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

Jason D. Lee