}_i) + \varsigma_i, \quad f_*(\boldsymbol{x}_i) = \sigma_*(\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle), \quad \boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_d), \tag{1.1}$$

where ς_i is i.i.d. label noise, $\boldsymbol{\theta} \in \mathbb{R}^d$ is the direction of index features, and we assume the link function $\sigma_* : \mathbb{R} \to \mathbb{R}$ has information exponent $p \in \mathbb{N}_+$ defined as the index of the first non-zero coefficient in the Hermite expansion (see Definition 1).

Equation (1.1) requires the estimation of the one-dimensional link function σ_* and the relevant direction $\boldsymbol{\theta}$; it is known that learning is information theoretically possible with $n \gtrsim d$ training examples [DH24, DPVLB24]. Indeed, when σ_* is polynomial, such statistical complexity can be achieved up to logarithmic factors by a tailored algorithm that exploit the structure of low-dimensional target [CM20]. On the other hand, for gradient-based training of two-layer NNs, existing works established a sample complexity of $n \gtrsim d^{\Theta(p)}$ [BAGJ21, BBSS22, DNGL23], which presents a gap
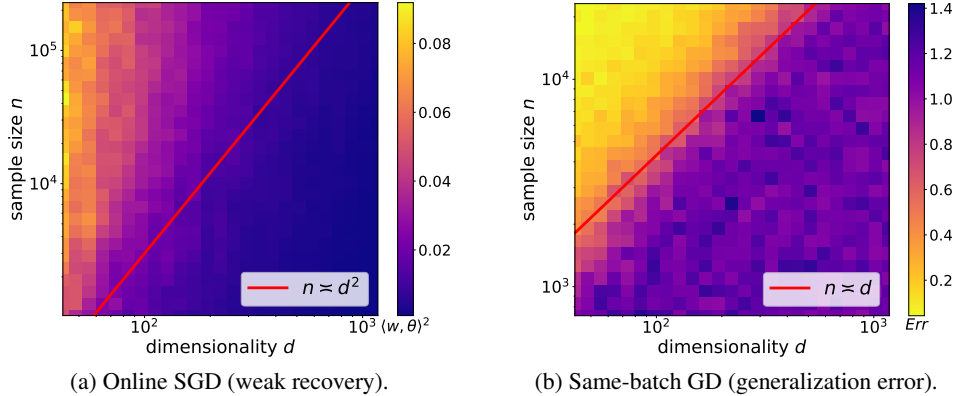
(a) Online SGD (weak recovery).

(b) Same-batch GD (generalization error).

Figure 1: We train a ReLU NN (3.1) with $N = 1024$ neurons using SGD (squared loss) with step size $\eta = 1/d$ to learn a single-index target $f_*(\boldsymbol{x}) = \mathsf{He}_3(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)$; heatmaps are values averaged over 10 runs. (a) online SGD with batch size $B = 8$; (b) GD on the same batch of size n for $T = 2^{14}$ steps. For online SGD we only report weak recovery (i.e., averaged overlap between neuron $\boldsymbol{w}$ and target $\boldsymbol{\theta}$) since the test error does not drop.

between the information theoretic limit and what is computationally achievable by (S)GD. Such a gap is also predicted by the correlational statistical query (CSQ) lower bound [DLS22, AAM23], which roughly states that for a CSQ algorithm to learn (isotropic) Gaussian single-index models using less than exponential compute, a sample size of $n \gtrsim d^{p/2}$ is necessary.

Although CSQ lower bounds are frequently cited to imply a fundamental barrier of learning via SGD (with the squared/correlation loss), strictly speaking, the CSQ model does not include empirical risk minimization with gradient descent, due to the non-adversarial noise and existence of non-correlational terms in the gradient computation. Very recently, [DTA+24] exploited higher-order terms in the gradient update arising from the reuse of the same training data, and showed that for certain link functions with high information exponent ($p > 2$), two-layer NNs may still achieve weak recovery (i.e., nontrivial overlap with $\boldsymbol{\theta}$) after two GD steps with $\Theta(d)$ batch size. While this presents evidence that GD-trained NNs can learn f_* beyond the sample complexity suggested by the CSQ lower bound, the weak recovery statement in [DTA+24] may not translate to statistical guarantees; moreover, the class of functions where SGD can achieve vanishing generalization error is not fully characterized, as only a few specific examples of link functions are discussed.

Given the existence of (non-NN) algorithms that learn any single-index polynomials in $n = \tilde{O}(d)$ samples [CM20] regardless of the information exponent p, and more generally, non-CSQ algorithms with a sample complexity surpassing the CSQ lower bound [DPVLB24], it is natural to ask if gradient-based training of NNs can achieve similar statistical efficiency for this function class. Motivated by observations in [DTA+24] that SGD with reused data may break the "curse of information exponent", we aim to address the question:

Can NN optimized by SGD with reused batch learn single-index f_* beyond the CSQ lower bound? And for polynomial σ_*, can learning succeed near the information-theoretic limit $n \simeq d$?

Empirically, the separation between one-pass (online) and multi-pass SGD is clearly observed in Figure 1, where we trained the same two-layer ReLU neural network to learn a single-index polynomial with information exponent $p = 3$. We see that SGD with reused data (Figure 1(b)) reaches low test error using roughly $n \simeq d$ samples, whereas online SGD fails to achieve even weak recovery with much larger sample size $n = \Omega(d^2)$. Our main contribution is to establish this improved statistical complexity for two-layer NNs trained by a variant of SGD with reused training data.

1.1 Our Contributions

We answer the above question in the affirmative by showing that SGD training (with the squared loss) on a natural class of shallow NNs can achieve small generalization error using polynomial compute and a sample complexity that is not governed by the information exponent, if we employ a layer-wise optimization procedure (analogous to that in [BES+22, DLS22, AAM23]) and reuse of the same minibatch. The core insight is that SGD can implement a full statistical query (SQ) algorithm that goes beyond CSQ, despite the correlational structure of the squared loss. Our main finding is summarized by the following theorem.
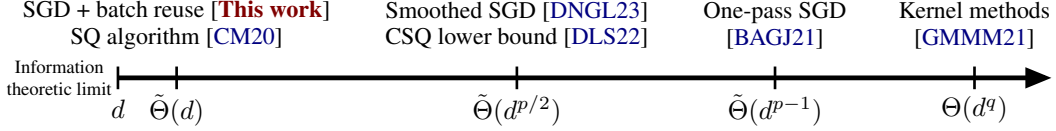
2

SGD + batch reuse [**This work**]
SQ algorithm [CM20]
Smoothed SGD [DNGL23]
CSQ lower bound [DLS22]
One-pass SGD
[BAGJ21]
Kernel methods
[GMMM21]

Information
theoretic limit

d $\quad \tilde{\Theta}(d)$ $\qquad\qquad\qquad \tilde{\Theta}(d^{p/2})$ $\qquad\qquad \tilde{\Theta}(d^{p-1})$ $\qquad\quad \Theta(d^q)$

Figure 2: Complexity of learning single-index model where the link function σ_* is a degree-q polynomial with information exponent p. For the CSQ lower bound, we translate the tolerance to sample complexity using the i.i.d. concentration heuristic $\tau \approx n^{-1/2}$. We restrict ourselves to algorithms using polynomial compute; this excludes the sphere-covering procedure in [DPVLB24] or exponential-width neural network in [Bac17, TS24].

Theorem (informal). *A shallow NN with $N = \tilde{\Theta}_d(1)$ neurons can learn arbitrary single-index models up to small population loss: $\mathbb{E}_{\boldsymbol{x}}[(f_{\boldsymbol{\Theta}}(\boldsymbol{x}) - f_*(\boldsymbol{x}))^2] = o_{d,\mathbb{P}}(1)$, if we employ an SGD-based algorithm (with reused training data) to minimize the squared loss objective, with a sample and runtime complexity of $n, T = \tilde{\Theta}_d(d^{(p_*-1)\vee 1})$, where p_* is the generative exponent of the link σ_*.*

Note that the generative exponent [DPVLB24] is defined as the *minimum* information exponent of the link function σ_* after arbitrary L^2 transformation, and hence by definition $p_* \leq p$ (equality is achieved by the identity transformation). We make the following remarks on our main result.

- We know that $p_* \leq 2$ for arbitrary *polynomial* link functions. Therefore, the theorem suggests that NN + SGD with reused batch can learn single-index polynomials with a sample complexity $n = \tilde{O}_d(d)$ which is information theoretically optimal up to polylogarithmic factors, hence matching the efficiency of SQ algorithms tailored for low-dimensional polynomial regression [CM20].

- For non-polynomial σ_* with high generative exponent $p_* > 2$, our sample complexity $n \gtrsim d^{p_*-1}$ can be interpreted as an SQ version of the online SGD result in [BAGJ21]. Since the information exponent p can be arbitrarily larger than the generative exponent p_*, our main theorem disproves a conjecture in [AAM23] stating that $n \asymp d^{p/2}$ is the optimal sample complexity for empirical risk minimization with SGD on the squared loss / correlation loss.

- A key observation in our analysis is that with suitable activation function, SGD with reused batch can go beyond correlational queries and implement (a subclass of) SQ algorithms. This enables polynomial transformations to the labels that reduce the information exponent, and therefore optimization can escape the high-entropy "equator" at initialization in polylogarithmic time.

Upon completion of this work, we became aware of the preprint [ADK$^+$24] showing weak recovery (for polynomial targets with $p_* \leq 2$) with similar sample complexity, also by exploiting the reuse of training data. Our work was conducted independently and simultaneously.

2 Problem Setting and Prior Works

Notations. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and the $\ell_2 \to \ell_2$ operator norm for matrices. $O_d(\cdot)$ and $o_d(\cdot)$ stand for the big-O and little-o notations, where the subscript highlights the asymptotic variable d and suppresses dependence on p, q; we write $\tilde{O}(\cdot)$ when (poly-)logarithmic factors are ignored. $\mathcal{O}_{d,\mathbb{P}}(\cdot)$ (resp. $o_{d,\mathbb{P}}(\cdot)$) represents big-O (resp. little-o) in probability as $d \to \infty$. $\Omega(\cdot), \Theta(\cdot)$ are defined analogously. γ is the standard Gaussian distribution in $\mathbb{R}$. We denote the L^2-norm of a function f with respect to the data distribution (which will be specified) as $\|f\|_{L^2}$. For $g : \mathbb{R} \to \mathbb{R}$, we denote g^i as its i-th exponentiation, and $g^{(i)}$ is the i-th derivative. We say an event happens *with high probability* when the failure probability is bounded by $\exp(-C \log d)$ for large constant C.

2.1 Complexity of Learning Single-index Models

We aim to learn a single-index model (1.1) where the link function $\sigma_* : \mathbb{R} \to \mathbb{R}$ has information exponent p defined as follows [DH18, BAGJ21].

Definition 1 (Information exponent). *Let $\{\mathsf{He}_j\}_{j=0}^\infty$ denote the normalized Hermite polynomials. The information exponent of $g \in L^2(\gamma)$, denoted by $\mathrm{IE}(g) := p \in \mathbb{N}_+$, is the index of the first non-zero Hermite coefficient of g, that is, given $g(z) = \sum_{i=0}^\infty \alpha_i \mathsf{He}_i(z)$, $p := \min\{i > 0 : \alpha_i \neq 0\}$.*

By definition, when σ_* is a degree-q polynomial, we always have $p \leq q$. Note that f_* contains $\Theta(d)$ parameters to be estimated, and hence *information theoretically $n \gtrsim d$ samples are both suf-*

ficient and necessary for learning [MM18, BKM$^+$19, DPVLB24]; however, the sample complexity achieved by different (polynomial time) algorithms depends on structure of the link function.

- **Kernel Methods.** Rotationally invariant kernels cannot adapt to the low-dimensional structure of single-index f_* and hence suffer from the curse of dimensionality [YS19, GMMM21, DWY21, BES$^+$22]. By a standard dimension argument [KMS20, HSSVG21, AAM22], we know that in the isotropic data setting, kernel methods (including neural networks in the lazy regime [JGH18, COB19]) require $n \gtrsim d^q$ samples to learn degree-q polynomials in $\mathbb{R}^d$.

- **Gradient-based Training of NNs.** While NNs can easily approximate a single-index model [Bac17], the sample complexity of gradient-based learning established in prior works typically scales as $n \gtrsim d^{\Theta(p)}$: in the well-specified setting, [BAGJ21] proved a sample complexity of $n = \tilde{\Theta}(d^{p-1})$ for online SGD, which is later improved to $\tilde{\Theta}(d^{p/2})$ by a smoothed objective [DNGL23]; as for the misspecified setting, [BBSS22, DKL$^+$23] showed that $n \gtrsim d^p$ samples suffice, and in some cases a $\tilde{\Theta}(d^{p-1})$ complexity is achievable [AAM23, OSSW24a]. Consequently, at the information-theoretic limit ($n \asymp d$), existing results can only cover the learning of low information exponent targets [AAM22, BMZ23, BES$^+$23]. This exponential dependence on p also appears in the CSQ lower bounds [DLS22, AAM22], which is often considered to be indicative of the performance of SGD learning with the squared loss (see Section 2.2).

Statistical Query Learners. If we do not restrict ourselves to correlational queries, the sample complexity of learning (1.1) can be drastically improved. Specifically, for polynomial σ_*, [CM20] gave an SQ algorithm that achieves low generalization error in $n = \tilde{O}(d)$ samples, which is near the information-theoretic limit; the key ingredient is to construct nonlinear transformations to the labels that lowers the information exponent to 2; similar preprocessing also appeared in context of phase retrieval [MM18, BKM$^+$19]. Such transformations do not belong to CSQ, but can be utilized by a full SQ learner to enhance the statistical efficiency. Recently, [DPVLB24] introduced the *generative exponent* which governs the complexity of SQ algorithms.

Definition 2 (Generative exponent). *The generative exponent (GE) of $g \in L^2(\gamma)$ is defined as the lowest information exponent (IE) after arbitrary L^2 transformation, that is,*

$$p_* =: \text{GE}(g) = \inf_{\mathcal{T} \in L^2(P_y)} \text{IE}(\mathcal{T} \circ g).$$

The generative exponent is the smallest information exponent obtained by all possible label transformations. By definition we always have $p^* \leq p$, and the gap between the two indices can be arbitrarily large; for example, for the Hermite polynomials we have $\text{IE}(\text{He}_k) = k$ whereas $\text{GE}(\text{He}_k) \leq 2$.

[DPVLB24] established a sample complexity lower bound of $n = \Omega(d^{p_*/2 \vee 1})$ for full SQ learners with polynomial compute (assuming $\tau \approx n^{-1/2}$), and obtained matching upper bound by a tensor partial-trace algorithm. Our goal is to show that SGD training of two-layer neural network can also achieve a sample and runtime complexity that scales with $n \simeq d^{\Theta(p_*)}$, where the dimension dependence is governed by the generative exponent p_* instead of the information exponent p.

2.2 Can Gradient Descent Go Beyond Correlational Queries?

Correlational statistical query. A statistical query (SQ) learner [Kea98, Rey20] accesses the target f_* through noisy queries $\tilde{\phi}$ with error tolerance τ: $|\tilde{\phi} - \mathbb{E}_{\boldsymbol{x},y}[\phi(\boldsymbol{x}, y)]| \leq \tau$. Lower bound on the performance of SQ algorithm is a classical measure of computational hardness. In the context of gradient-based optimization, an often-studied subclass of SQ is the *correlational* statistical query (CSQ) [BF02] where the query is restricted to (noisy version of) $\mathbb{E}_{\boldsymbol{x},y}[\phi(\boldsymbol{x})y]$. To see the connection between CSQ and SGD, consider the gradient of expected squared loss for one neuron $f_{\boldsymbol{w}}(\boldsymbol{x})$:

$$\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x},y}(f_{\boldsymbol{w}}(\boldsymbol{x}) - y)^2 \propto -\mathbb{E}_{\boldsymbol{x},y}[\underbrace{y \cdot \nabla_{\boldsymbol{w}} f_{\boldsymbol{w}}(\boldsymbol{x})}_{\text{correlational query}}] + \mathbb{E}_{\boldsymbol{x}}[\underbrace{f_{\boldsymbol{w}}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{w}} f_{\boldsymbol{w}}(\boldsymbol{x})}_{\text{can be evaluated without } y}].$$

One can see that information of the target function is encoded in the correlation term in the gradient. To infer the statistical efficiency of GD in the empirical risk minimization setting, we replace the population gradient with the empirical average $\nabla_{\boldsymbol{w}}(\frac{1}{n} \sum_{i=1}^n (f_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i)^2)$, and heuristically equate the CSQ tolerance τ with the scale of i.i.d. concentration error $n^{-1/2}$.

For the Gaussian single-index model class with information exponent p, [DLS22] proved a lower bound stating that a CSQ learner either has access to queries with tolerance $\tau \lesssim d^{-p/4}$, or exponentially many queries are needed to learn f_* with small population loss. Using the heuristic $\tau \approx n^{-1/2}$, this suggests a sample complexity lower bound $n \gtrsim d^{p/2}$ for polynomial time CSQ algorithm. This lower bound can be achieved by a landscape smoothing procedure [DNGL23] (in the well-specified setting), and is conjectured to be optimal for empirical risk minimization with SGD [AAM23].

SGD with reused data. As previously discussed, the gap between SQ and CSQ algorithms primarily stems from the existence of label transformations that decrease the information exponent. While such transformation cannot be utilized by a CSQ learner, [DTA$^+$24] argued that they may arise from two consecutive gradient updates using the same minibatch. For illustrative purposes, consider one neuron $f_{\boldsymbol{w}}(\boldsymbol{x}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$ updated by two GD steps using the same data point $(\boldsymbol{x}, y)$, starting from zero initialization $\boldsymbol{w}^0 = \boldsymbol{0}$ (we focus on the correlational term in the loss for simplicity):

$$\boldsymbol{w}^2 = \boldsymbol{w}^1 + \eta \cdot y\sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^1\rangle)\boldsymbol{x} = \eta\sigma'(0) \underbrace{y \cdot \boldsymbol{x}}_{\text{CSQ term}} + \eta \underbrace{y\sigma'(\eta\sigma'(0)\|\boldsymbol{x}\|^2 \cdot y)\boldsymbol{x}}_{\text{non-CSQ term}}. \tag{2.1}$$

Under appropriate learning rate scaling $\eta \cdot \|\boldsymbol{x}\|^2 = \Theta(1)$, one can see that in the second gradient step, the label y is transformed by the nonlinearity σ', even though the loss function itself is not modified. Based on this observation, [DTA$^+$24] showed that if the non-CSQ term in (2.1) reduces the information exponent to 1, then *weak recovery* (i.e., nontrivial overlap between the first-layer parameters $\boldsymbol{w}$ and index features $\boldsymbol{\theta}$) can be achieved after two GD steps with $n = \Theta(d)$ samples.

2.3 Challenges in Establishing Statistical Guarantees

Importantly, the analysis in [DTA$^+$24] does not lead to concrete learnability guarantees for the class of single-index polynomials for the following reasons: (i) it is not clear if an appropriate nonlinear transformation that lowers the information exponent can always be extracted from SGD with reused data, and (ii) the weak recovery guarantee may not translate to a sample complexity for the trained NN to achieve small generalization error. We elaborate these technical challenges below.

SGD decreases information exponent. To show weak recovery, [DTA$^+$24, Definition 3.1] assumed that the student activation σ can reduce the information exponent of the labels to 1; while a few examples are given, the existence of such transformations in SGD is not guaranteed:

- The label transformation employed in prior SQ algorithms [CM20] is based on thresholding, which reduces the information exponent to 2 for any polynomial σ_*; however, isolating such function from SGD updates on the squared loss is challenging. Instead, we make use of monomial transformation which can be extracted from SGD via Taylor expansion.

- If the link function satisfies $p_* \geq 2$, its information exponent after arbitrary nonlinear transformation is at least 2; such functions are predicted not be not learnable by SGD in the $n \asymp d$ regime [DTA$^+$24]. To handle this setting, we analyze the SGD update up to $\mathrm{poly}(d)$ time, at which a nontrivial overlap can be established by a Grönwall-type argument similar to [BAGJ21]. For $p_* = 2$, this recovers results on phase retrieval when $\sigma_*(z) = z^2$ which requires $n = \Omega(d\log d)$ samples.

From weak recovery to sample complexity. Note that weak recovery (i.e., $|\langle \boldsymbol{w}, \boldsymbol{\theta}\rangle| > \varepsilon$ for some small constant $\varepsilon > 0$) is generally insufficient to establish low generalization error of the trained NN. Therefore, we need to show that starting from a nontrivial overlap, subsequent gradient steps can achieve *strong recovery* of the index features (i.e., $|\langle \boldsymbol{w}, \boldsymbol{\theta}\rangle| > 1 - \varepsilon$), despite the link misspecification. After the first-layer parameters align with the target function, we train the second-layer parameters with SGD to learn the link function σ_* with the aid of random bias units [DLS22].

3 Learning Polynomial f_* in Linear Sample Complexity

We first consider the setting where σ_* is polynomial with degree q specified as follows.

Assumption 1. *The target function is given as $f_*(\boldsymbol{x}) = \sigma_*(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle)$, where the link function $\sigma_* : \mathbb{R} \to \mathbb{R}$ admits the Hermite decomposition $\sigma_*(z) = \sum_{i=p}^{q} \alpha_i \mathsf{He}_i(z)$.*

For single-index polynomials, we do not expect a computational-to-statistical gap under the SQ class [CM20] — indeed, we will establish learning guarantees near the information theoretic limit $n \asymp d$.

Algorithm 1: Gradient-based training of two-layer neural network
---
Input : Step sizes η^t; momentum parameters ξ^t; training time T_1, T_2; ℓ_2 regularization λ.
Initialize $\boldsymbol{w}_j^0 \sim \mathbb{S}^{d-1}(1)$, $a_j \sim \mathrm{Unif}\{\pm c_a\}$.
Phase I: normalized SGD on first-layer parameters
     for $t = 0$ ***to*** T_1 **do**
         if *t is even* **then**
             $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $y = f_*(\boldsymbol{x}) + \varsigma$;   `// Draw i.i.d. data` $(\boldsymbol{x}, y)$
             $\boldsymbol{w}_j^t \leftarrow \boldsymbol{w}_j^t - \xi_j^t(\boldsymbol{w}_j^t - \boldsymbol{w}_j^{t-2})$, (when $t > 0$) ;   `// Interpolation step`
             $\boldsymbol{w}_j^t \leftarrow \boldsymbol{w}_j^t / \|\boldsymbol{w}_j^t\|$;   `// Normalization`
         end
         $\boldsymbol{w}_j^{t+1} \leftarrow \boldsymbol{w}_j^t - \eta^t \tilde{\nabla}_{\boldsymbol{w}}(f_{\boldsymbol{\Theta}}(\boldsymbol{x}) - y)^2$, $(j = 1, \ldots, N)$;   `// SGD step`
     end
Initialize $b_j \sim \mathrm{Unif}([-C_b, C_b])$.
Phase II: SGD on second-layer parameters
     $\hat{\boldsymbol{a}} \leftarrow \mathrm{argmin}_{\boldsymbol{a} \in \mathbb{R}^N} \frac{1}{T_2} \sum_{i=1}^{T_2} (f_{\boldsymbol{\Theta}}(\boldsymbol{x}_i) - y_i)^2 + \lambda\|\boldsymbol{a}\|^2$;   `// Ridge regression`
Output: Prediction function $\boldsymbol{x} \mapsto f_{\hat{\boldsymbol{\Theta}}}(\boldsymbol{x})$ with $\hat{\boldsymbol{\Theta}} = (\hat{a}_j, \boldsymbol{w}_j^{T_1}, b_j)_{j=1}^N$.

3.1 Training Algorithm

We train the following two-layer network with N neurons using SGD to minimize the squared loss:

$$f_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \frac{1}{N} \sum_{j=1}^{N} a_j \sigma_j(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle + b_j), \tag{3.1}$$

where $\boldsymbol{\Theta} = (\boldsymbol{w}_j, a_j, b_j)_{j=1}^N$ are trainable parameters, and $\sigma_j : \mathbb{R} \to \mathbb{R}$ is the activation function defined as the sum of Hermite polynomials up to degree C_σ: $\sigma_j(z) := \sum_{i=0}^{C_\sigma} \beta_{j,i} \mathsf{He}_i(z)$, where C_σ only depends on the degree of link function σ_*. Note that we allow each neuron to have a different nonlinearity as indicated by the subscript in σ_j; this subscript is omitted when we focus on the dynamics of one single neuron. Our SGD training procedure is described in Algorithm 1, and below we outline the key ingredients of the algorithm.

- Algorithm 1 employs a layer-wise training strategy common in the recent feature learning theory literature [DLS22, BES$^+$22, BBSS22, AAM23, MHWSE23], where in the first stage, we optimize the first-layer parameters $\{\boldsymbol{w}_j\}_{j=1}^N$ with normalized SGD to learn the low-dimensional latent representation (index features $\boldsymbol{\theta}$), and in the second phase, we train the second-layer $\{a_j\}_{j=1}^N$ to fit the unknown link function σ_*.

- The most crucial part in Phase I of Algorithm 1 is the reuse of the same minibatch in the gradient computation. Specifically, we sample a fresh batch of training examples in *every two GD steps*; this enables us to extract non-CSQ terms from two consecutive gradient updates outlined in (2.1).

- We introduce an *interpolation step* between the current and previous iterates with hyperparameter ξ to stabilize the training dynamics; this resembles a negative momentum often seen in optimization algorithms [AZ18, ZLBH19]; the role of this interpolation is discussed in Section 4.2. We use a projected gradient update $\tilde{\nabla}_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) = (\boldsymbol{I}_d - \boldsymbol{w}^{2t} \boldsymbol{w}^{2t\top}) \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w})$ for steps $2t$ and $2t+1$, where $\nabla_{\boldsymbol{w}}$ is the Euclidean gradient; similar use of projection also appeared in [DNGL23, AAM23].

3.2 Convergence and Sample Complexity

Weak Recovery Guarantee. We first consider the "search phase" of SGD, and show that after running Phase I of Algorithm 1 for $T = \mathrm{polylog}(d)$ steps, a subset of parameters $\boldsymbol{w}$ achieve nontrivial overlap with the target direction $\boldsymbol{\theta}$. We denote $H(g; j)$ as the j-th Hermite coefficient of some $g \in L^2(\gamma)$. Our main theorems handle polynomial activations satisfying the following condition.

Assumption 2. *We require the activation function to be a polynomial $\sigma(z) = \sum_{i=0}^{C_\sigma} \beta_i \mathsf{He}_i(z)$ and its degree C_σ to be sufficiently large so that $C_\sigma \geq C_q$ holds (C_q is defined in Proposition 6). For all $2 \leq \ell \leq C_\sigma$ and $k = 0, 1$, we assume that $H(\sigma^{(\ell)}(\sigma^{(1)})^{\ell-1}; k) > 0$.*

As discussed in Appendix B.1, for a given σ_*, the above assumption only needs to be met for one pair of (k, ℓ). Appendix B.1.3 states that $H(\sigma^{(\ell)}(\sigma^{(1)})^{\ell-1}; k) \neq 0$ also suffices if we set the momentum parameter ξ differently. Now we verify this condition for a wide range of polynomial activations.

Lemma 3. *Given $\ell \geq 2$ and $k \geq 0$. For $C_\sigma \geq \frac{2\ell+k-1}{\ell}$, if we choose $\{\beta_i\}_{i=0}^{C_\sigma}$ where β_i is randomly drawn from some non-empty interval $[a_i, b_i]$, then $H(\sigma^{(\ell)}(\sigma^{(1)})^{\ell-1}; k) \neq 0$ with probability 1.*

The next theorem states that $n = \tilde{\Theta}(d)$ samples are sufficient for SGD to achieve weak recovery.

Theorem 1. *Under Assumptions 1 and 2, for suitable choices of hyperparameters $\eta^t = \tilde{O}_d(Nd^{-1})$ and $1 - \xi^t = o_d(1)$, there exists constant $C(q)$ such that after Phase I of Algorithm 1 is run for $2T_{1,1} = C(q) \cdot d\mathrm{polylog}(d)$ steps, with high probability, there exists a subset of neurons $\boldsymbol{w}_j^{2T_1} \in \mathcal{W}$ with $|\mathcal{W}| = \tilde{\Theta}(N)$ such that $\left|\langle \boldsymbol{w}_j^{2T_1}, \boldsymbol{\theta} \rangle\right| > c$ for some $c \gtrsim 1/\mathrm{polylog}(d)$.*

Recall that at random initialization we have $\langle \boldsymbol{w}, \boldsymbol{\theta} \rangle \approx d^{-1/2}$ with high probability. The theorem hence implies that SGD "escapes from mediocrity" after seeing $n = \tilde{O}(d)$ samples, analogous to the information exponent $p = 2$ setting studied in [BAGJ21]. We remark that due to the small second-layer initialization, the squared loss is dominated by the correlation loss, which allows us to track the evolution of each neuron independently; similar use of vanishing initialization also appeared in [BES+22, AAM23].

Strong recovery and sample complexity. After weak recovery is achieved, we continue Phase I to amplify the alignment. Due to the nontrivial overlap between $\boldsymbol{w}$ and $\boldsymbol{\theta}$, the objective is no longer dominated by the lowest degree in the Hermite expansion. Therefore, to establish strong recovery ($\langle \boldsymbol{w}, \boldsymbol{\theta} \rangle > 1 - \varepsilon$), we place an additional assumption on the activation function.

Assumption 3. *Given the Hermite expansions $\sigma_*(z) = \sum_{i=p}^q \alpha_i \mathsf{He}_i(z)$, $\sigma_j(z) = \sum_{i=0}^{C_\sigma} \beta_{j,i} \mathsf{He}_i(z)$, we assume the coefficients satisfy $\alpha_i \beta_{j,i} \geq 0$ for $p \leq i \leq q$.*

This assumption is easily verified in the well-specified setting $\sigma_* = \sigma$ [BAGJ21] since $\alpha_i = \beta_i$, and under link misspecification, it has been directly assumed in prior work [MHWSE23]. We follow [OSSW24a] and show that by randomizing the Hermite coefficients of the activation function, a subset of neurons satisfy the above assumption for any degree-q polynomial link function σ_*.

Lemma 4. *If we set $\sigma_j(z) = \sum_{i=0}^{C_\sigma} \beta_{j,i} \mathsf{He}_i(z)$, where for each neuron we sample $\beta_{j,i} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{\pm r_i\})$ with appropriate constant r_i, then Assumption 2 and 3 are satisfied in $\exp(-\Theta(q))$-fraction of neurons.*

Note that in our construction of activation functions for both assumptions, we do not exploit knowledge of the link function σ_* other than its degree q which decides the constant C_σ; see Appendix B.1 for more discussion of Assumption 3 and Lemma 4. The next theorem shows that by running Phase I for $\tilde{\Theta}(d)$ more steps, a subset of neurons achieves sufficiently large overlap with the index features.

Theorem 2. *For student neurons satisfying Assumptions 2, 3 and parameter $\boldsymbol{w}_j$ starting from non-trivial overlap $c > 0$ specified in Theorem 1, if Phase I of Algorithm 1 continues for $2T_{1,2} = \tilde{\Theta}_d(d\varepsilon^{-2})$ steps with hyperparameters $\eta^t = \tilde{O}_d(Nd^{-1}\varepsilon)$, $\xi^t = 1$, we achieve $\langle \boldsymbol{w}_j^{2(T_{1,1}+T_{1,2})}, \boldsymbol{\theta} \rangle > 1 - \varepsilon$ with high probability.*

The following proposition shows that after strong recovery, training the second-layer parameters in Phase II is sufficient for the NN model (3.1) to achieve small generalization error.

Proposition 5. *After Phase I terminates, for suitable $\lambda > 0$, the output of Phase II satisfies*
$$\mathbb{E}_{\boldsymbol{x}}[(f_{\hat{\boldsymbol{\Theta}}}(\boldsymbol{x}) - f_*(\boldsymbol{x}))^2] \lesssim \varepsilon^2.$$
with probability 1 as $d \to \infty$, if we set $T_2 = C(q)N^4\mathrm{polylog}(d)\varepsilon^{-4}$, $N = C(q)\mathrm{polylog}(d)\varepsilon^{-1}$ for some constant $C(q)$ depending on the target degree q.

Putting things together. Combining the above theorems, we conclude that in order for two-layer NN (3.1) trained by Algorithm 1 to achieve ε population squared loss, it is sufficient to set
$$n = T_1 + T_2 \asymp C(q) \cdot (d\varepsilon^{-2} \vee \varepsilon^{-8})\mathrm{polylog}(d), \quad N \asymp C(q) \cdot \varepsilon^{-1}\mathrm{polylog}(d),$$
where constant $C(q)$ only depends on the target degree q (although exponentially). Hence we may set $\varepsilon^{-1} \asymp \mathrm{polylog} d$ to conclude an almost-linear sample and computational complexity for learning arbitrary single-index polynomials up to $o_d(1)$ population error.

4 Proof Sketch

In this section we outline the high-level ideas and key steps in our derivation.

4.1 Monomial Transformation Reduces Information Exponent

To prove the main theorem, we first establish the existence of nonlinear label transformation that (i) reduces the information exponent, and (ii) can be easily extracted from SGD updates. If we ignore desideratum (ii), then for polynomial link functions, transformations that decrease the information exponent to at most 2 have been constructed in [CM20, Section 2.1]. However, prior results are based on the thresholding function, and it is not clear if such function naturally arises from SGD with batch reuse. The following proposition shows that the effect of thresholding can also be achieved by a simple monomial transformation where the required degree can be uniformly upper bounded.

Proposition 6. *Let $g : \mathbb{R} \to \mathbb{R}$ be any polynomial with degree up to p and $\|g\|_{L^2(\gamma)}^2 = 1$, then*

(i) *There exists some $i \leq C_q \in \mathbb{N}_+$ such that $\mathrm{IE}(g^i) \leq 2$, where constant C_q only depends on q.*

(ii) *Let $g^{\mathrm{odd}} : \mathbb{R} \to \mathbb{R}$ be the odd part of g with $\|g^{\mathrm{odd}}\|_{L^2(\gamma)}^2 \geq \rho > 0$. Then there exists some $i \leq C_{q,\rho} \in \mathbb{N}_+$ such that $\mathrm{IE}(g^i) = 1$, where constant $C_{q,\rho}$ only depends on q and ρ.*

The proof can be found in Appendix A. We make the following remarks.

- The proposition implies that for any polynomial link function that is not even, there exists some $i \in \mathbb{N}_+$ only depending on the degree of σ_* such that raising the function to the i-th power reduces the information exponent to 1 (this implies the generative exponent $p_* = 1$). For even σ_*, the information exponent after arbitrary transformation is at least 2 ($p_* = 2$), which can also be attained by monomial transformation. Furthermore, we provide a *uniform* upper-bound on the required degree of transformation i via a compactness argument.

- The advantage of working with monomial transformations is that they can be obtained from two GD steps on the same training example, by Taylor expanding the activation σ'. In Section 4.2, we build upon this observation to show that Phase I of Algorithm 1 achieves weak recovery using $n \gtrsim d \operatorname{polylog}(d)$ samples.

Intuition behind the analysis. Our proof is inspired by [CM20] which introduced a (non-polynomial) label transformation that reduces the information exponent of any degree-q polynomial to at most 2. To prove the existence of monomial transformation for the same purpose, we first show that for a fixed link function σ_*, there exists some i such that the i-th power of the link function has information exponent 2, which mirrors the transformation used in [CM20]. Then, we make use of the compactness of the space of link functions to define a test function and obtain a uniform bound on i. As for the polynomial transformation for non-even functions, we exploit the asymmetry of σ_* to further reduce the information exponent to 1.

4.2 SGD with Batch Reuse Implements Polynomial Transformation

Now we present a more formal discussion of (2.1) to illustrate how polynomial transformation can be utilized in batch reuse SGD. We let $\eta^t \equiv \eta$. When one neuron $f_w(x) = \sigma(\langle x, w \rangle)$ is updated by two GD steps using the same sample (x, y), starting from $w^0 := \omega$, the alignment with θ becomes

$$\langle \boldsymbol{\theta}, \boldsymbol{w}^2 \rangle = \langle \boldsymbol{\theta}, [\boldsymbol{w}^1 + \eta \cdot y\sigma'(\langle \boldsymbol{x}, \boldsymbol{w}^1 \rangle)\boldsymbol{x}] \rangle = \langle \boldsymbol{\theta}, \boldsymbol{\omega} \rangle +$$

$$\eta \left[y\sigma'(\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle)\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle + \sum_{i=0}^{C_\sigma - 1} \underbrace{(\eta\|\boldsymbol{x}\|^2)^i y^{i+1}(i!)^{-1}(\sigma'(\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle))^i \sigma^{(i+1)}(\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle)\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle}_{=: \psi_i} \right]. (4.1)$$

We take $\eta \leq c_\eta d^{-1}$ with a small constant c_η so that $\eta\|\boldsymbol{x}\|^2 \ll 1$ with high probability. Crucially, the strength of each term in (4.1) can vary depending on properties of the unknown link function σ_*. Hence a careful analysis is required to ensure that a suitable monomial transformation is always singled out from the gradient. We establish the following lemma on the evolution of alignment.

Lemma 7. *Under the assumptions per Theorem 1, the following holds for $p_* = 1, 2$:*

$$\langle \boldsymbol{\theta}, \boldsymbol{w}^{2(t+1)} \rangle \geq \langle \boldsymbol{\theta}, \boldsymbol{w}^{2t} \rangle + c_\eta^I c_\xi c_\sigma d^{-\frac{p_*}{2} \vee 1}(\kappa^{2t})^{p_* - 1} + c_\eta c_\xi d^{-\frac{p_*}{2} \vee 1}\nu^{2t}.$$

See Lemma 16 for the formal version. For $p_* = 1$, taking expectation immediately yields that weak recovery within $(\eta(1-\xi)\gamma)^{-1} = O(d)$ steps. For $p_* = 2$, $\langle \boldsymbol{\theta}, \boldsymbol{w}_j^{2t} \rangle =: \kappa^t$ can be approximated by a differential equation $\frac{\mathrm{d}\kappa^t}{\mathrm{d}t} = \eta(1-\xi)\gamma\kappa^t$. Solving this yields $\kappa^t = \kappa^0 \exp(\eta(1-\xi)\gamma t) \approx d^{-\frac{1}{2}}\exp(\eta(1-\xi)\gamma t)$, and weak recovery is obtained within $t \lesssim (\eta(1-\xi)\gamma)^{-1} \cdot \log d = O(d \log d)$ steps, similar to the analysis in [BAGJ21].

Why interpolation is needed.   In our setting, the signal strength may not dominate the error from discarding the effect of normalization. In prior analyses for online SGD, given the gradient $-\boldsymbol{g}$ and projection $P_{\boldsymbol{w}} = \boldsymbol{I}_d - \boldsymbol{w}\boldsymbol{w}^\top$, the spherical gradient changes the alignment as $\langle \boldsymbol{\theta}, \boldsymbol{w}^{t+1} \rangle = \left\langle \boldsymbol{\theta}, \frac{\boldsymbol{w}^t + \eta P_{\boldsymbol{w}}\boldsymbol{g}}{\|\boldsymbol{w}^t + \eta P_{\boldsymbol{w}}\boldsymbol{g}\|} \right\rangle \geq \langle \boldsymbol{\theta}, \boldsymbol{w}^t \rangle + \eta\langle \boldsymbol{\theta}, \boldsymbol{g} \rangle - \frac{1}{2}\eta^2\|\boldsymbol{g}\|^2\langle \boldsymbol{\theta}, \boldsymbol{w}^t \rangle + $ (h.o.t.), see [BAGJ21, DNGL23]. Here $\eta\langle \boldsymbol{\theta}, \boldsymbol{g} \rangle$ corresponds to the signal, and $-\frac{1}{2}\eta^2\|\boldsymbol{g}\|^2\langle \boldsymbol{\theta}, \boldsymbol{w}^t \rangle$ comes from normalization. Thus, taking η sufficiently small, the normalization error shrinks faster than the signal. However, in our case the signal shrinks at the rate of c_η^I (recall that $\eta = c_\eta d^{-1}$), and hence taking a smaller step may not improve the signal-to-noise ratio when the degree of transformation I is large. The interpolation step in Algorithm 1 reduces the effect of normalization without shrinking the signal too much. In particular, by setting $\xi = 1 - \tilde{\eta}$, we see that the signal is affected by a factor of $\tilde{\eta}$ whereas the normalization error shrinks by $\tilde{\eta}^2$; this allows us to boost the signal-to-noise ratio by taking $\tilde{\eta}$ small.

4.3   Analysis of Phase II and Statistical Guarantees

Once strong recovery is achieved for the first-layer parameters, we turn to Phase II and optimize the second-layer with ℓ_2 regularization. Since the objective is strongly convex, gradient-based optimization can efficiently minimize the empirical loss. In Appendix B.6, the learnability guarantee follows from standard analysis analogous to that in [AAM22, DLS22, BES+22], where we construct a "certificate" second-layer $\boldsymbol{a}^* \in \mathbb{R}^N$ that achieves small loss and small norm:

$$\mathbb{E}_{\boldsymbol{x}}\left(f_*(\boldsymbol{x}) - \frac{1}{N}\sum_{j=1}^N a_j^*\sigma_j\big(\langle \boldsymbol{w}_j^{T_1}, \boldsymbol{x} \rangle + b_j\big)\right)^2 \leq \varepsilon^*, \quad \|\boldsymbol{a}^*\| \lesssim r^*,$$

from which the population loss of the regularized empirical risk minimizer can be bounded via standard Rademacher complexity argument. To construct such a certificate, we make use of the random bias units $\{b_j\}_{j=1}^N$ to approximate the link function σ_* as done in [DLS22, BBSS22, OSSW24a].

5   Beyond Polynomial Link Functions

Thus far we have shown that for polynomial single-index target functions (which satisfy $p_* \leq 2$), SGD with data reuse can implement a polynomial transformation to the labels that reduces the information exponent to at most 2; consequently, the trained two-layer neural network can achieve small generalization error with $n = d\,\mathrm{polylog}(d)$ samples. However, as shown in [DPVLB24], there exists (non-polynomial) σ_* with generative exponent $p_* > 2$ (i.e., label transformations cannot lower the information exponent to 2) and thus not learnable by SQ algorithms in linear sample complexity.

Nevertheless, for a single-index model with generative exponent p_*, we know there exists an "optimal" label transformation that reduces the information exponent to p_*. If SGD can make use of such transformation, then from the arguments in [BAGJ21], it is natural to conjecture that a sample size of $n \simeq d^{p_*-1}$ is sufficient. In this section we confirm this intuition by proving that SGD with data reuse (Algorithm 1) indeed matches this complexity. The following lemma is an analogue of Proposition 6 stating that polynomial transformations are sufficient to lower the information exponent.

Lemma 8. *Given σ_* with generative exponent $p_* \in \mathbb{N}_+$. Suppose we can take an orthonormal polynomial basis $\{\phi_k\}_k$ for the space $L^2(P_y)$ with inner product $\langle f, g \rangle = \mathbb{E}_{y=\sigma_*(z)}[f(y)g(y)]$. Then there exists some degree of transformation $I \in \mathbb{N}_*$ such that $\mathrm{IE}(\sigma_*^I) = p_*$.*

We outline the differences and additional technical challenges to handle the $\mathrm{GE}(\sigma_*) > 2$ setting.

- For general L^2 link functions σ_*, we can no longer make use of the compactness argument (see proof of Proposition 6) to upper bound the degree of monomial transformation. Hence in Lemma 8 we do not state a uniform upper bound on the required degree I, unlike the polynomial setting.

- Any link function with $p_* > 2$ cannot be polynomial, and hence we cannot achieve low generalization error using a neural network with polynomial nonlinearity. We therefore need to use an activation function with universal function approximation ability.

5.1 Sample Complexity for Weak Recovery

We first show that Algorithm 1 achieves weak recovery (i.e., nontrivial overlap with the ground truth $\boldsymbol{\theta}$) with a complexity governed by the generative exponent of the link function $p_* = \mathrm{GE}(\sigma_*)$. Similar to Section 3.2, we make use of randomized activation functions to ensure the desired label transformation is encoded — we defer the conditions on the student activation to Appendix B.1.2.

Proposition 9. *Suppose the link function σ_* has generative exponent p_*, and let $I \in \mathbb{N}_+$ be the smallest degree of monomial transformation that lowers the information exponent to p_* (i.e., $\mathrm{IE}(\sigma_*^I) = p$). We can find a student activation function σ depending only on p, p_* and I, such that if we take $\eta^{2t}, \eta^{2t+1} = c_\eta N d^{-1}, \xi^{2(t+1)} = 1 - c_\xi d^{-(p_*-2)_+/2}$ for small $c_\eta, c_\xi = o_d(1)$, and set*

$$T_{1,1} \simeq c_\xi^{-1} \begin{cases} d & (\textit{if } p_* = 1) \\ d(\log d) & (\textit{if } p_* = 2) \\ d^{p_*-1} & (\textit{if } p_* \geq 3), \end{cases}$$

then if the initial alignment $\langle \boldsymbol{w}^0, \boldsymbol{\theta} \rangle \geq 2c_\eta^{-1} d^{-1/2}$, there exists $\tau_* \leq T_{1,1}$ such that for all $\tau \geq \tau_*$,

$$\langle \boldsymbol{w}^{2\tau}, \boldsymbol{\theta} \rangle \geq \tilde{\Theta}(1), \quad \textit{with probability } 1 - o_d(1).$$

Proposition 9 is a generalization of Theorem 1 beyond polynomial σ_* (the proof of both results are presented in Appendix B.3,B.4), and can be interpreted as an SQ counterpart to [BAGJ21]: we establish a sufficient sample size of $n \simeq d^{(p_*-1)\vee 1}$ for Algorithm 1 to exit the search phase, which is parallel to the $n \simeq d^{(p-1)\vee 1}$ rate for one-pass SGD (note that our rates are slightly sharper due to logarithmic factors removed, since c_ξ^{-1} can grow arbitrarily slowly with d). For high generative exponent σ_* with $p_* > 2$, we no longer match the information theoretically optimal sample complexity $n \asymp d$, which is consistent with the computational-to-statistical gap observed in [DPVLB24].

5.2 Generalization Error Guarantee

After Phase I of Algorithm 1, we learn the unknown link function σ_* via training the second-layer. To approximate non-polynomial functions, we introduce a ReLU component in the student nonlinearity σ (see Lemma 12 for discussions), and make use of the approximation result for the (univariate) ReLU kernel in [BBSS22], which handles general σ_* whose second derivative has bounded 4th moment. Combining the above, we arrive at the following end-to-end guarantee for learning single-index models with arbitrary generative exponent using SGD training of neural network.

Proposition 10 (Informal). *Suppose the link function σ_* has generative exponent $p_* \in \mathbb{N}_*$ and satisfies $\sigma_*, \sigma_*'' \in L^4(\gamma)$. For appropriately chosen activation function σ (see Appendix B.1.2), a neural network (3.1) with $N = \tilde{\Theta}(1)$ neurons optimized by Algorithm 1 achieves small population loss $\mathbb{E}_{\boldsymbol{x}}[(f_{\hat{\boldsymbol{\Theta}}}(\boldsymbol{x}) - f_*(\boldsymbol{x}))^2] = o_{d,\mathbb{P}}(1)$ with a sample complexity of $n = \tilde{\Theta}(d^{(p_*-1)\vee 1})$.*

See Appendix B.6 for the full statement with ε dependence. This proposition confirms that the sample complexity for weak recovery (Proposition 9) is the bottleneck in single-index learning, as the total sample size required for Algorithm 1 to achieve low test error also scales with $d^{(p_*-1)\vee 1}$.

6 Conclusion and Future Directions

We showed that a two-layer neural network (3.1) trained by SGD with reused batch can learn single-index model (with generative exponent p_*) using $n \simeq d^{(p_*-1)\vee 1}$ samples and compute; in particular, when the link function σ_* is polynomial, we established a sample complexity of $n = \tilde{O}(d\varepsilon^{-2})$ to achieve ε population loss, which is almost information theoretically optimal. Our analysis is based on the observation that by reusing the same training data twice in the gradient computation, a non-correlational term arises in the SGD update that transforms the labels (despite the loss function not modified). We proved that monomial transformations that lower the information exponent of σ_* can be extracted by Taylor-expanding the SGD update; then we showed via careful analysis of the trajectory that strong recovery and low population loss is achieved under suitable activation function.

Interesting future directions include extension to multi-index models [BAGJ22, BBPV23, CWPPS23], hierarchical target functions [AZL19, NDL23], and in-context learning [OSSW24b]. Also, the SGD algorithm that we employ requires a layer-wise training procedure and a specific batch reuse schedule; one may therefore ask if standard multi-pass SGD training of all parameters simultaneously [Gla23] (as reported in Figure 1) also achieves the same statistical efficiency.

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 *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[AAM23] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

[ADK+24] Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.

[AZ18] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.

[AZL19] Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *Advances in Neural Information Processing Systems*, 32, 2019.

[Bac17] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[BAGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

[BAGJ22] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

[BBPV23] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

[BBSS22] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 35:9768–9783, 2022.

[BES+22] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

[BES+23] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[BF02] Nader H Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2(Feb):359–395, 2002.

[BKM+19] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[BL20] Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*, 2020.

[BMZ23] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

[CCM11] Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein. Chernoff-type bounds for the Gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

[CM20] Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *Conference on Learning Theory*, pages 1161–1227. PMLR, 2020.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[CWPPS23] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *arXiv preprint arXiv:2308.08977*, 2023.

[DH18] Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pages 1887–1930. PMLR, 2018.

[DH24] Rishabh Dudeja and Daniel Hsu. Statistical-computational trade-offs in tensor pca and related problems via communication complexity. *The Annals of Statistics*, 52(1):131–156, 2024.

[DKL+23] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

[DLS22] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

[DNGL23] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[DPVLB24] Alex Damian, Loucas Pillaud-Vivien, Jason D Lee, and Joan Bruna. The computational complexity of learning gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

[DTA+24] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024.

[DWY21] Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

[Gla23] Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. *arXiv preprint arXiv:2309.15111*, 2023.

[GMMM21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

[HSSVG21] Daniel Hsu, Clayton H Sanford, Rocco Servedio, and Emmanouil Vasileios Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. In *Conference on Learning Theory*, pages 2423–2461. PMLR, 2021.

[JGH18]   Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[Kea98]   Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

[KMS20]   Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Conference on Learning Theory*, pages 2236–2262. PMLR, 2020.

[MHPG+23]   Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023.

[MHWSE23]   Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A. Erdogdu. Gradient-based feature learning under structured data. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[MM18]   Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450. PMLR, 2018.

[MZD+23]   Arvind Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. *Advances in Neural Information Processing Systems*, 36, 2023.

[NDL23]   Eshaan Nichani, Alex Damian, and Jason D Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. *Advances in Neural Information Processing Systems*, 36, 2023.

[O'D14]   Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[OSSW24a]   Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. In *Conference on Learning Theory*. PMLR, 2024.

[OSSW24b]   Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. *arXiv preprint arXiv:2411.02544*, 2024.

[Rey20]   Lev Reyzin. Statistical queries and statistical algorithms: Foundations and applications. *arXiv preprint arXiv:2004.00557*, 2020.

[Sch80]   Jacob T Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *Journal of the ACM (JACM)*, 27(4):701–717, 1980.

[TS24]   Shokichi Takakura and Taiji Suzuki. Mean-field analysis on two-layer neural networks from a kernel perspective. *arXiv preprint arXiv:2403.14917*, 2024.

[WWF24]   Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *Conference on Learning Theory*. PMLR, 2024.

[YS19]   Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[ZLBH19]   Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems*, 32, 2019.

Table of Contents

A   Polynomial Transformation

Proof of Proposition 6.   We use a thresholding and compactness argument inspired by [CM20].

A.1   Proof for Even Functions (i)

We divide the analysis into the following steps.

(i-1): Monomials reducing the information exponent. Define $\tau(f) = \max_{-2 \leq t \leq 2} |f(t)|$. This entails that if $|f(t)| \geq \tau(f)$, then we have $|t| > 2$.

Consider the following expectation:

$$\mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\left(\frac{f(t)}{2\tau(f)}\right)^i (t^2 - 1)\right]. \tag{A.1}$$

We evaluate the case when i is even. (A.1) can be lower bounded as

$$(A.1) = \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|f(t)| \geq 2\tau(f)]\left(\frac{f(t)}{2\tau(f)}\right)^i (t^2 - 1)\right]$$

$$+ \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[\tau(f) \leq |f(t)| < 2\tau(f)]\left(\frac{f(t)}{2\tau(f)}\right)^i (t^2 - 1)\right]$$

$$+ \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|f(t)| < \tau(f)]\left(\frac{f(t)}{2\tau(f)}\right)^i (t^2 - 1)\right]$$

$$\geq \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|f(t)| \geq 2\tau(f)]\left(\frac{2\tau(f)}{2\tau(f)}\right)^i (2^2 - 1)\right]$$

$$+ \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[\tau(f) \leq |f(t)| < 2\tau(f)]\left(\frac{f(t)}{2\tau(f)}\right)^i (2^2 - 1)\right]$$

$$+ \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|f(t)| < \tau(f)]\left(\frac{\tau(f)}{2\tau(f)}\right)^i (0^2 - 1)\right]$$

$$\geq 3\mathbb{P}_{t \sim \mathcal{N}(0,1)}[|f(t)| \geq 2\tau(f)] - 2^{-i}.$$

Note that $\mathbb{P}[|f(t)| \geq 2\tau(f)]$ is positive (since f is polynomial) and independent of i, while 2^{-i} decays to 0 as i increases. Therefore, for sufficiently large $i \in \mathbb{N}$, (A.1) is positive and hence $\mathrm{IE}(f^i) \leq 2$. The subsequent analysis aims to provide an upper bound on i.

(i-2): Construction of test function.   We introduce the notation $H(\cdot; j)$ which takes any function (in L^1) and returns its j-th Hermite coefficient. We consider the following test function:

$$\mathscr{H}(f) := \sum_{i=2}^{\infty} \left(\frac{H(f^i; 2)}{2^{\frac{i}{2}}(2i-1)^{\frac{iq}{2}}}\right)^2. \tag{A.2}$$

(i-3): Lower bound of test function via compactness.   Let $\mathcal{F}_q$ be a set of polynomials with degree up to q with unit L^2 norm. Because $\mathscr{H}(f)$ is positive for any $f \in \mathcal{F}_q$, $H(f^i; 2)$ is continuous with respect to f, and $\mathcal{F}_q$ is a compact set, $\inf_{f \in \mathcal{F}_q} \mathscr{H}(f)$ admits a minimum value $\mathscr{H}_0$ which is positive.

(i-4): Conclusion via hypercontractivity.   Because f is a polynomial with degree at most q, Gaussian hypercontractivity [O'D14] yields that

$$2H(f^i; 2)^2 \leq \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[(f(t))^{2i}\right] \leq (2i-1)^{iq}\left(\mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[f(t)^2\right]\right)^i = (2i-1)^{iq}.$$

Therefore, for all polynomials in $\mathcal{F}_q$, a partial sum of (A.2) is uniformly bounded by

$$\left|\sum_{i=j}^{\infty} \left(\frac{H(f^i; 2)}{2^{\frac{i}{2}}(2i-1)^{\frac{iq}{2}}}\right)^2\right| \leq \sum_{i=j}^{\infty} 2^{-i-1} = 2^{-j} \to 0 \quad (j \to \infty).$$

Combining this with the fact that $\mathscr{H}(f) \geq \mathscr{H}_0 > 0$, we know that there exists some $C_q \leq 1 + \log_2(\mathscr{H}_0^{-1})$ such that

$$\sum_{i=2}^{C_q} \left(\frac{H(f^i; 2)}{2^{\frac{i}{2}}(2i-1)^{\frac{iq}{2}}}\right)^2 > \frac{1}{2}\mathscr{H}_0 > 0,$$

for all polynomials in $\mathcal{F}_q$. This means that there is at least one $i \leq C_q$ such that $H(f^i; 2) \neq 0$.

A.2 Proof for Non-even Functions (ii)

(ii-1): Monomials reducing the information exponent. We prove that some exponentiation of $g := f^2$ has non-zero first Hermite coefficient. Denote g^{odd} as the odd part of g, and similarly g^{even}. Let $\upsilon(g) \in \mathbb{R}_+$ be the value at which the followings hold:

(a) $g^{\mathrm{odd}}(t) > 0$ for all $t \geq \upsilon(g)$ and $g^{\mathrm{odd}}(t) < 0$ for all $t \leq -\upsilon(g)$.

(b) $g^{\mathrm{even}}(t) > |g^{\mathrm{odd}}(t)|$ for all $t \geq \upsilon(g)$ and $t \leq -\upsilon(g)$.

(c) For for all $t \geq \upsilon(g)$ and $t \leq -\upsilon(g)$, $g(s) = g(t)$ (as an equation of s) only has two real-valued solutions with opposing signs.

Such threshold $\upsilon(g)$ exists because the tail of $g = f^2$ is dominated by the highest degree which is even. Then, we let $\tau(g) = \max_{-\upsilon(g) \leq t \leq \upsilon(g)} |g(t)|$.

Consider the following expectation:

$$\mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\left(\frac{g(t)}{2\tau(g)}\right)^i t\right]. \tag{A.3}$$

(A.3) is decomposed as

$$(\text{A.3}) = \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|g(t)| \geq 3\tau(f)]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right]$$

$$+ \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[2\tau(g) \leq |g(t)| < 3\tau(g)]\left(\frac{g(t)}{3\tau(f)}\right)^i t\right]$$

$$+ \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|g(t)| < 2\tau(g)]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right]. \tag{A.4}$$

We first evaluate the first term. Because of (c), $g(t) = 3\tau(f)$ has two real-valued solutions $\alpha < 0 < \beta$. Because of (a) and (b), $g(\beta) = g^{\mathrm{even}}(\beta) + g^{\mathrm{odd}}(\beta) = 3\tau(f) > g^{\mathrm{even}}(-\beta) + g^{\mathrm{odd}}(-\beta) = g^{\mathrm{odd}}(-\beta)$. Because $\lim_{t \to -\infty} g^{\mathrm{odd}}(t) = +\infty$, and α is the only solution in $t < 0$, we have $\alpha < -\beta$. Moreover, for all $t > \beta$, we have $g(t) = g^{\mathrm{even}}(t) + g^{\mathrm{odd}}(t) > g^{\mathrm{even}}(-t) + g^{\mathrm{odd}}(-t) = g^{\mathrm{odd}}(-t)$. Combining the above, the first term of (A.4) is bounded as

$$\mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|g(t)| \geq 3\tau(f)]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right]$$

$$= \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[\beta \leq t \leq -\alpha]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right] + \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[t \geq -\alpha]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right]$$

$$+ \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[t \leq \alpha]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right]$$

$$= \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[\beta \leq t \leq -\alpha]t\right] + \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[t \geq -\alpha]\left(\left(\frac{g(t)}{3\tau(g)}\right)^i - \left(\frac{g(-t)}{3\tau(g)}\right)^i\right)t\right]$$

$$> \beta \mathbb{P}_{t \sim \mathcal{N}(0,1)}\left[\beta \leq t \leq -\alpha\right].$$

Following the exact same reasoning, we know that the second term of (A.4) is positive. Finally, the third term which is bounded by

$$\mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|g(t)| < 2\tau(g)]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right] \geq -\mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|g(t)| < 2\tau(g)]|t|\right]\left(\frac{2}{3}\right)^i.$$

Putting things together,

$$(\text{A.4}) > \beta \mathbb{P}_{t \sim \mathcal{N}(0,1)}\left[\beta \leq t \leq -\alpha\right] - \mathbb{E}_{t \sim \mathcal{N}(0,1)}\left[\mathbb{1}[|g(t)| < 2\tau(g)]|t|\right]\left(\frac{2}{3}\right)^i.$$

The first term is independent of i and positive, while the second term goes to zero as i grows. Therefore, there exists some i such that $\mathrm{IE}(g^i; 1) = 1$.

(ii-2): Construction of test function. This time we consider the following function:

$$\mathscr{H}(f) := \sum_{i=2}^{\infty} \left(\frac{H(f^i; 1)}{2^{\frac{i}{2}} (2i-1)^{\frac{iq}{2}}} \right)^2.$$

(ii-3): Lower bound of test function via compactness. Let $\mathcal{F}_q$ be a set of unit L^2-norm polynomials with degree up to q and $\mathbb{E}_{t \sim \mathcal{N}(0,1)}[f^{\mathrm{odd}}(t)^2] \geq c$. Since $\mathscr{H}(f)$ is always positive for $\mathcal{F}_q$, $\mathscr{H}(f)$ is continuous with respect to f, and $\mathcal{F}_q$ is a compact set, $\inf_{f \in \mathcal{F}_q} \mathscr{H}(f)$ has the minimum value $\mathscr{H}_0$ that is positive. Note that $\mathscr{H}(f)$ might depends on c.

(ii-4): Conclusion via hypercontractivity. Using the same argument as in (i), we conclude that there exists some $C_{q,c}$ such that

$$\sum_{i=2}^{C_q} \left(\frac{H(f^i; 1)}{2^i (2i-1)^{\frac{iq}{2}}} \right)^2 > \frac{1}{2} \mathscr{H}_0 > 0.$$

Because $\mathscr{H}_0$ depends on c, $C_{q,c}$ depends on c as well as q. □

A.3 Proof for Non-Polynomial Functions

For non-polynomial link functions, we note that similar to [DPVLB24], the existence of polynomial basis is needed to exclude extreme cases, and we cannot upper bound the required degree I because general link functions are not included in a compact space.

Proof of Lemma 8. The derivation is analogous to [DPVLB24, Lemma F.14]. Let $z \sim \mathcal{N}(0,1)$ and $y = \sigma_*(z)$. We define $\zeta_{p_*}(y) = \mathbb{E}[\frac{1}{\sqrt{p_*!}} \mathsf{He}_{p_*}(z)|y]$ and its basis expansion $\zeta_{p_*}(y) = \sum_{k=0}^{\infty} v_k \phi_k$. Let K be a smallest integer such that $v_k \neq 0$. Then, there exists an integer with $I \leq K$ such that $\mathrm{IE}(y^I) = p_*$. Indeed,

$$\mathbb{E}[\phi_K(y) \mathsf{He}_{p_*}(z)] = \mathbb{E}_y[\Phi_K(y) \mathbb{E}_{z|y}[\mathsf{He}_{p_*}(z)|y]]$$

$$= \mathbb{E}_y \left[\Phi_K(y) \sum_{k=0}^{K} v_k \phi_k(y) \right] = v_K \neq 0,$$

which means that at least one of $y, y^2, \cdots, y^K$ yields a non-zero p_*-th Hermite coefficient. □

B SGD with Reused Batch

In this section we show that Algorithm 1 learns single-index models in $\tilde{O}(d^{1 \vee (p_*-1)})$ samples with high probability. The algorithm trains the first layer for T_1 SGD steps, where we sample a new data point in every two steps. The first layer training is further divided into two phases: weak recovery ($\boldsymbol{w}^\top \boldsymbol{\theta} \gtrsim 1$) and strong recovery ($\|\boldsymbol{w} - \boldsymbol{\theta}\| \lesssim \varepsilon$). Then, we learn the second layer parameters.

Specifically, Section B.2 shows that at initialization, a (nearly) constant fraction of neurons has alignment $\boldsymbol{w}^\top \boldsymbol{\theta}$ beyond a certain threshold. We focus on such neurons in the first phase of training. Section B.3 lower bounds the expected update of alignment $\boldsymbol{w}^\top \boldsymbol{\theta}$ of two gradient steps, and Section B.4 establishes that the neurons achieve weak recovery within $2T_{1,1} = \tilde{O}(d^{1 \vee (p_*-1)})$ steps. Section B.5 discusses how to convert weak recovery to strong recovery using $2T_{1,2} = \tilde{O}(d\varepsilon^{-2})$ more steps. We let $T_1 = 2T_{1,1} + 2T_{1,2}$. Finally, Section B.6 analyzes second layer training and concludes the proof.

In the following proofs, we use several constants, which depends on d at most at most polylogarithmically. Specifically, asymptotic strength of the constants is ordered as follows.

$$1 \simeq c_\sigma \simeq C_1 \lesssim \begin{Bmatrix} c_\eta^{-1} \simeq C_2 \lesssim \mathrm{poly}(c_\eta^{-1}) \lesssim \begin{Bmatrix} c_1^{-1} \simeq C_3 \\ c_2^{-1} \end{Bmatrix} \\ \delta^{-1} \end{Bmatrix} \lesssim \begin{Bmatrix} \delta^{-1}\mathrm{poly}(c_\eta^{-1}) \lesssim c_\xi^{-1} \\ \mathrm{poly}(c_1^{-1}) \lesssim \bar{c}_\eta^{-1} \end{Bmatrix}$$

$$\lesssim \mathrm{polylog}(d) = C_4.$$

Here, c_η and δ should satisfy $\lim_{d\to\infty} c_\eta = \lim_{d\to\infty}\delta = 0$, but the convergence can be arbitrarily slow, (e.g., as slow as $1/\log\log\log\cdots\log d$). This requirement comes from the fact that we do not know the exact value of $H(\sigma_*^I; p_*)$. To ensure that one signal term (from the Taylor series) is isolated, taking $\eta \asymp d^{-1}$ with a sufficiently small constant is insufficient but $\eta \asymp c_\eta d^{-1}$ with arbitrarily slow c_η suffices. Also, to guarantee that the failure probability is $o_d(1)$, we require δ to be $o_d(1)$. c_ξ can also decay arbitrarily slowly, as long as it satisfies $c_\xi \lesssim \delta\,\mathrm{poly}(c_\eta^{-1})$. $C_4 = \mathrm{polylog}(d)$ will be used to represent any polylogarithmic factor that comes from high probability bounds.

For the first-layer training, we can reduce the argument into training of one neuron using the correlation loss as follows. At each step, the gradient update (Line 8 of Algorithm 1) is written as

$$
\begin{aligned}
\boldsymbol{w}_j^{t+1} &\leftarrow \boldsymbol{w}_j^t - \eta^t \tilde{\nabla}_{\boldsymbol{w}}\big((f_\Theta(\boldsymbol{x}) - y)^2\big) \\
&= \boldsymbol{w}_j^t - \eta^t \tilde{\nabla}_{\boldsymbol{w}}\bigg(\frac{1}{N}\sum_{j=1}^N a_j \sigma_j(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\bigg)^2 + 2\eta_j^t \tilde{\nabla}_{\boldsymbol{w}}\bigg(y\frac{1}{N}\sum_{j=1}^N a_j\sigma_j(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\bigg) \\
&= \boldsymbol{w}_j^t - \frac{2\eta^t c_a^2}{N}\bigg(\frac{1}{N}\sum_{j=1}^N \sigma_j(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\bigg)\big(\tilde{\nabla}_{\boldsymbol{w}}\sigma_j(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\big) + \frac{2\eta^t c_a}{N}y\big(\tilde{\nabla}_{\boldsymbol{w}}\sigma_j(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\big). \quad\text{(B.1)}
\end{aligned}
$$

While the second term scales with $\eta^t c_a^2 N^{-1}$, the third term scales with $\eta^t c_a N^{-1}$. Thus, by setting c_a sufficiently small, we can ignore the interaction between neurons. We will show that the strength of the signal in the direction of $\boldsymbol{\theta}$ is at least $(\kappa_j^t)^{p_*-1} \gtrsim d^{-\frac{p_*-1}{2}}$ (up to a polylogarithmic factor, and $p_* = \mathrm{GE}(\sigma_*)$). On the other hand, we can easily see that $\boldsymbol{\theta}^\top\big(\frac{1}{N}\sum_{j=1}^N \sigma_j(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\big)\big(\tilde{\nabla}_{\boldsymbol{w}}\sigma_j(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\big)$ is bounded by $\tilde{O}(1)$ with high probability. Therefore, by simply letting $c_a = \tilde{\Theta}(d^{-\frac{p_*-1}{2}})$, we can ignore the effect of the second term in (B.1). Moreover, for simplicity, we will reparameterize $\frac{2\eta^t c_a}{N}$ as η^t below. Consequently, we may analyze the following update

$$
\boldsymbol{w}_j^{t+1} \leftarrow \boldsymbol{w}_j^t + \eta^t \tilde{\nabla}_{\boldsymbol{w}}\big(y\sigma_j(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\big),
$$

instead of Line 8 of Algorithm 1. Since there is no interaction between neurons now, we omit the subscript j when the context is clear.

B.1 Assumptions on Link Function

The analysis consists of three different phases: weak recovery and strong recovery of the first-layer weights, and approximation of the link function (ridge regression of the second-layer). Each phase requires different assumptions on the activation functions, depending on the link function. Before starting the analysis, we decompose Assumptions 2 and 3 and clarify which conditions are needed in each phase. We prove that instead of using a specific activation function tailored to different link functions, a randomized activation function satisfies all required assumptions with probability $\Omega(1)$.

In the following, we write the student activation function as

$$
\sigma_j(s) := \sum_{i=0}^\infty \beta_{j,i}\mathsf{He}_i(s)
$$

with coefficients $\{\beta_{j,i}\}_{i=0}^{C_\sigma}$ (sometimes the subscript j, which is the index of the neurons, is omitted).

B.1.1 For polynomial link functions

In the following, we summarize the precise conditions to be satisfied by the activation functions (these conditions are weaker than Assumptions 2 and 3). For polynomial link functions, we focus on polynomial activation functions (with bounded degree) for simplicity, but non-polynomial activation functions would not change the proof significantly.

Let p and q be the minimum and maximum degree of non-zero Hermite coefficients of σ_*. Note that $\mathrm{GE}(\sigma_*) = 1$ or 2 holds (see Proposition 6). Let $I \leq C_q$ (according to Proposition 6) be the smallest integer such that $\mathrm{IE}(\sigma_*^I) = \mathrm{GE}(\sigma_*) = p_*$ and C_σ be the degree of the activation function.

(I) If $I = 1 \Leftrightarrow \mathrm{IE}(\sigma_*) = \mathrm{GE}(\sigma_*) = p_*$.

    Weak recovery: $\alpha_{p_*}\beta_{p_*} > 0$ (covered by Assumption 3).

Strong recovery: $\sum_{j=p_*}^{q} j! \alpha_j \beta_j s^{j-1} > 0$ for all $s > 0$ (covered by Assumption 2).

Approximation (ridge regression): $\beta_i \neq 0$ for some $i \geq q$ (covered by Assumption 3).

(II) If $2 \leq I = \{\min i \mid \mathrm{IE}(\sigma_*^I) = \mathrm{GE}(\sigma_*) = p_*\} \leq C_\sigma$.

Weak recovery: $H((\sigma_*)^I; p_*) H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_* - 1) > 0$ (covered by Assumption 3).

Strong recovery: $\sum_{j=p_*+1}^{q} j! \alpha_j \beta_j s^{j-1} > 0$ for all $s > 0$ (covered by Assumption 2).

Approximation: $\beta_i \neq 0$ for some $i \geq q$ (covered by Assumption 3).

Note that it is difficult to construct a deterministic activation function that satisfies all of the assumptions for any link function σ_* (the simplest counterexample is to consider $-\sigma_*$ which flips the Hermite coefficients). Instead, we show the existence of randomized construction of such an activation function that satisfies all of the assumptions on the activation function simultaneously with constant probability, which entails that a subset of neurons can achieve strong recovery. The construction does not depend on properties of the link function itself except for its degree q.

Lemma 11. *There exists a randomized activation function sampled from a discrete set such that the above conditions hold with constant probability.*

Proof. Let c be a sufficiently small constant only used in this proof and C_σ be the minimum odd integer with $C_\sigma \geq \max\{C_q + 1, q + 2, 3\}$, where C_q was introduced in Proposition 6. With probability $\frac{1}{2}$, we let $\beta_1 \sim \mathrm{Unif}(\{\pm 1\})$, and $\beta_j \sim \mathrm{Unif}(\{\pm c\})$ for $2 \leq j \leq C_\sigma$. With probability $\frac{1}{2}$, we let $\beta_j \sim \mathrm{Unif}(\{\pm c\})$ for $1 \leq j \leq C_\sigma - 2$ and $\beta_{C_\sigma - 1} = \beta_{C_\sigma} \sim \mathrm{Unif}(\{\pm 1\})$.

We first consider (I). When $\beta_1 \sim \mathrm{Unif}(\{\pm 1\})$, and $\beta_j \sim \mathrm{Unif}(\{\pm c\})$ for $2 \leq j \leq C_\sigma$, it is easy to see $\mathrm{sign}(\alpha_j) = \mathrm{sign}(\beta_j)$ for all $j = 1, \cdots, q$ hold with probability at least 2^{-q}, which is sufficient to satisfy (I).

We then consider (II). First focus on the case when $p_* = 1$ and I is even. When $\beta_1 \sim \mathrm{Unif}(\{\pm 1\})$ and $\beta_j \sim \mathrm{Unif}(\{\pm c\})$ for $2 \leq j \leq C_\sigma$, by taking c sufficiently small, we have

$$H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 0) = \underbrace{I! \beta_I (\beta_1)^{I-1}}_{\asymp c} + O(c^2). \tag{B.2}$$

When I is even, by adjusting the sign of β_1, $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 0)$ is non-zero and has the same sign as $H((\sigma_*)^I; 1)$ with probability $\frac{1}{2}$. Note that the sign of β_1 is independent from whether $\sum_{j=2}^{q} j! \alpha_j \beta_j s^{j-1} > 0$ for all $s > 0$ holds. This holds with probability at least 2^{-q+1}. Thus we verified (II) for $p_* = 1$ and even I.

For $p_* = 1$ and odd I, consider $\beta_j \sim \mathrm{Unif}(\{\pm c\})$ for $1 \leq j \leq C_\sigma - 2$ and $\beta_{C_\sigma - 1} = \beta_{C_\sigma} \sim \mathrm{Unif}(\{\pm 1\})$. Note that $\sum_{j=2}^{q} j! \alpha_j \beta_j s^{j-1} > 0$ for all $s > 0$ (this is the condition for strong recovery) and the condition for ridge regression also holds. Furthermore, the sign of $H((\mathrm{He}_{C_\sigma} + \mathrm{He}_{C_\sigma - 1})^{(I)}((\mathrm{He}_{C_\sigma} + \mathrm{He}_{C_\sigma - 1})^{(1)})^{I-1}; 0)$ is ± 1 with equiprobability, independent of $\beta_2, \ldots, \beta_q$. Therefore, by taking c sufficiently small, we can obtain the desired sign of $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 0)$. Thus we proved (II) for $p_* = 1$ and odd I.

Regarding (II) for $p_* = 2$ and even I, when $\beta_1 \sim \mathrm{Unif}(\{\pm 1\})$ and $\beta_j \sim \mathrm{Unif}(\{\pm c\})$ for $2 \leq j \leq C_\sigma$, we have

$$H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 1) = \underbrace{(I+1)! \beta_{I+1} (\beta_1)^{I-1}}_{\asymp c} + O(c^2).$$

Thus, similar to (II) with $p_* = 1$ and even I, we get (II) for $p_* = 2$ and even I.

Finally, consider (II) for $p_* = 2$ and odd I. When $\beta_j \sim \mathrm{Unif}(\{\pm c\})$ for $1 \leq j \leq C_\sigma - 2$ and $\beta_{C_\sigma - 1} = \beta_{C_\sigma} \sim \mathrm{Unif}(\{\pm 1\})$, the sign of $H((\mathrm{He}_{C_\sigma} + \mathrm{He}_{C_\sigma - 1})^{(I)}((\mathrm{He}_{C_\sigma} + \mathrm{He}_{C_\sigma - 1})^{(1)})^{I-1}; 1)$ is ± 1 with equiprobability when I is odd, and this term dominates the others in $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 1)$. Thus, (II) for $p_* = 2$ and odd I holds similarly to (II) for $p_* = 1$ and odd I.

Now we have obtained the assertion for all cases. $\qquad\square$

B.1.2 For general link functions

Now we consider non-polynomial link functions with potentially large generative exponent $p_* = \mathrm{GE}(\sigma_*) \geq 2$. For weak and strong recovery to succeed, the conditions on the activation function are essentially the same as those for polynomial link functions:

(I) If $I = 1 \Leftrightarrow \mathrm{IE}(\sigma_*) = \mathrm{GE}(\sigma_*) = p_*$**.**

    Weak recovery: $\alpha_{p_*}\beta_{p_*} > 0$.

    Strong recovery: $\sum_{j=p_*}^{\infty} j!\alpha_j\beta_j s^{j-1} > 0$ for all $s > 0$,

(II) If $2 \leq I = \{\min\ i\ |\ \mathrm{IE}(\sigma_*^I) = \mathrm{GE}(\sigma_*) = p_*\} \leq C_\sigma$**.**

    Weak recovery: $H((\sigma_*)^I; p_*)H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_* - 1) > 0$,

    Strong recovery: $\sum_{j=p_*+1}^{\infty} j!\alpha_j\beta_j s^{j-1} > 0$ for all $s > 0$.

Due to the proof strategy (which uses Taylor expansion), we also require that all differentials and sum of expectations appearing in the following proofs are well-defined and bounded.

To approximate a non-polynomial σ_*, we introduce the following condition on the activation function. We sample σ_j from a discrete set (with bounded cardinality). Let J be an index set such that the coefficients of σ_j ($j \in J$) satisfy the conditions above. Because we are selecting σ_j from a discrete set, $|J| \simeq N$ holds. We introduce the following condition, which states that the target single-index model can be well-approximated by a linear combination of student neurons.

Assumption 4. *When* $b_j \sim \mathrm{Unif}([-C_b, C_b])$ *where* $(C_b = \mathrm{polylog}(d))$ *and* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T_2} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, *there exists a set of coefficients* $a_1, \ldots, a_{|J|}$ *such that*

$$\frac{1}{T_2}\sum_{i=1}^{T_2}\left(\frac{1}{|J|}\sum_{j\in J} a_j\sigma_j(\boldsymbol{\theta}_j^\top \boldsymbol{x}_i + b_j) - \sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}_i)\right)^2 \lesssim \varepsilon^2,$$

holds with coefficients of reasonable magnitudes $\sum_{j\in J} a_j^2 = \Theta(|J|)$ *with high probability (w.r.t. the randomness of* b_j *and* $\boldsymbol{x}_i$*). Moreover,* $\mathbb{E}_{\boldsymbol{x}}[\sigma_j(\boldsymbol{\theta}_j^\top \boldsymbol{x} + b_j)^4] \leq \mathrm{polylog}(d)$ *for all* j *with high probability (w.r.t. the randomness of* b_j*).*

The following lemma states that we can design a randomized activation function that satisfies all of the above assumptions with probability $\Omega(1)$, as long as the link function σ satisfies Assumption 4 for $\sigma = \mathrm{ReLU}$. In other words, we are able to cover the class of link functions σ that can be efficiently approximated by a two-layer ReLU network. Since the general link functions are not included in a compact space, we do not have an upper bound of exponent to obtain $\mathrm{IE}(\sigma_*^I) = \mathrm{GE}(\sigma_*)$ as we had C_q in the polynomial case. Consequently, our student activation is not entirely agnostic to the link function σ_*, as we require knowledge of p (information exponent), p_* and I.

Lemma 12. *Suppose the target link function* σ_* *satisfies Assumption 4 for* $\sigma_j = \mathrm{ReLU}$. *There exists a randomized activation sampled from a discrete set such that the above conditions hold with constant probability.*

Before we sketch the design of activation function, we present the following approximation result from [BBSS22], which establishes that Assumption 4 with $\sigma_j = \mathrm{ReLU}$ is satisfied for broad class of functions, according to Lemma 4.4 and 4.5 of [BBSS22]. Specifically, taking $\tau = 1/2$ and $\lambda = N^{-1}$ yields that $\mathbb{E}_{\boldsymbol{x}}[(\frac{1}{|J|}\sum_{j\in J} a_j\sigma_j(\boldsymbol{\theta}^\top \boldsymbol{x} + b_j) - \sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^2] \leq N^{-\frac{2}{7}}$. Although they sample b_j from Gaussian $\mathcal{N}(0, 2)$, the result translates to uniform sampling of biases from $[-C_b, C_b]$ by introducing additional logarithmic factor.

Lemma 13 (Lemma 4.4, 4.5 of [BBSS22]). *Suppose that* $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma_*(z)^4]$, $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma_*''(z)^4] < \infty$. *Then, Assumption 4 with* $\sigma_j = \mathrm{ReLU}$ *holds with* $\varepsilon = N^{-\frac{1}{7}}$ *and* $C_b \simeq \sqrt{\log d}$.

Proof of Lemma 12. We show the existence of suitable σ in two steps: first we construct a randomized polynomial activation function that satisfies conditions (I)(II) with constant probability; then we add a small ReLU perturbation so that the activation can approximate non-polynomial σ_*.

Recall $p \in \mathbb{N}_+$ is the information exponent of σ_*. We first show that there exists a randomized polynomial activation that satisfies the conditions for weak and strong recovery with probability $\Omega(1)$. Note that the issue of differentiability and bounded moment is avoided when we focus on the polynomial activation functions. We specify the following two distributions. With probability $\frac{1}{2}$, let $\beta_1 \sim \mathrm{Unif}(\{-1, 1\})$, $\beta_j \sim \mathrm{Unif}(\{-c, c\})$ for $j = 1, \cdots, p_* + I - 1$ and $\beta_j = 0$ otherwise, where $c > 0$ is a sufficiently small constant. With probability $\frac{1}{2}$, let $\beta_1 = \mathrm{Unif}(\{-1, 1\})$, $\beta_2 = \mathrm{Unif}(\{-c, c\})$, $\beta_j = \mathrm{Unif}(\{-c^2, c^2\})$ for all $2 \leq j \leq (p_* + I) \vee p$ for a sufficiently small constant $c > 0$, and $\beta_j = 0$ otherwise.

Regarding (I), consider the case when the coefficients are sampled from the first distribution, and $|\beta_j| \ll 1$ except for $j = p_*$. Then, $\sum_{j=p_*}^{\infty} j! \alpha_j \beta_j s^{j-1} \approx p_*! \alpha_{p_*} \beta_{p_*} s^{p_*}$. Choosing the sign of β_{p_*}, we have that the assumption holds with probability $\Omega(1)$.

Regarding (II) with even I, consider coefficients sampled from the first distribution, and $\mathrm{Sign}(\beta_j) = \mathrm{Sign}(\alpha_j)$ for $j \leq (p_* + I - 1) \vee p$. Then, $\sum_{j=p_*+1}^{\infty} j! \alpha_j \beta_j s^{j-1} > 0$ for all $s > 0$. Also, similarly to (B.2),

$$H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_* - 1) = \underbrace{i! \beta_{I+p_*-1}(\beta_1)^{I-1}}_{\asymp c} + O(c^2).$$

By flipping the sign of β_1, we can change the sign of $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_* - 1)$. Thus, (II) for even I is satisfied by a randomized choice of β_1.

For (II) with odd I, consider coefficients sampled from the second distribution, and $\mathrm{Sign}(\beta_j) = \mathrm{Sign}(\alpha_j)$ for $j \leq (p_* + I) \vee p$. Then, $\sum_{j=p_*+1}^{\infty} j! \alpha_j \beta_j s^{j-1} > 0$ for all $s > 0$.

$$H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_* - 1) = \frac{1}{(p_* - 1)!} \mathbb{E}[\sigma^{(I)}(\sigma^{(1)})^{I-1} \mathsf{He}_{p_*-1}]$$

$$= \frac{1}{(p_* - 1)!} \mathbb{E}[(I-1)(\beta_{p_*+I} \mathsf{He}_{p_*+I})^{(I)}(\beta_2 \mathsf{He}_2)^{(1)}(\beta_1)^{I-2} \mathsf{He}_{p_*-1}] + O(c^4).$$

$$= \underbrace{\frac{2(I-1)\beta_{p_*+I} \beta_2 (\beta_1)^{I-2}(p_* + I)!}{(p_* - 1)!}}_{\asymp c^3} + O(c^4).$$

By flipping the sign of β_1, we can change the sign of $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_* - 1)$. Thus, (II) for odd I is satisfied by a randomized choice of β_1.

Therefore, we have constructed a randomized polynomial activation σ that satisfies all of the conditions for weak and strong recovery. Now we provide a sketch of reasoning that when the link function σ_* is well-approximated by ReLU as specified in Assumption 4, we can find some σ that additionally satisfies Assumption 4 by introducing a small ReLU component. Specifically, we add $c_{\mathrm{R}} \cdot \mathrm{ReLU}$ to the activation function with probability $\frac{1}{2}$, with a sufficiently small $c_{\mathrm{R}} = \tilde{\Omega}(1)$, e.g., $c_{\mathrm{R}} = (\log d)^{-C}$ for some $C > 0$. When a two-layer ReLU network approximates σ_* that satisfies Assumption 4, by using the neurons with added ReLU component, σ_* can be approximated up to some polynomial residual with degree $(p_* + I) \vee p$. And by using the remaining polynomial neurons, we can approximate the additional polynomial terms in σ_* (see Lemma 22,23). Subtracting the latter from the former, we obtain the desired approximation result. When c_{R} is sufficiently small, this additional term does not impact the conditions for weak and strong recovery and the moment calculations; similarly, since $c_{\mathrm{R}} \ll 1$ we may discard this non-smooth term before Taylor expansion without affecting the analysis of optimization dynamics. We remark that to avoid such unnatural design of activation function, we can also train the first-layer parameters using a polynomial activation specified above, and then perturb it before the second-layer training to enhance the approximation ability — such strategy has also been employed in prior layer-wise training analysis [AAM22]. $\square$

B.1.3 More Discussion on Assumption 2

Assumption 2 requires $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_* - 1)$ is not zero and has the same sign as $H(\sigma_*^I; p_*)$. We remark that if we allow a negative momentum parameter larger than 1, i.e., setting $\xi^{2(t+1)} = 1 + c_\xi d^{-\frac{(p_*-2)_+}{2}}$, we can negate the opposite sign of $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_* - 1)$ (see Lemma 16), and the subsequent analysis still holds. Therefore, what we essentially need is $H(\sigma^{(i)}(\sigma^{(1)})^{i-1}; k) \neq 0$. Lemma 3 confirms that it is satisfied by almost all polynomials:

Proof of Lemma 3. We note that $H(\sigma^{(i)}(\sigma^{(1)})^{i-1}; k) = \mathbb{E}[\sigma^{(i)}(\sigma^{(1)})^{i-1} \mathsf{He}_k]$ is a polynomial of $\{\beta_j\}_{j=0}^{C_\sigma}$. This polynomial is not identically equal to zero. To confirm this, consider $\sigma = x^{C_\sigma} + x^{C_\sigma - 1}$. Because $\sigma^{(i)}(\sigma^{(1)})^{i-1}$ is expanded as a sum of $x^l(i(C_\sigma - 3) \leq l \leq i(C_\sigma - 2) + 1$ with positive coefficients and each x^l is a sum of $\mathsf{He}_l, \mathsf{He}_{l-2} \cdots$ with positive coefficients, $\sigma^{(i)}(\sigma^{(1)})^{i-1}$ has all positive Hermite coefficients for degree $0, 1, \cdots, i(C_\sigma - 2) + 1$. If $k \leq i(C_\sigma - 2) + 1$,

this choice of σ yields $H(\sigma^{(i)}(\sigma^{(1)})^{i-1}; k) > 0$, which confirms that $H(\sigma^{(i)}(\sigma^{(1)})^{i-1}; k)$ as a polynomial of $\{\beta_j\}_{j=0}^{C_\sigma}$ is not identically equal to zero. Hence the assertion follows from so-called Schwartz–Zippel Lemma [Sch80], or the fact that zeros of a non-zero polynomial form a measure-zero set. $\qquad\square$

B.2 Initialization

We first consider the initial alignment. In the following sections, we focus on the neurons that satisfy $\kappa_j^0 = \boldsymbol{\theta}^\top \boldsymbol{w}_j^0 \geq 2c_\eta^{-1}d^{-\frac{1}{2}}$ at the initialization. The following lemma states that roughly a constant portion of the neurons satisfy the initial alignment condition upon random initialization. In particular, if we take $c_\eta = \Omega((\log\log d)^{-\frac{1}{2}})$, the fraction of neurons that satisfy the initial alignment condition is at least $e^{-16c_\eta^{-2}} = \tilde{\Omega}(1)$. Let us write $C_2 = c_\eta^{-1}$ for simplicity in the following.

Lemma 14. *At the time of initialization, $\kappa_j^0 = \boldsymbol{\theta}^\top \boldsymbol{w}^0$ satisfies the following:*

$$\mathbb{P}[\kappa_j^0 \geq 2C_2 d^{-\frac{1}{2}}] = \mathbb{P}[\kappa_j^0 \leq -2C_2 d^{-\frac{1}{2}}] \gtrsim e^{-16C_2^2} = \tilde{\Omega}(1).$$

We make use of the following lemma.

Lemma 15 (Theorem 2 of [CCM11]). *For any $\beta > 1$ and $s \in \mathbb{R}$, we have*

$$\frac{\sqrt{2e(\beta-1)}}{2\beta\sqrt{\pi}} e^{-\frac{\beta s^2}{2}} \leq \int_s^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \,\mathrm{d}t$$

Proof of Lemma 14. Because $\kappa^0 = \boldsymbol{v}^\top \boldsymbol{w} \overset{\mathrm{d}}{=} \frac{\boldsymbol{e}_1^\top \boldsymbol{g}}{\|\boldsymbol{g}\|}$, where $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$,

$$\begin{aligned}
\mathbb{P}[\kappa_j^0 \geq 2C_2 d^{-\frac{1}{2}}] &= \mathbb{P}_{\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\boldsymbol{e}_1^\top \boldsymbol{g} \geq 4C_2 \wedge \|\boldsymbol{g}\| \leq 2d^{\frac{1}{2}}\right] \\
&\geq \mathbb{P}_{\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\boldsymbol{e}_1^\top \boldsymbol{g} \geq 4C_2\right] - \mathbb{P}_{\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\|\boldsymbol{g}\| \geq 2d^{\frac{1}{2}}\right] \\
&\gtrsim \frac{\sqrt{2e(\beta-1)}}{2\beta\sqrt{\pi}} e^{-8\beta C_2^2} - e^{-\Omega(d)},
\end{aligned}$$

where we used Lemma 15 for the final inequality. By letting $\beta = 2$, we have that $\mathbb{P}[\kappa_j^0 \geq C_2 d^{-\frac{1}{2}}] \gtrsim e^{-16C_2^2}$. Because of the symmetry, $\mathbb{P}[\kappa_j^0 \leq 2C_2 d^{-\frac{1}{2}}] = \mathbb{P}[\kappa_j^0 \geq 2C_2 d^{-\frac{1}{2}}]$. $\qquad\square$

B.3 Weak Recovery: Population Update

We divide the first layer training into the first phase (weak recovery) and the second phase (strong recovery). We first evaluate the expected update of two gradient steps with the same training example.

Lemma 16. *Let $\eta^{2t}, \eta^{2t+1} = \eta = c_\eta d^{-1}$, $\xi^{2(t+1)} = \xi = 1 - c_\xi d^{-\frac{(p_*-2)_+}{2}}$. Suppose that the link function satisfies $\mathrm{IE}(\sigma_*^I) = \mathrm{GE}(\sigma_*) = p_*$ (we choose the smallest such I) and activation functions satisfy all of the assumptions in Section B.1 for weak recovery. Then, for $\boldsymbol{w}^{2t}$ with $c_\eta^{-1}d^{-\frac{1}{2}} \leq \boldsymbol{\theta}^\top \boldsymbol{w}^{2t} \leq c_\eta^I$, the alignment $\boldsymbol{\theta}^\top \boldsymbol{w}^{2(t+1)}$ can be evaluated as,*

$$\boldsymbol{\theta}^\top \boldsymbol{w}^{2(t+1)} \geq \boldsymbol{\theta}^\top \boldsymbol{w}^{2t} + c_\eta^I c_\xi c_\sigma d^{-\frac{p_*}{2}\vee 1}(\kappa^{2t})^{p_*-1} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\nu^{2t}.$$

Here $c_\sigma = p_*!\alpha_{p_*}\beta_{p_*}$ (when $\mathrm{IE}(\sigma_*) = \mathrm{GE}(\sigma_*)$) or $c_\sigma = \frac{p_*!H(\sigma_*^I; p_*)H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_*-1)}{2(I-1)!}$ (otherwise), and ν^{2t} is a mean-zero sub-exponential random variable.

Proof. The expected alignment $\boldsymbol{\theta}^\top \boldsymbol{w}^{2(t+1)}$ after two gradient steps from $\boldsymbol{w}^{2t} = \boldsymbol{\omega}$ using the same sample $(\boldsymbol{x}, y)$, step size $\eta^{2t} = \eta^{2t+1} = \eta = c_\eta d^{-1}$ and momentum parameter $\xi^{2(t+1)} = \xi =$

$1 - c_\xi d^{-\frac{(p_*-2)_+}{2}}$ is evaluated as follows. With a projection matrix $\boldsymbol{P}_{\boldsymbol{\omega}} = \boldsymbol{I} - \boldsymbol{\omega}\boldsymbol{\omega}^\top$, the first step updates the weight as

$$\boldsymbol{w}^{2t+1} \leftarrow \boldsymbol{w}^{2t} + \eta\tilde{\nabla}_{\boldsymbol{w}} y\sigma(\boldsymbol{w}^{2t^\top}\boldsymbol{x}) = \boldsymbol{\omega} + \eta y\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}, \tag{B.3}$$

and the next gradient step with the same sample is computed as

$$\begin{aligned}
\tilde{\nabla}_{\boldsymbol{w}} y\sigma(\boldsymbol{w}^{2t+1^\top}\boldsymbol{x}) &= y\sigma'(\boldsymbol{w}^{2t+1^\top}\boldsymbol{x})\boldsymbol{x} \\
&= y\sigma'\big((\boldsymbol{\omega} + \eta y\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x})^\top\boldsymbol{x}\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x} \\
&= y\sigma'\big(\boldsymbol{\omega}^\top\boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}, \tag{B.4}
\end{aligned}$$

here we used the notation $\|\boldsymbol{\theta}\|_A^2 = \boldsymbol{\theta}^\top A\boldsymbol{\theta}$ for a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and a positive symmetric matric $A \in \mathbb{R}^{d\times d}$. From (B.3) and (B.4), the parameter after the two steps is obtained as

$$\begin{aligned}
\boldsymbol{w}^{2(t+1)} &\leftarrow \boldsymbol{w}^{2t+1} + \eta\tilde{\nabla}_{\boldsymbol{w}} y\sigma(\boldsymbol{w}^{2t+1^\top}\boldsymbol{x}) \\
&= \boldsymbol{\omega} + \eta y\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x} + \eta y\sigma'\big(\boldsymbol{\omega}^\top\boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}. \\
&= \boldsymbol{\omega} + \eta\boldsymbol{g}^{2t},
\end{aligned}$$

where

$$\boldsymbol{g}^{2t} = y\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x} + y\sigma'\big(\boldsymbol{\omega}^\top\boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}.$$

Finally, the normalization step yields

$$\boldsymbol{w}^{2(t+1)} \leftarrow \frac{\boldsymbol{w}^{2(t+1)} - \xi^{2(t+1)}(\boldsymbol{w}^{2(t+1)} - \boldsymbol{w}^{2t})}{\|\boldsymbol{w}^{2(t+1)} - \xi^{2(t+1)}(\boldsymbol{w}^{2(t+1)} - \boldsymbol{w}^{2t})\|} = \frac{\boldsymbol{\omega} + \eta\xi\boldsymbol{g}^{2t}}{\|\boldsymbol{\omega} + \eta\xi\boldsymbol{g}^{2t}\|} = \frac{\boldsymbol{\omega} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\boldsymbol{g}^{2t}}{\|\boldsymbol{\omega} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\boldsymbol{g}^{2t}\|}.$$

Therefore, by writing $\boldsymbol{\theta}^\top\boldsymbol{w}^{2t} = \kappa^{2t}$, the update of the alignment is

$$\begin{aligned}
\kappa^{2(t+1)} &= \boldsymbol{\theta}^\top\boldsymbol{w}^{2(t+1)} \\
&= \frac{\kappa^{2t} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\boldsymbol{\theta}^\top\boldsymbol{g}^{2t}}{\|\boldsymbol{\omega} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\boldsymbol{g}^{2t}\|} \\
&\geq \kappa^{2t} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\boldsymbol{\theta}^\top\boldsymbol{g}^{2t} - \frac{1}{2}\kappa^{2t}c_\eta^2 c_\xi^2 d^{-p_*\vee 2}\|\boldsymbol{g}^{2t}\|^2 - \frac{1}{2}c_\eta^3 c_\xi^3 d^{-\frac{3p_*}{2}\vee 3}|\boldsymbol{\theta}^\top\boldsymbol{g}^{2t}|\|\boldsymbol{g}^{2t}\|^2. \quad \text{(B.5)}
\end{aligned}$$

We can easily see that $\mathbb{E}[\|\boldsymbol{g}^{2t}\|^2] \lesssim d$ and $\mathbb{E}[|\boldsymbol{\theta}^\top\boldsymbol{g}^{2t}|\|\boldsymbol{g}^{2t}\|^2] \lesssim d$, which implies that the expectation of the last two terms of (B.5) is bounded by $\lesssim \kappa^{2t}c_\eta^2 c_\xi^2 d^{-(p_*-1)\vee 1} \vee c_\eta^3 c_\xi^3 d^{-(\frac{3p_*}{2}-1)\vee 2} \leq c_\eta^2 c_\xi^2 d^{-(p_*-1)\vee 1}\kappa^{2t}$.

Now we bound $\mathbb{E}[\boldsymbol{\theta}^\top\boldsymbol{g}^{2t}]$ by $\gtrsim c_\eta^{I-1}\kappa^{p_*-1}$. Let C_σ be the maximum degree of the activation function with non-zero coefficients of Hermite expansion, which may be infinity when we consider general link functions, and there appear some infinite sums. For these cases we simply assume the sums converge – we discuss the validity of this condition in Section B.1.2. We omit the subscript $2t$ in the following for simplicity. We divide the analysis into the two cases.

(I) If $I = 1 \Leftrightarrow \text{IE}(\sigma_*) = \text{GE}(\sigma_*) = p_*$. For the first term of $\mathbb{E}[\boldsymbol{\theta}^\top\boldsymbol{g}]$, we have

$$\begin{aligned}
\boldsymbol{\theta}^\top\mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] &= \boldsymbol{\theta}^\top\boldsymbol{P}_{\boldsymbol{\omega}}\mathbb{E}\bigg[\bigg(\sum_{j=p_*}^\infty \alpha_j\mathsf{He}_j(\boldsymbol{\theta}^\top\boldsymbol{x})\bigg)\bigg(\sum_{j=1}^{C_\sigma} j\beta_j\mathsf{He}_{j-1}(\boldsymbol{\omega}^\top\boldsymbol{x})\bigg)\boldsymbol{x}\bigg] \\
&= \boldsymbol{\theta}^\top\boldsymbol{P}_{\boldsymbol{\omega}}\sum_{j=p_*}^\infty \bigg[j!\alpha_j\beta_j(\boldsymbol{\theta}^\top\boldsymbol{\omega})^{j-1}\boldsymbol{\theta} + (j+2)!\alpha_j\beta_{j+2}(\boldsymbol{\theta}^\top\boldsymbol{\omega})^j\boldsymbol{\omega}\bigg] \\
&= \sum_{j=p_*}^{C_\sigma} j!\alpha_j\beta_j(\boldsymbol{\theta}^\top\boldsymbol{\omega})^{j-1}\boldsymbol{\theta}^\top\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{\theta} \\
&= p_*!\alpha_{p_*}\beta_{p_*}\kappa^{p_*-1} + O(\kappa^{p_*}). \tag{B.6}
\end{aligned}$$

For the second term of $\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}]$, the following decomposition can be made.

$$\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2 \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}]$$

$$= \sum_{i=1}^{C_\sigma - 1} (i!)^{-1}\boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}}\mathbb{E}\left[y\sigma^{(i+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big(\eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2 y\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^i \boldsymbol{x}\right] + \boldsymbol{\theta}^\top \mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}]$$

$$= \sum_{i=1}^{C_\sigma - 1} (i!)^{-1}\eta^i \boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}}\mathbb{E}\left[\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^{2i} y^{i+1}\sigma^{(i+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big(\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^i \boldsymbol{x}\right] \tag{B.7}$$

$$+ p_*!\alpha_{p_*}\beta_{p_*}\kappa^{p_*-1} + O(\kappa^{p_*}). \tag{B.8}$$

We evaluate each term in the summation. We need to show that although $\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2$ is a function of $\boldsymbol{\theta}^\top \boldsymbol{x}$ and $\boldsymbol{\omega}^\top \boldsymbol{x}$, it is mostly independent from the two quantities. To verify this, let $\boldsymbol{e} = \frac{\boldsymbol{\theta} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}}{\|\boldsymbol{\theta} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}\|}$ be the orthogonal component of $\boldsymbol{\theta}$ to $\boldsymbol{\omega}$. Then, we have that

$$\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2 = \boldsymbol{x}^\top(I - \boldsymbol{\omega}\boldsymbol{\omega}^\top - \boldsymbol{e}\boldsymbol{e}^\top)\boldsymbol{x} + (\boldsymbol{e}^\top \boldsymbol{x})^2$$

$$= \underbrace{\boldsymbol{x}^\top(I - \boldsymbol{\omega}\boldsymbol{\omega}^\top - \boldsymbol{e}\boldsymbol{e}^\top)\boldsymbol{x}}_{\sim\chi_{d-2}^2,\text{ independent from }\boldsymbol{\omega}^\top \boldsymbol{x}\text{ and }\boldsymbol{\theta}^\top \boldsymbol{x}} + \left(\frac{\boldsymbol{\theta}^\top \boldsymbol{x} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}^\top \boldsymbol{x}}{\|\boldsymbol{\theta} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}\|}\right)^2.$$

With $P_{\boldsymbol{\omega},\boldsymbol{\theta}} = \boldsymbol{I}_d - \boldsymbol{\omega}\boldsymbol{\omega}^\top - \boldsymbol{e}\boldsymbol{e}^\top$, (B.7) is expanded as

$$(\text{B.7}) = \sum_{i=1}^{C_\sigma-1}\sum_{j=0}^{i}\sum_{l=0}^{i+1}\frac{\binom{i}{j}\binom{i+1}{l}\eta^i}{i!\|\boldsymbol{\theta} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}\|^{2j}}$$

$$\boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}}\mathbb{E}\left[(\boldsymbol{x}^\top P_{\boldsymbol{\omega},\boldsymbol{\theta}}\boldsymbol{x})^{i-j}\varsigma^{i+1-l}(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l(\boldsymbol{\theta}^\top \boldsymbol{x} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}^\top \boldsymbol{x})^{2j}\sigma^{(i+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big(\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^i \boldsymbol{x}\right]$$

$$= \sum_{i=1}^{C_\sigma-1}\sum_{j=0}^{i}\sum_{l=0}^{i+1}\sum_{k=0}^{2j}\frac{\binom{i}{j}\binom{i+1}{l}\binom{2j}{k}\eta^i\kappa^k(-1)^k\mathbb{E}[\varsigma^{i+1-l}]\mathbb{E}_{z\sim\chi_{d-2}^2}[z^{i-j}]}{i!\|\boldsymbol{\theta} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}\|^{2j}}$$

$$\mathbb{E}\left[(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l(\boldsymbol{\theta}^\top \boldsymbol{x})^{2j-k}(\boldsymbol{\omega}^\top \boldsymbol{x})^k\sigma^{(i+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big(\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^i(\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\theta}^\top \boldsymbol{\omega}\boldsymbol{\omega}^\top \boldsymbol{x})\right] \tag{B.9}$$

For a general differentiable function $g(\boldsymbol{x})$, we have $\mathbb{E}[\mathsf{He}_t(x_1)g(\boldsymbol{x})] = \mathbb{E}[\frac{\mathrm{d}^t}{\mathrm{d}x_1^t}g(\boldsymbol{x})]$. If $g(\boldsymbol{x})$ is a polynomial (with a bounded coefficients) of x_1 and $\boldsymbol{u}^\top \boldsymbol{x}$ and its degree with respect to x_1 is at most $s(\leq t)$, $|\mathbb{E}[\mathsf{He}_t(x_1)g(\boldsymbol{x})]| \lesssim |u_1|^{t-s}$, because differentiation of $g(\boldsymbol{x}) = \bar{g}(x_1, \boldsymbol{u}^\top \boldsymbol{x})$ is taken with respect to the first variable at most s times. Each term of (B.9) is an expectation of $(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l$, multiplied by the polynomial of $\boldsymbol{\theta}^\top \boldsymbol{x}$ and $\boldsymbol{\omega}^\top \boldsymbol{x}$, where its degree with respect to $\boldsymbol{\theta}^\top \boldsymbol{x}$ is at most $2j - k$. Thus each term of (B.9) is evaluated as (here we omit the constants)

$$\frac{\eta^i\kappa^k(-1)^k\mathbb{E}_{z\sim\chi_{d-2}^2}[z^{i-j}]}{i!\|\boldsymbol{\theta} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}\|^{2j}}\mathbb{E}\left[\underbrace{(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l}_{\text{IE}\geq p_*}\underbrace{(\boldsymbol{\theta}^\top \boldsymbol{x})^{2j-k}(\boldsymbol{\omega}^\top \boldsymbol{x})^k\sigma^{(i+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big(\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^i(\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\theta}^\top \boldsymbol{\omega}\boldsymbol{\omega}^\top \boldsymbol{x})}_{\text{degree w.r.t. }\boldsymbol{\theta}^\top \boldsymbol{x}\text{ is at most }2j-k+1}\right]$$

$$\lesssim c_\eta^i d^{-i}d^{i-j}\kappa^k\kappa^{((p_*-2j+k-1)\vee 0)} \lesssim c_\eta d^{-j}\kappa^{((p_*-2j-1)\vee 0)} \leq c_\eta\kappa^{p_*-1}(d/\kappa^2)^{-j} \leq c_\eta\kappa^{p_*-1}.$$

The lower bound follows in the same fashion. Therefore,

$$|(\text{B.9})| \lesssim c_\eta\kappa^{p_*-1}.$$

Now, $\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}]$ can be evaluated as

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}] = (\text{B.6}) + (\text{B.7}) + (\text{B.8}) = 2p_*!\alpha_{p_*}\beta_{p_*}\kappa^{p_*-1} + O(c_\eta\kappa^{p_*-1} + \kappa^{p_*}).$$

(II) If $I = \{\min i \mid \mathrm{IE}(\sigma_*^i) = \mathrm{GE}(\sigma_*) = p_*\} \geq 2$. Note that $\alpha_j = 0$ for all $j \leq p_*$ from the assumption. Following (B.6), the first term of $\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}]$ is evaluated as

$$\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P_\omega}\boldsymbol{x}] = \sum_{j=p}^{C_\sigma} j! \alpha_j \beta_j (\boldsymbol{\theta}^\top \boldsymbol{\omega})^{j-1} \boldsymbol{\theta}^\top \boldsymbol{P_\omega}\boldsymbol{\theta} = O(\kappa^{p_*}). \tag{B.10}$$

For the second term of $\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}]$, similarly to (B.9), the following decomposition can be made.

$$\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P_\omega}}^2 \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y)\boldsymbol{P_\omega}\boldsymbol{x}]$$
$$= \sum_{i=0}^{C_\sigma-1}\sum_{j=0}^{i}\sum_{l=0}^{i+1}\sum_{k=0}^{2j} \frac{\binom{i}{j}\binom{i+1}{l}\binom{2j}{k}\eta^i\kappa^k(-1)^k \mathbb{E}[\varsigma^{i+1-l}]\mathbb{E}_{z\sim\chi_{d-2}^2}[z^{i-j}]}{i!\|\boldsymbol{\theta} - (\boldsymbol{\theta}^\top \boldsymbol{\omega})\boldsymbol{\omega}\|^{2j}}$$
$$\mathbb{E}\left[(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l(\boldsymbol{\theta}^\top \boldsymbol{x})^{2j-k}(\boldsymbol{\omega}^\top \boldsymbol{x})^k \sigma^{(i+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})(\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x}))^i(\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\theta}^\top \boldsymbol{\omega}\boldsymbol{\omega}^\top \boldsymbol{x})\right] \tag{B.11}$$

Each term of (B.11) (omitting constants) is evaluated as

$$\eta^i\kappa^k\mathbb{E}_{z\sim\chi_{d-2}^2}[z^{i-j}]\mathbb{E}\left[(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l(\boldsymbol{\theta}^\top \boldsymbol{x})^{2j-k}(\boldsymbol{\omega}^\top \boldsymbol{x})^k \sigma^{(i+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})(\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x}))^i(\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\theta}^\top \boldsymbol{\omega}\boldsymbol{\omega}^\top \boldsymbol{x})\right]$$

$$\tag{B.12}$$

$$= c_\eta^i\kappa^k d^{-j}\mathbb{E}\left[\underbrace{(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l}_{\mathrm{IE}\geq\begin{cases}p_* (l \geq I)\\ p_*+1 (l < I)\end{cases}}\underbrace{(\boldsymbol{\theta}^\top \boldsymbol{x})^{2j-k}(\boldsymbol{\omega}^\top \boldsymbol{x})^k \sigma^{(i+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})(\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x}))^i(\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\theta}^\top \boldsymbol{\omega}\boldsymbol{\omega}^\top \boldsymbol{x})}_{\text{degree w.r.t. } \boldsymbol{\theta}^\top \boldsymbol{x} \text{ is at most } 2j-k+1}\right]$$

$$\lesssim c_\eta^i\kappa^k d^{-j}\kappa^{(\mathrm{IE}(\sigma_*^l)-2j+k-1)\vee 0} \tag{B.13}$$

When $i \leq I-2$ and $\eta = c_\eta d^{-1}$, we have $l \leq i+1 < I$ and $\mathrm{IE}((\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l) \geq p_* + 1$. Thus

$$(\text{B.13}) \lesssim \kappa^{p_*}$$

When $i \geq I$, $\mathrm{IE}((\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^l) \geq p_*$ and we get

$$(\text{B.13}) \lesssim c_\eta^I\kappa^{p_*-1}.$$

Now the case of $i = I-1$. When $i = I-1$ and $j \neq 0$, and using the assumption that $\kappa \leq c_\eta$,

$$(\text{B.13}) \lesssim c_\eta^{I-1}\kappa^{p_*-1}(\kappa^{-2}/d) \leq c_\eta^I\kappa^{p_*-1}.$$

When $i = I-1$, $j = 0$, and $k \neq 0$,

$$(\text{B.13}) \lesssim c_\eta^{I-1}\kappa^{p_*}.$$

When $i = I-1$, $j = 0$, $k = 0$, and $l \leq I-1$,

$$(\text{B.13}) \lesssim c_\eta^{I-1}\kappa^{p_*}.$$

Therefore, except for $i = I-1$, $j = 0$, $k = 0$, and $l \leq I-1$, we can bound (B.13) by $\lesssim c_\eta^I\kappa^{p_*-1} + \kappa^{p_*}$. The lower bound follows in the same way. Finally, consider the case of $i = I-1$, $j = 0$, $k = 0$, and $l = I$.

$$(\text{B.12}) = \eta^{I-1}\mathbb{E}_{z\sim\chi_{d-2}^2}[z^{I-1}]\mathbb{E}\left[(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^I\sigma^{(I+1)}(\boldsymbol{\omega}^\top \boldsymbol{x})(\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x}))^{I-1}(\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\theta}^\top \boldsymbol{\omega}\boldsymbol{\omega}^\top \boldsymbol{x})\right]$$

$$= \eta^{I-1}\mathbb{E}_{z\sim\chi_{d-2}^2}[z^{I-1}]\sum_{m=p_*}^{C_\sigma I} m!H(\sigma_*^I;m)H(\sigma^{(I)}(\sigma^{(1)})^{I-1};m-1)(1-\kappa^2)\kappa^{m-1}$$

$$= c_\eta^{I-1} p_*! d^{-(I-1)} \mathbb{E}_{z\sim\chi^2_{d-2}}[z^{I-1}] H(\sigma_*^I; p_*) H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_*-1)(1-\kappa^2)\kappa^{p_*-1} + O(c_\eta^{I-1}\kappa^{p_*}).$$

Putting it all together (recovering the constants omitted in (B.12) again),
(B.11)

$$= c_\eta^{I-1} \frac{p_*! d^{-(I-1)} \mathbb{E}_{z\sim\chi^2_{d-2}}[z^{I-1}]}{(I-1)!} H(\sigma_*^I; p_*) H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_*-1)\kappa^{p_*-1} + O(c_\eta^I\kappa^{p_*-1} + \kappa^{p_*}),$$

and

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}] = (B.10) + (B.11)$$

$$= c_\eta^{I-1} \underbrace{\frac{p_*! d^{-(I-1)} \mathbb{E}_{z\sim\chi^2_{d-2}}[z^{I-1}]}{(I-1)!} H(\sigma_*^I; p_*) H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_*-1)}_{=\Theta(1)} \kappa^{p_*-1} + O(c_\eta^I\kappa^{p_*-1} + \kappa^{p_*}).$$

Combining (i) and (ii), we have

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}] \geq 2c_\eta^{I-1} c_\sigma \kappa^{p_*-1} + O(c_\eta^I \kappa^{p_*-1} + \kappa^{p_*})$$

for a positive constant $c_\sigma = \Theta(1)$. Here $c_\sigma > 0$ satisfies $2c_\sigma = 2p_*!\alpha_{p_*}\beta_{p_*}$ (for (i)) or $2c_\sigma = \frac{p_*! H(\sigma_*^I; p_*) H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; p_*-1)}{(I-1)!}$ (for (ii)). Going back to (B.5), by setting $\nu^{2t} = (\boldsymbol{\theta}^\top \boldsymbol{g}^{2t} - \mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}^{2t}])$, we have

$$\kappa^{2(t+1)} \geq \kappa^{2t} + 2c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1} \mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}^{2t}] + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}(\boldsymbol{\theta}^\top \boldsymbol{g}^{2t} - \mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}^{2t}]) + O(c_\eta^2 c_\xi^2 d^{-(p_*-1)\vee 1}\kappa^{2t})$$

$$= \kappa^{2t} + 2c_\eta^I c_\xi d^{-\frac{p_*}{2}\vee 1} c_\sigma (\kappa^{2t})^{p_*-1} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\nu^{2t}$$
$$+ O\left(c_\eta^2 c_\xi^2 d^{-(p_*-1)\vee 1}(\kappa^{2t})^{2t} + c_\eta^{I+1} c_\xi d^{-\frac{p_*}{2}\vee 1}(\kappa^{2t})^{p_*-1} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}(\kappa^{2t})^{p_*}\right).$$

When $c_\xi \leq c_\eta^I$ and $c_\eta^{-1} d^{-\frac{1}{2}} \leq \kappa \leq c_\eta^I$, terms in the big-$O$ notation is smaller than $c_\eta^I c_\xi d^{-\frac{p_*}{2}\vee 1} c_\sigma(\kappa^{2t})^{p_*-1}$ and we have

$$\kappa^{2(t+1)} \geq \kappa^{2t} + c_\eta^I c_\xi c_\sigma d^{-\frac{p_*}{2}\vee 1}(\kappa^{2t})^{p_*-1} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\nu^{2t}.$$

It is straightforward to check ν^{2t} has sub-Weibull tail. $\qquad\square$

B.4 Weak Recovery: Stochastic Update

This subsection proves weak recovery using the results on population update from the previous section. Specifically, from the previous section, we know that

$$\boldsymbol{\theta}^\top \boldsymbol{w}^{2(t+1)} \geq \boldsymbol{\theta}^\top \boldsymbol{w}^{2t} + c_\eta^I c_\xi c_\sigma d^{-\frac{p_*}{2}\vee 1}(\kappa^{2t})^{p_*-1} + c_\eta c_\xi d^{-\frac{p_*}{2}\vee 1}\nu^{2t},$$

with the mean-zero sub-Weibull random variable ν^{2t} and a positive $c_\sigma = \Theta(1)$. For notational simplicity we write $c_\eta^I c_\sigma = c_1$. The following lemma is a detailed version of Proposition 9.

Lemma 17. *Take* $\eta^{2t}, \eta^{2t+1} = \eta = c_\eta d^{-1}$, $\xi^{2(t+1)} = \xi = 1 - c_\xi d^{-\frac{(p_*-2)_+}{2}}$. *Suppose that the link function satisfies* $\mathrm{IE}(\sigma_*^I) = \mathrm{GE}(\sigma_*) = p_*$ *(we choose the smallest such I) and activation functions satisfy all of the assumptions in Section B.1 for weak recovery. Let*

$$T_{1,1} = C_3 c_\xi^{-1} \begin{cases} d & (\text{if } p_* = \mathrm{GE}(\sigma_*) = 1) \\ d(\log d) & (\text{else if } p_* = \mathrm{GE}(\sigma_*) = 2) \\ d^{p_*-1} & (\text{else } p_* = \mathrm{GE}(\sigma_*) \geq 3), \end{cases}$$

and take $c_\xi \lesssim \delta \mathrm{poly}(c_\eta)$, $c_2 \gtrsim \mathrm{poly}(c_\eta)$, *and* $C_3 \simeq c_1^{-1}$. *If* $\kappa^0 \geq 2c_\eta^{-1} d^{-\frac{1}{2}}$, *there exists some* $\tau_* \leq T_{1,1}$ *such that*

$$\kappa^{2\tau_*} \geq 2c_2,$$

with probability at least $1 - \delta$, *and* $\kappa^{2\tau} \geq 2c_2$ *for all* $\tau_* \leq \tau \leq T_{1,1}$, *with high probability.*

26

We may take $\delta = o_d(1)$ with arbitrarily slow decay. The proof is adapted from [BAGJ21], but our bound on $T_{1,1}$ is slightly sharper (by a $\log d$ factor for $p_* = 2$ and by a $(\log d)^2$ factor for $p_* \geq 3$). For $p_* = 2$, this is because of a trick that we carefully "restart" the dynamics, whose failure probability exponentially decays.

Proof. We divide the proof into the following cases.

(i) When $p_* = 1$. Note that $\{\sum_{s=0}^{\tau} \nu^{2s}\}_{\tau}$ is Martingale with $\mathbb{E}[(\nu^{2s})^2] \lesssim 1$. By Doob's maximal inequality and Markov's inequality, with probability $1 - \delta$, we have

$$\max_{0 \leq \tau \leq T} \left| \sum_{s=0}^{\tau} \nu^{2s} \right|^2 \leq \delta^{-1} \mathbb{E}[(\sum_{s=0}^{T} \nu^{2s})^2] \leq \delta^{-1} \sum_{s=0}^{T} \mathbb{E}[(\nu^{2s})^2] \leq C_1 \delta^{-1}(T+1) \quad \text{(B.14)}$$

for any fixed $T \geq 0$, with a sufficiently large constant $C_1 = \Theta(1)$. In the following we consider the case when (B.14) holds for $T = c_1^{-1} c_\xi^{-1} d - 1$.

If $c_\eta^{-1} d^{-\frac{1}{2}} \leq \kappa^{2t} \leq c_\eta^I$ for all $t = 0, 1, \cdots, \tau$, we have

$$\kappa^{2(\tau+1)} \geq \kappa^{2\tau} + c_1 c_\xi d^{-1} + c_\eta c_\xi d^{-1} \nu^{2\tau}$$

$$\geq 2c_\eta^{-1} d^{-\frac{1}{2}} + c_1 c_\xi (\tau+1) d^{-1} \gamma - c_\eta c_\xi d^{-1} \left| \sum_{s=0}^{\tau} \nu^{2s} \right|. \quad \text{(B.15)}$$

Now, applying (B.14) to get

$$\kappa^{2(\tau+1)} \geq \text{(B.15)} \geq 2c_\eta^{-1} d^{-\frac{1}{2}} + c_1 c_\xi (\tau+1) d^{-1} - c_\eta c_\xi^{\frac{1}{2}} c_1^{-\frac{1}{2}} C_1^{\frac{1}{2}} \delta^{-\frac{1}{2}} d^{-\frac{1}{2}},$$

when $\tau \leq c_1^{-1} c_\xi^{-1} d - 1$. By letting $c_\xi \leq c_\eta^{-4} c_1 C_1^{-1} \delta$, we have $c_\eta^{-1} d^{-\frac{1}{2}} \leq c_\eta c_\xi^{\frac{1}{2}} c_1^{-\frac{1}{2}} C_1^{\frac{1}{2}} \delta^{-\frac{1}{2}} d^{-\frac{1}{2}}$, and

$$\kappa^{2(\tau+1)} \geq c_\eta^{-1} d^{-\frac{1}{2}} + c_1 c_\xi (\tau+1) d^{-1},$$

which verifies $c_\eta^{-1} d^{-\frac{1}{2}} \leq \kappa^{2t}$ for $t = \tau + 1$. Thus, there exists some $\tau_* \leq c_1^{-1} c_\xi^{-1} d$ such that

$$\kappa^{2\tau_*} \geq 4c_2,$$

for $c_1 \leq \frac{1}{4} c_\eta^I$, with probability $1 - \delta$.

Now we prove that $\kappa^{2t} \geq 2c_2$ holds for all $\tau_* \leq t \leq T_{1,1} = C_3 c_\xi^{-1} d$. Because ν^{2t} are mean-zero sub-Weibull random variables, we also have that $\left| \sum_{s=\tau}^{\tau+\tau'-1} \nu^{2s} \right| \leq C_4 \sqrt{\tau'}$ for all $0 \leq \tau, \tau' \leq T_{1,1}$ with high probability. Also, because $\eta^t \ll d^{-1}$ and $|1 - \xi^t| \ll 1$, we can easily see that $|\kappa^{2(\tau+1)} - \kappa^{2\tau}| = \tilde{O}(d^{-1})$ for all $\tau = 0, 1, \cdots, T_{1,1} - 1$, with high probability. Thus, when there exists $\tau \geq \tau_*$ such that $\kappa^{2(\tau-1)} \geq 4c_2$ and $\kappa^{2\tau} < 4c_2$, we have $\kappa^{2\tau} \geq 3c_2$ with high probability. Moreover, following the above argument, we can inductively show that

$$\kappa^{2(\tau+\tau')} \geq 3c_2 + c_1 c_\xi \tau' d^{-1} - c_\eta c_\xi d^{-1} C_4 \sqrt{\tau'}$$

$$\geq 3c_2 + c_1 c_\xi \tau' d^{-1} - \begin{cases} c_2 & (\tau' \leq c_\eta^{-2} c_\xi^{-2} C_4^{-2} c_2^2 d^2) \\ c_1 c_\xi \tau' d^{-1} & (\tau' \geq c_\eta^2 c_1^{-2} C_4^2) \end{cases}.$$

$$\geq 2c_2,$$

for $\tau' \leq T_{1,1} = C_3 c_\xi^{-1} d$ or until $\kappa^{2(\tau+\tau')} \geq 4c_2$ holds again. By repeating this argument (if there are multiple such τ), we see that $\kappa^{2t} \geq 2c_2$ holds for all $\tau_* \leq t \leq T_{1,1} = C_3 c_\xi^{-1} d$ with high probability.

(ii) When $p_* = 2$. We define $\iota_0 = 0, \iota_1 = \log_{(1+c_1 c_\xi d^{-1})}(4), \iota_2 = 2\log_{(1+c_1 c_\xi d^{-1})}(4), \ldots$. We show that, for each i, if $\kappa^{2\iota_i} \geq 2c_\eta^{-1} d^{-\frac{1}{2}}$, we have $\kappa^{2(\iota_{i+1})} \geq 2\kappa^{2\iota_i}$, with probability at least $1 - \delta 4^{-i}$, or there exists some t ($\iota_i < t \leq \iota_{i+1}$) with $\kappa^{2t} > c_\eta^I$.

Assume that the above statement holds until some $i - 1 \geq 0$ (we do not need to assume anything for $i = 0$). Then, we have $\kappa^{2\iota_i} \geq 2^i \kappa^0 \geq 2c_\eta^{-1} d^{-\frac{1}{2}}$. Similarly to (B.15), if $c_\eta^{-1} d^{-\frac{1}{2}} \leq \kappa^{2t} \leq c_\eta^I$ for all $t = \iota_i, \iota_i + 1, \cdots, \tau$, we have

$$\kappa^{2(\tau+1)} \geq \kappa^{2\iota_i} + c_1 c_\xi d^{-1} \sum_{s=\iota_i}^{\tau} \kappa^{2s} - c_\eta c_\xi d^{-1} \left| \sum_{s=\iota_i}^{\tau} \nu^{2s} \right|.$$

27

Applying (B.14) with $\delta = \delta/4^i$ and $T = \frac{1}{4}c_\eta^{-2}c_\xi^{-2}C^{-1}(\delta/4^i)(\kappa^{2\iota_i})^2 d^2 - 1$ to get

$$\kappa^{2(\tau+1)} \geq \kappa^{2\iota_i} + c_1 c_\xi d^{-1} \sum_{s=\iota_i}^{\tau} \kappa^{2s} - c_\eta c_\xi d^{-1} C^{\frac{1}{2}} \delta^{-\frac{1}{2}} \sqrt{\tau + 1 - \iota_i}$$

$$\geq \kappa^{2\iota_i} + c_1 c_\xi d^{-1} \sum_{s=\iota_i}^{\tau} \kappa^{2s} - \frac{1}{2}\kappa^{2\iota_i}$$

when $\tau \leq \iota_i + \frac{1}{4}c_\eta^{-2}c_\xi^{-2}C^{-1}(\delta/4^i)(\kappa^{2\iota_i})^2 d^2 - 1$, which verifies $c_\eta^{-1}d^{-\frac{1}{2}} \leq \frac{1}{2}\kappa^{2\iota_i} \leq \kappa^{2t}$ for $t = \tau + 1$.

This implies that, with probability $1 - \delta/4^i$, we have

$$\kappa^{2(\tau+1)} \geq \frac{1}{2}\kappa^{2\iota_i} + c_1 c_\xi d^{-1} \sum_{s=\iota_i}^{\tau} \kappa^{2s}$$

for all $\tau = \frac{1}{4}c_\eta^{-2}c_\xi^{-2}C^{-1}(\delta/4^i)(\kappa^{2\iota_i})^2 d^2 - 1$, which is equivalent to

$$\kappa^{2\tau} \geq (1 + c_1 c_\xi d^{-1})^{\tau - \iota_i} \frac{1}{2}\kappa^{2\iota_i}$$

for all $\tau = \iota_i, \iota_i + 1, \cdots, \iota_i + \frac{1}{4}c_\eta^{-2}c_\xi^{-2}C^{-1}(\delta/4^i)(\kappa^{2\iota_i})^2 d^2$. By taking $c_\xi \ll c_1 c_\eta^{-2}C^{-1}(\delta/4^i)(\kappa^{2\iota_i})^2 d$, we have $\frac{1}{4}c_\eta^{-2}c_\xi^{-2}C^{-1}(\delta/4^i)(\kappa^{2\iota_i})^2 d^2 \geq \log_{(1+c_1 c_\xi d^{-1})}(4)$, and we get

$$\kappa^{2\iota_{i+1}} \geq 2\kappa^{2\iota_i}$$

with probability $1 - \delta/4^i$ (or there exists $t \leq \iota_{i+1}$ such that $\kappa^{2t} > c_\eta^I$).

Thus, by induction, for all i, we have that

$$\kappa^{2\iota_i} \geq 2^i \kappa^0, \tag{B.16}$$

or that there exists some $t \leq \iota_i$ such that κ^{2t} is larger than c_η^I, with probability $1 - \delta$.

The LHS of (B.16) becomes larger than c_η^I for some $i \leq \log d$. Because $\iota_i = \Theta(i c_1^{-1} c_\xi^{-1} d)$, within $O(c_1^{-1} c_\xi d \log d)$ steps, there exists at least one $\tau_* = O(c_1^{-1} c_\xi^{-1} d \log d)$ such that $\kappa^{2\tau_*} \geq 4c_2$ for $c_2 \leq \frac{1}{4}c_\eta^I$, with probability $1 - \delta$.

Once such τ_* is obtained, following the last paragraph of (i), we can see that $\kappa^{2t} \geq 2c_2$ holds until $t = T_{1,1}$ with high probability.

(iii) When $p_* \geq 3$. We apply (B.14) with $T = \frac{1}{p_*-2}c_1^{-1}c_\xi^{-1}d^{\frac{p_*}{2}}(\kappa^0)^{-(p_*-2)}$ to obtain that

$$c_\eta c_\xi d^{-\frac{p_*}{2}} \left| \sum_{s=0}^{\tau} \nu^{2s} \right| \leq c_\eta c_\xi^{\frac{1}{2}} c_1^{-\frac{1}{2}} C^{\frac{1}{2}} \delta^{-\frac{1}{2}} d^{-\frac{p_*}{4}}(\kappa^0)^{-\frac{p_*-2}{2}} \tag{B.17}$$

for all $\tau = 0, 1, \cdots, T - 1$, with probability $1 - \delta$.

We take $c_\xi \ll c_\eta^{-2}c_1 C^{-1}\delta d^{\frac{p_*}{4}}(\kappa^0)^{\frac{p_*}{2}}$ so that (B.17) is bounded by $c_1^{-1}d^{-\frac{1}{2}}$. Then,

$$\kappa^{2(\tau+1)} \geq \kappa^0 + c_1 c_\xi d^{-\frac{p_*}{2}} \sum_{s=0}^{\tau} (\kappa^{2s})^{p_*-1} + c_\eta c_\xi d^{-\frac{p_*}{2}} \sum_{s=0}^{\tau} \nu^{2s}$$

$$\geq c_\eta^{-1}d^{-\frac{1}{2}} + c_1 c_\xi d^{-\frac{p_*}{2}} \sum_{s=0}^{\tau} (\kappa^{2\tau})^{p_*-1}.$$

It is easy to see that $\kappa^{2(\tau+1)}$ is lower bounded by $a^{\tau+1}$, where $a^0 = c_\eta^{-1}d^{-\frac{1}{2}}$ and $a^{\tau+1} = a^\tau + c_1 c_\xi d^{-\frac{p_*}{2}}(a^\tau)^{p_*-1}$. By applying Lemma 18, we have

$$\kappa^{2\tau} \geq \frac{\kappa^0}{\left(1 - c_1 c_\xi d^{-\frac{p_*}{2}}(p_* - 2)(\kappa^0)^{(p_*-2)}t\right)^{\frac{1}{p_*-2}}}.$$

Thus, until $\tau \leq \left(c_1 c_\xi d^{-\frac{p_*}{2}} (p_* - 2)(\kappa^0)^{(p_* - 2)}\right)^{-1} \leq T + 1 \ll d^{p_* - 1}$, with probability at least $1 - \delta$, there exists some τ_* such that

$$\kappa^{2\tau_*} \geq 4c_2 \geq c_\eta^I$$

when $c_2 \leq \frac{1}{4} c_\eta^I$.

Once such τ_* is obtained, following the last paragraph of (i), we can see that $\kappa^{2t} \geq 2c_2$ holds until $t = T_{1,1}$ with high probability. $\qquad \square$

In the above proof we used the (discrete version of) Bihari–LaSalle inequality from [BAGJ22].

Lemma 18. *For $p \geq 3$ and $c > 0$, consider a positive sequence $(a^t)_{t \geq 0}$ such that*

$$a^{t+1} = a^t + c(a^t)^{p-1}.$$

Then, we have

$$a^t \geq \frac{a^0}{\left(1 - c(p-2)(a^0)^{(p-2)} t\right)^{\frac{1}{p-2}}}.$$

Proof. From definition, we have

$$c = \frac{a^{t+1} - a^t}{(a^t)^{p-1}} \leq \int_{t=a^t}^{a^{t+1}} \frac{1}{x^{p-1}} \leq \frac{1}{p-2} \left[\frac{1}{(a^t)^{p-2}} - \frac{1}{(a^{t+1})^{p-2}} \right].$$

Taking the summation and re-arranging the terms yield

$$(a^t)^{-(p-2)} \leq (a^0)^{-(p-2)} - c(p-2)t,$$

$$\therefore a^t \geq \frac{a^0}{\left(1 - c(p-2)(a^0)^{(p-2)} t\right)^{\frac{1}{p-2}}},$$

which gives the lower bound. $\qquad \square$

B.5 From Weak Recovery to Strong Recovery

In the previous subsection, we proved that after $t = 2T_{1,1} = \tilde{\Theta}(d)$ steps, with probability $\tilde{\Omega}(1)$ over the randomness of initialization, we obtain nontrivial alignment $\kappa_j^{2T_{1,1}} \geq 2c_2$. This subsection discusses how to convert the weak recovery into the strong recovery.

Lemma 19. *Suppose the neuron satisfies $\kappa^{2T_{1,1}} \geq 2c_2$. Take $\eta^{2t} = \eta = \bar{c}_\eta \varepsilon d^{-1}$, $\eta^{2t+1} = 0$, $\xi^{2(t+1)} = 0$ for all $t \geq T_{1,1}$, where $\bar{c}_\eta \lesssim \mathrm{poly}(c_1)$. If the activation functions satisfy all of the assumptions in Section B.1 for strong recovery, then we have*

$$\boldsymbol{\theta}^\top \boldsymbol{w}^{2(T_{1,1} + \tau_*)} \geq 1 - \varepsilon,$$

with high probability, where $\tau_* \leq T_{1,2} = C_3 d\varepsilon^{-2}$. Moreover, $\boldsymbol{\theta}^\top \boldsymbol{w}^{2(T_{1,1} + t)} \geq 1 - \varepsilon$ for all $\tau_* \leq t \leq T_{1,2} = C_3 d\varepsilon^{-2}$, with high probability.

Proof. Consider the Hermite expansions of σ_* and σ. Let p be the smallest degree that both σ_* and σ have non-zero coefficients. First we compute the population gradient (of the correlation term) as

$$\mathbb{E}\left[\tilde{\nabla}_{\boldsymbol{w}} y \sigma(\boldsymbol{w}^{2t\top} \boldsymbol{x})\right] = \mathbb{E}\left[\tilde{\nabla}_{\boldsymbol{w}} \left(\sum_{j=p}^{\infty} \alpha_j \mathsf{He}_j(\boldsymbol{\theta}^\top \boldsymbol{x}) \right) \left(\sum_{j=0}^{\infty} \beta_j \mathsf{He}_j(\boldsymbol{w}^{2t\top} \boldsymbol{x}) \right) \right]$$

$$= \sum_{j=p}^{\infty} \left[j! \alpha_j \beta_j (\boldsymbol{\theta}^\top \boldsymbol{w}^{2t})^{j-1} \boldsymbol{\theta} + (j+2)! \alpha_j \beta_{j+2} (\boldsymbol{\theta}^\top \boldsymbol{w}^{2t})^j \boldsymbol{w}^{2t} \right]. \quad \text{(B.18)}$$

Applying $P_{\boldsymbol{w}^{2t}}$, we have

$$\mathbb{E}\left[P_{\boldsymbol{w}^{2t}} \tilde{\nabla}_{\boldsymbol{w}} y \sigma(\boldsymbol{w}^{2t\top} \boldsymbol{x})\right] = (\boldsymbol{\theta} - (\boldsymbol{w}^{2t\top} \boldsymbol{\theta}) \boldsymbol{w}^{2t}) \sum_{j=p}^{\infty} j! \alpha_j \beta_j (\boldsymbol{\theta}^\top \boldsymbol{w}^{2t})^{j-1}. \quad \text{(B.19)}$$

Thus, the update of the alignment $\kappa^{2t} = \boldsymbol{\theta}^\top \boldsymbol{w}^{2t}$ is

$$\kappa^{2(t+1)} \geq \kappa^{2t} + \eta\boldsymbol{\theta}^\top\boldsymbol{g} - \frac{1}{2}\eta^2\kappa^{2t}\|\boldsymbol{g}\|^2 - \frac{1}{2}\eta^3\tilde{\eta}^3|\boldsymbol{\theta}^\top\boldsymbol{g}|\|\boldsymbol{g}\|^2,$$

where

$$\boldsymbol{g} = P_{\boldsymbol{w}^{2t}} y\sigma'(\boldsymbol{w}^{2t\top}\boldsymbol{x})\boldsymbol{x}.$$

From (B.18), the expectation of (B.19) is bounded by

$$\mathbb{E}[\kappa^{2(t+1)}] \geq \kappa^{2t} + \eta(1 - (\kappa^{2t})^2)\sum_{j=p}^{\infty} j!\alpha_j\beta_j(\boldsymbol{\theta}^\top\boldsymbol{w}^{2t})^{j-1} - \eta^2 C_4 d(\kappa^{2t} + \eta)$$

$$\geq \kappa^{2t} + \eta(1 - (\kappa^{2t})^2)\sum_{j=p}^{\infty} j!\alpha_j\beta_j(\kappa^{2t})^{p-1} - \eta^2 C_4 d(\kappa^{2t} + \eta).$$

By letting $\eta \leq c_1^{p-1}\varepsilon d^{-1}$, when $\kappa^{2t} \leq 1 - \varepsilon$, we have

$$\mathbb{E}[\kappa^{2(t+1)}] \geq \kappa^{2t} + \frac{1}{2}\eta\varepsilon\sum_{j=p}^{\infty} j!\alpha_j\beta_j(\kappa^{2t})^{p-1} \geq \kappa^{2t} + \eta\varepsilon c_1^p.$$

It is easy to see that the noise ν^{2t} has sub-Weibull tail and we obtain that

$$\kappa^{2(t+1)} \geq \kappa^{2t} + \frac{1}{2}\eta\varepsilon\sum_{j=p}^{\infty} j!\alpha_j\beta_j(\kappa^{2t})^{p-1} + \eta\nu^{2t} \geq \kappa^{2t} + \eta\varepsilon c_1^p + \eta\nu^{2t}. \tag{B.20}$$

Suppose that $2c_2 \leq \kappa^{2(T1,1+\tau)} \leq 1 - \varepsilon$ for all $t = 0, 1, \ldots, \tau - 1$. By taking the summation of (B.20), we have

$$\kappa^{2(T1,1+\tau)} \geq \kappa^{2T1,1} + \eta\varepsilon t c_1^p + \eta\sum_{s=T1,1}^{T1,1+\tau-1} \nu^{2t} \geq 2c_2 + \eta\varepsilon\tau c_1^p - C_4\eta\sqrt{\tau}, \tag{B.21}$$

with high probability. The third term is bounded by $C_4\eta\sqrt{\tau} \leq c_2$ when $\tau \leq c_2^2 C_4^{-2}\eta^{-2} = c_2^2 C_4^{-2}\bar{c}_\eta^{-2}\varepsilon^{-2}d^2$ and by $\frac{1}{2}\eta\varepsilon\tau c_1^p$ when $\tau \geq 4\varepsilon^{-2}c_1^{-2p}C_4^2$. Because $c_2^2 C_4^{-2}\bar{c}_\eta^{-2}\varepsilon^{-2}d^2 \geq 4\varepsilon^{-2}c_1^{-2p}C_4^2$, we can bound (B.21) by

$$\kappa^{2(T1,1+\tau)} \geq c_2 + \frac{1}{2}\eta\varepsilon\tau c_1^p, \tag{B.22}$$

which verifies $2c_2 \leq \kappa^{2(T1,1+\tau)}$.

Therefore, by induction, until $\kappa^{2t} \geq 1 - \varepsilon$, we have the lower bound (B.22), whose RHS exceeds $1 - \varepsilon$ when $\tau \geq 2\eta^{-1}\varepsilon^{-1}c_1^{-p} \leq C_3 d\varepsilon^{-2}$. Thus, there exists $\tau_* \leq T_{1,2} = C_3 d\varepsilon^{-2}$ such that $\kappa^{2(T1,1+\tau_*)} \geq 1 - \varepsilon$, with high probability.

Now, what remains is to prove that $\kappa^{2(T1,1+\tau)} \geq 1 - 3\varepsilon$ holds for all $\tau_* \leq t \leq T_{1,2} = C_3 d\varepsilon^{-2}$. Because ν^{2t} are mean-zero sub-Weibull random variables, we have that $|\sum_{s=\tau}^{\tau+\tau'-1} \nu^{2s}| \leq C_4\sqrt{\tau'}$ for all $0 \leq \tau, \tau' \leq T_{1,1}$ with high probability. Also, because $\eta^t \ll \varepsilon d^{-1}$, we can easily see that $|\kappa^{2(\tau+1)} - \kappa^{2\tau}| = \tilde{O}(\varepsilon d^{-1})$ for all $\tau = 0, 1, \cdots, T_{1,1} - 1$, with high probability. Thus, when there exists $\tau \geq \tau_*$ such that $\kappa^{2(T_{1,1}+\tau-1)} \geq 1 - \varepsilon$ and $\kappa^{2(T_{1,1}+\tau)} < 1 - \varepsilon$, we have $\kappa^{2(T_{1,1}+\tau)} \geq 1 - 2\varepsilon$ with high probability. Moreover, following the above argument, we can inductively show that

$$\kappa^{2(T1,1+\tau+\tau')} \geq 1 - 2\varepsilon + \eta\varepsilon\tau'c_1^p - C_4\eta\sqrt{\tau'}$$

$$\geq 1 - 2\varepsilon + \eta\varepsilon\tau'c_1^p - \begin{cases} \varepsilon & (\tau' \leq \bar{c}_\eta^{-2}C_4^{-2}d^2) \\ \eta\varepsilon\tau'c_1^p & (\tau' \geq \varepsilon^{-2}C_4^2 c_1^{-2p}) \end{cases}.$$

$$\geq 1 - 3\varepsilon,$$

for $\tau' \leq T_{1,2}$ or until $\kappa^{2(T1,1+\tau+\tau')} \geq 1 - \varepsilon$ holds again. Note that the last inequality follows from $\bar{c}_\eta^{-2}C_4^{-2}d^2 \geq \varepsilon^{-2}C_4^2 c_1^{-2p}$. By repeating this argument (if there are multiple such τ), we can see that $\kappa^{2(T1,1+t)} \geq 1 - \varepsilon$ holds for all $\tau_* \leq t \leq T_{1,2} = C_3 d\varepsilon^{-2}$ with high probability.

Adjusting hidden constants to remove a factor of 3 from 3ε yields the desired result. □

B.6 Second Layer Training

From the previous analysis, we know that at least $\Omega(1)$ portion of the neurons will satisfy the weak and strong recovery conditions (Appendix B.1), at least $\tilde{\Omega}(1)$ portion of the neurons (independent from the choice of σ_j) satisfy initial alignment conditions (Appendix B.2), and at least $1 - o(1)$ fraction of them achieves strong recovery. This subsection proves a generalization error bound after second-layer training. Let $f_{\boldsymbol{a}}(\boldsymbol{x}) = f_{\boldsymbol{\Theta}}(\boldsymbol{x})$ for $\boldsymbol{\Theta} = (\hat{\boldsymbol{w}}_j, a_j, \hat{b}_j)_{j=1}^N$ where $\boldsymbol{a} \in \mathbb{R}^N$ and $(\hat{\boldsymbol{w}}_j, \hat{b}_j)_{j=1}^N$ are the parameters trained in the first stage. Let $\boldsymbol{a}^* \in \mathbb{R}^N$ be the "certificate" with $\|\boldsymbol{a}^*\|^2 = \tilde{O}(N)$ that is shown to exist in Lemma 22.

Polynomial Link Functions. The following lemma is a complete version of Proposition 5.

Lemma 20. *There exists a $4q$-th order polynomial $Q(R_{\boldsymbol{w}}, b, q')$ of $R_{\boldsymbol{w}} = \max_j \|\boldsymbol{w}_j\|$, $b = (b_j)_{j=1}^N$ such that, if we set $\lambda = \Theta\left(\sqrt{\frac{2}{T_2 \delta_0} N^2 Q(R_w, b, q')}\right)$ for some $\delta_0 > 0$, the ridge estimator $\hat{\boldsymbol{a}}$ satisfies*

$$\|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_x)} \lesssim (N^{-2} + \varepsilon^2) + \frac{1}{T_2 \lambda \delta_0} \left(2N^2 Q(R_w, b, q') + \mathbb{E}_{\boldsymbol{x}}[(f_*)^4]\right) + \frac{3\lambda}{2}\|\boldsymbol{a}^*\|^2 \quad \text{(B.23)}$$

with probability $1 - \delta_0$. Hence taking $T_2 = \tilde{\Theta}((N^4 Q^2(R_w, b, q') + \mathbb{E}[f_*(\boldsymbol{x})^4]^2)\varepsilon^{-4})$ and $N = \tilde{\Theta}(\varepsilon^{-1})$, we have

$$\mathbb{E}_{\boldsymbol{x}}[(f_{\hat{\boldsymbol{a}}}(\boldsymbol{x}) - f_*(\boldsymbol{x}))^2] \lesssim \varepsilon^2.$$

Proof. Let P_{T_2} be the empirical distribution of the second stage: $P_{T_2} := \frac{1}{T_2} \sum_{i=1}^{T_2} \delta_{\boldsymbol{x}_i}$. Let $\psi(\boldsymbol{x}) = (\sigma(\langle \boldsymbol{x}, \hat{\boldsymbol{w}}_j \rangle) + b_j))_{j=1}^N$ so that $f_{\boldsymbol{a}}(\boldsymbol{x}) = \langle \boldsymbol{a}, \psi(\boldsymbol{x}) \rangle$.

Part (1). We first bound the term $\|f_{\hat{\boldsymbol{a}}} - f_*\|_{L^2(P_{T_2})}$. Since $\hat{\mathcal{L}}(f_{\hat{\boldsymbol{a}}}) + \lambda \|\hat{\boldsymbol{a}}\|^2 \le \hat{\mathcal{L}}(f_{\boldsymbol{a}^*}) + \lambda \|\boldsymbol{a}^*\|^2$, we have

$$\|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_{T_2})} + \lambda \|\hat{\boldsymbol{a}}\|^2 \quad \text{(B.24)}$$

$$\le \|f_{\boldsymbol{a}^*} - f_*\|^2_{L^2(P_{T_2})} + \frac{2}{T_2} \sum_{i=1}^{T_2} (f_{\boldsymbol{a}^*}(\boldsymbol{x}_i) - f_{\hat{\boldsymbol{a}}}(\boldsymbol{x}_i))\varepsilon_i + \lambda \|\boldsymbol{a}^*\|^2.$$

Now, by the Cauchy-Schwarz inequality, we have

$$\frac{2}{T_2} \sum_{i=1}^{T_2} (f_{\boldsymbol{a}^*}(x_i) - f_{\hat{\boldsymbol{a}}}(x_i))\varepsilon_i = (\boldsymbol{a}^* - \hat{\boldsymbol{a}})^\top \frac{2}{T_2} \sum_{i=1}^{T_2} \psi(\boldsymbol{x}_i)\varepsilon_i$$

$$\le 2\|\boldsymbol{a}^* - \hat{\boldsymbol{a}}\| \sqrt{\frac{\sum_{i,j} \varepsilon_i \varepsilon_j \psi(\boldsymbol{x}_i)^\top \psi(\boldsymbol{x}_j)}{T_2^2}}.$$

By applying Markov's inequality to the right hand side, it can be further bounded by

$$\|\boldsymbol{a}^* - \hat{\boldsymbol{a}}\| \sqrt{\frac{\mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^2]}{T_2 \delta_1}} \le \frac{\lambda}{2}\|\hat{\boldsymbol{a}}\|^2 + \frac{\lambda}{2}\|\boldsymbol{a}^*\|^2 + \frac{\mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^2]}{T_2 \delta_1 \lambda},$$

with probability $1 - \delta_1$. Thus, by combining with (B.24), we arrive at

$$\|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_{T_2})} + \frac{\lambda}{2}\|\hat{\boldsymbol{a}}\|^2 \le \|f_{\boldsymbol{a}^*} - f_*\|^2_{L^2(P_{T_2})} + \frac{\mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^2]}{T_2 \delta_1 \lambda} + \frac{3\lambda}{2}\|\boldsymbol{a}^*\|^2.$$

Here, by using the evaluation $\|f_{\boldsymbol{a}^*} - f_*\|_{L^2(P_{T_2})} = \tilde{O}(N^{-1} + \varepsilon)$ in Lemma 22, the right hand side can be further bounded by

$$\|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_{T_2})} + \frac{\lambda}{2}\|\hat{\boldsymbol{a}}\|^2 \le \tilde{O}(N^{-2} + \varepsilon^2) + \frac{\mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^2]}{T_2 \delta_1 \lambda} + \frac{3\lambda}{2}\|\boldsymbol{a}^*\|^2.$$

Part (2). Next we lower bound $\|f_{\hat{a}} - f_*\|^2_{L^2(P_{T_2})}$ by noticing that

$$
\|f_{\hat{a}} - f_*\|^2_{L^2(P_{T_2})}
$$
$$
= \|f_{\hat{a}} - f_*\|^2_{L^2(P_{T_2})} - \|f_{\hat{a}} - f_*\|^2_{L^2(P_x)} + \|f_{\hat{a}} - f_*\|^2_{L^2(P_x)}
$$
$$
= \|f_{\hat{a}}\|^2_{L^2(P_{T_2})} - \|f_{\hat{a}}\|^2_{L^2(P_x)} - 2\left(\frac{1}{T_2} \sum_{i=1}^{T_2} f_{\hat{a}}(\boldsymbol{x}_i) f_*(\boldsymbol{x}_i) - \mathbb{E}[f_{\hat{a}}(\boldsymbol{x}_i) f_*(\boldsymbol{x}_i)] \right)
$$
$$
+ \|f_*\|^2_{L^2(P_{T_2})} - \|f_*\|^2_{L^2(P_x)} + \|f_{\hat{a}} - f_*\|^2_{L^2(P_x)}. \tag{B.25}
$$

The first two terms of Eq. (B.25) can be bounded by

$$
\left| \|f_{\hat{a}}\|^2_{L^2(P_{T_2})} - \|f_{\hat{a}}\|^2_{L^2(P_x)} \right| = \left| \hat{\boldsymbol{a}}^\top \left(\frac{\sum_{i=1}^{T_2} \psi(\boldsymbol{x}_i)\psi(\boldsymbol{x}_i)^\top}{T_2} - \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x})\psi(\boldsymbol{x})^\top] \right) \hat{\boldsymbol{a}} \right|
$$
$$
\leq \|\hat{\boldsymbol{a}}\|^2 \sup_{\boldsymbol{a}:\|\boldsymbol{a}\|\leq 1} \left| \|f_{\boldsymbol{a}}\|^2_{L^2(P_{T_2})} - \|f_{\boldsymbol{a}}\|^2_{L^2(P_x)} \right|.
$$

The standard Rademacher complexity bound yields that

$$
\mathbb{E}_{(x_i)_{i=1}^{T_2}} \left[\sup_{\boldsymbol{a}\in\mathbb{R}^N:\|\boldsymbol{a}\|\leq 1} \left| \|f_{\boldsymbol{a}}\|^2_{L^2(P_x)} - \|f_{\boldsymbol{a}}\|^2_{L^2(P_{T_2})} \right| \right]
$$
$$
\leq 2\mathbb{E}_{(x_i,\sigma_t)_{t=1}^{T_2}} \left[\sup_{\boldsymbol{a}\in\mathbb{R}^N:\|\boldsymbol{a}\|\leq 1} \left| \frac{1}{T_2} \sum_{t=1}^{T_2} \sigma_t f_{\boldsymbol{a}}(x_i)^2 \right| \right]
$$
$$
\leq 2\sqrt{ \mathbb{E}_{(\boldsymbol{x}_i)_{i=1}^{T_2}} \left[\sup_{\boldsymbol{a}\in\mathbb{R}^N:\|\boldsymbol{a}\|\leq 1} \frac{1}{T_2^2} \sum_{i=1}^{T_2} (\boldsymbol{a}^\top \psi(\boldsymbol{x}_i))^4 \right] }
$$
$$
\leq 2\sqrt{ \mathbb{E}_{(\boldsymbol{x}_i)_{i=1}^{T_2}} \left[\frac{1}{T_2^2} \sum_{i=1}^{T_2} \|\psi(\boldsymbol{x}_i)\|^4 \right] }
$$
$$
= 2\sqrt{ \frac{1}{T_2} \mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^4] },
$$

where $(\sigma_i)_{i=1}^{T_2}$ is the i.i.d. Rademacher sequence independent of $(\boldsymbol{x}_i)_{i=1}^{T_2}$. Hence, Markov's inequality yields

$$
\left| \|f_{\hat{a}}\|^2_{L^2(P_{T_2})} - \|f_{\hat{a}}\|^2_{L^2(P_x)} \right| = 2\|\hat{\boldsymbol{a}}\|^2 \sqrt{ \frac{1}{T_2\delta_2} \mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^4] },
$$

with probability $1 - \delta_2$.

The third term in Eq. (B.25) can be evaluated as

$$
2\left(\frac{1}{T_2} \sum_{i=1}^{T_2} f_{\hat{a}}(\boldsymbol{x}_i) f_*(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[f_{\hat{a}}(\boldsymbol{x}) f_*(\boldsymbol{x})] \right)
$$
$$
= \hat{\boldsymbol{a}}^\top \left(\frac{1}{T_2} \sum_{i=1}^{T_2} (\psi(\boldsymbol{x}_i) f_*(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x}) f_*(\boldsymbol{x})]) \right)
$$
$$
\leq \|\hat{\boldsymbol{a}}\| \sqrt{ \frac{1}{T_2^2} \sum_{i=1}^{T_2} \sum_{j=1}^{T_2} (\psi(\boldsymbol{x}_i) f_*(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x}) f_*(\boldsymbol{x})])^\top (\psi(\boldsymbol{x}_j) f_*(\boldsymbol{x}_j) - \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x}) f_*(\boldsymbol{x})]) }
$$
$$
\leq \|\hat{\boldsymbol{a}}\| \sqrt{ \frac{1}{T_2\delta_3} \mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x}) f_*(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x}) f_*(\boldsymbol{x})]\|^2] }
$$
$$
\leq \|\hat{\boldsymbol{a}}\| \sqrt{ \frac{1}{T_2\delta_3} \mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^4 + \|f_*(\boldsymbol{x})\|^4] }
$$

$$\leq \frac{\lambda}{4}\|\hat{a}\|^2 + \frac{1}{\lambda T_2 \delta_3} \mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^4 + \|f_*(\boldsymbol{x})\|^4],$$

with probability $1 - \delta_3$ where we used Markov's inequality again in the second inequality. Finally, the fourth and fifth term in Eq. (B.25) can be bounded as

$$
\begin{aligned}
\left| \|f_*\|_{L^2(P_{T_2})}^2 - \|f_*\|_{L^2(P_x)}^2 \right| &= \sqrt{\left(\|f_*\|_{L^2(P_{T_2})}^2 - \|f_*\|_{L^2(P_x)}^2 \right)^2} \\
&\leq \sqrt{\frac{1}{T_2 \delta_4} \mathbb{E}_{\boldsymbol{x}}[(f^*(\boldsymbol{x})^4 - \|f_*\|_{L^2(P_x)}^2)^2]} \\
&\leq \sqrt{\frac{1}{T_2 \delta_4} \mathbb{E}_{\boldsymbol{x}}[(f^*(\boldsymbol{x}))^4]},
\end{aligned}
$$

with probability $1 - \delta_4$ where we used Markov's inequality in the last inequality. Combining these inequalities, we finally arrive at

$$
\begin{aligned}
\|f_{\hat{a}} &- f_*\|_{L^2(P_x)}^2 + \left(\frac{\lambda}{4} - \sqrt{\frac{2}{T_2 \delta_2} \mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^4]} \right) \|\hat{a}\|^2 \\
&\leq \tilde{O}(N^{-2} + \varepsilon^2) + \frac{1}{T_2 \lambda} \left(\frac{\mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^2]}{\delta_1} + \frac{\mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^2]}{\delta_3} + \frac{\mathbb{E}_{\boldsymbol{x}}[(f^*(\boldsymbol{x}))^4]}{\delta_3} \right) + \frac{3\lambda}{2} \|\boldsymbol{a}^*\|^2,
\end{aligned}
$$

with probability $1 - \sum_{j=1}^4 \delta_j$. Hence, by setting $\lambda \geq 8\sqrt{\frac{2}{T_2 \delta_2} \mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^4]}$, we have that

$$
\begin{aligned}
\|f_{\hat{a}} &- f_*\|_{L^2(P_x)}^2 \\
&\leq \tilde{O}(N^{-2} + \varepsilon^2) + \frac{1}{T_2 \lambda} \left(\frac{\mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^2]}{\delta_1} + \frac{\mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^4]}{\delta_3} + \frac{\mathbb{E}_{\boldsymbol{x}}[(f^*(\boldsymbol{x}))^4]}{\delta_3} \right) + \frac{3\lambda}{2} \|\boldsymbol{a}^*\|^2.
\end{aligned}
$$

When the activation function σ is a polynomial, then each $\psi_j(\boldsymbol{x}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle + b_j)$ is an order q-polynomial of a Gaussian random variable $\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle$ which has mean 0 and variance $\mathbb{E}[\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle^2] = \|\boldsymbol{w}_j\|^2 = \tilde{O}(1)$. Then, if we let $R_w := \max_j \|\boldsymbol{w}_j\| = \tilde{O}(1)$, the term $\max_j \max\{\mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x})_j^2], \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x})_j^4]\}$ can be bounded by a $4q$-th order polynomial of R_w and b, which can be denoted by $Q(R_w, b, 4q)$.

Part (3). By combining evaluations of (1) and (2) together, if we let $\lambda = 8\sqrt{\frac{2}{T_2 \delta_0} \mathbb{E}_{\boldsymbol{x}}[\|\psi(\boldsymbol{x})\|^4]}$ for some $\delta_0 > 0$, (by ignoring polylogarithmic factors) we obtain that

$$\|f_{\hat{a}} - f_*\|_{L^2(P_x)}^2 \lesssim (N^{-2} + \varepsilon^2) + \frac{1}{T_2 \lambda \delta_0} \left(2N^2 Q(R_w, b, q') + \mathbb{E}_{\boldsymbol{x}}[(f_*(\boldsymbol{x}))^4] \right) + \frac{3\lambda}{2} \|\boldsymbol{a}^*\|^2,$$

with probability $1 - 4\delta_0$. Thus, since $\|\boldsymbol{a}^*\|^2 = \tilde{O}(N)$, by setting $T_2 = \tilde{\Theta}((N^4 Q^2(R_w, b, q') + \mathbb{E}[f_*(\boldsymbol{x})^4]^2)\varepsilon^{-4})$, and $N = \tilde{\Theta}(\varepsilon^{-1})$, we obtain that (B.23) $\lesssim \varepsilon^2$. $\qquad \square$

Higher Generative Exponent Functions. For general link functions, under Assumption 4 and the bounded fourth moment of the link function, we have the following counterpart of Lemma 20, which provides the formal statement of Proposition 10.

Lemma 21. *Suppose that $\mathbb{E}[\sigma_*(\boldsymbol{\theta}^\top x)^4] < \infty$ and Assumption 4 hold. Then, by setting $\lambda = \tilde{\Theta}\left(\sqrt{\frac{N^2}{T_2 \delta_0}}\right)$ for some $\delta_0 > 0$, the ridge estimator $\hat{a}$ satisfies*

$$\|f_{\hat{a}} - f_*\|_{L^2(P_x)}^2 \lesssim \varepsilon^2 + \frac{1}{\sqrt{T_2 \delta_0}} \left(N^2 C_4 + \mathbb{E}_{\boldsymbol{x}}[(f_*)^4] \right) + \frac{1}{\sqrt{T_2 \delta_0}} \|\boldsymbol{a}^*\|^2,$$

with probability $1 - \delta_0$. By taking $T_2 = \tilde{\Theta}((N^4 + N^2)\varepsilon^{-4})$, we have

$$\mathbb{E}_{\boldsymbol{x}}[(f_{\hat{a}}(\boldsymbol{x}) - f_*(\boldsymbol{x}))^2] \lesssim \varepsilon^2.$$

33

Furthermore, applying Lemma 12 and 13 yields that, when $\sigma_* = \sum_{j=0}^{\infty} \alpha_j \mathsf{He}_j$ satisfies $\sum_{j=0}^{\infty} j^2 j! \alpha_j^2$ and $\mathbb{E}[\sigma_*(\boldsymbol{\theta}^\top x)^4]$ are bounded, with a properly designed randomized activation in Lemma 12, by taking $N = \tilde{\Theta}(\varepsilon^{-7})$ and $T_2 = \tilde{\Theta}(\varepsilon^{-32})$, Algorithm 1 yields

$$\mathbb{E}_{\boldsymbol{x}}[(f_{\hat{\boldsymbol{a}}}(\boldsymbol{x}) - f_*(\boldsymbol{x}))^2] \lesssim \varepsilon^2,$$

with probability $1 - o_d(1)$.

Proof. The proof is identical to that of Lemma 20, with the difference being that we replace the bounded moment assumptions with $\mathbb{E}[\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x})^4] < \infty$ or Assumption 4. $\qquad\square$

Approximation Guarantee. Note that for non-polynomial link function with generative exponent $p_* \geq 3$, the approximation error is already controlled in Assumption 4 based on [BBSS22, Lemma 4.4, 4.5] (using activation function with a ReLU component). If σ_* is a degree-q polynomial, we have the following approximation result using polynomial activation, which follows Lemmas 29 and 30 of [OSSW24a].

Lemma 22. *Suppose that there exist at least $N' = \tilde{\Theta}(N)$ neurons that satisfy $\|\boldsymbol{w}_j^{2T_1} - \boldsymbol{\theta}\| \leq \varepsilon$ and σ is a polynomial link function with degree at least q. Let $b_j \sim \mathrm{Unif}([-C_b, C_b])$ with $C_b = \tilde{O}(1)$, and consider approximation of a ridge function $h(\boldsymbol{\theta}^\top \boldsymbol{x})$ with its degree at most q. Then, there exists $a_1, \ldots, a_N$ such that*

$$\left| \frac{1}{N} \sum_{j=1}^{N} a_j \sigma_j \big(\boldsymbol{w}_j^{2T_1 \top} \boldsymbol{x} + b_j \big) - h(\boldsymbol{\theta}^\top \boldsymbol{x}) \right| = \tilde{O}(N^{-1} + \varepsilon)$$

with high probability, where $(\boldsymbol{x}, y)$ is a random sample, and we omit dependence on the degree q in the big-O notation. Moreover, we have $\sum_{j=1}^{N} a_j^2 = \tilde{O}(N)$.

Lemma 22 can be established from the following result in [OSSW24a].

Lemma 23. *Suppose that $C_b \geq q$. For any polynomial $h(s)$ with its degree at most q, there exists $\bar{v}(b; h)$ with $|\bar{v}(b; h)| \lesssim C_b$ such that for all s,*

$$\mathbb{E}[\bar{v}(b; h)\sigma(\delta s + b)] = h(s).$$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The paper's main contribution is studying the computational complexity of learning single-index models with polynomial link functions by training neural networks with stochastic gradient descent, which is what we claim in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

Justification: We discussed the limitations of our results in Section 4.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

   Justification: All assumptions required in our work are stated in Assumptions 1,2, and 3. All proofs are presented in the appendix.

   Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [NA]

   Justification: This paper is theoretical and experiments are only used for the purpose of illustration in the Introduction.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [NA]

   Justification: This paper is theoretical and experiments are only used for the purpose of illustration in the Introduction.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

Answer: [NA]

Justification: This is a theory paper.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [NA]

Justification: This is a theory paper.

Guidelines:

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

8. **Experiments Compute Resources**

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

Answer: [NA]

Justification: This is a theory paper.

Guidelines:

- The answer NA means that the paper does not include experiments.

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

9. **Code Of Ethics**

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

Answer: [Yes]

Justification: The authors have read and followed the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader Impacts**

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

Answer: [NA]

Justification: This paper carries out a theoretical study, and we believe our work does not have specific societal impacts that require a discussion.

Guidelines:

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

11. **Safeguards**

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

Answer: [NA]

Justification: This paper is theoretical and the experiment is only about two-layer neural network and synthetic data, which we do not have such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [NA]

Justification: This paper is theoretical and does not use any assets.

Guidelines:

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

13. **New Assets**

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

Answer: [NA]

Justification: This paper is theoretical and does not introduce new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

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

Answer: [NA]

Justification: This paper is theoretical and does not include crowdsourcing experiments nor research with human subjects.

Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

Answer: [NA]

Justification: This paper is theoretical and does not include crowdsourcing experiments nor research with human subjects.

Guidelines:

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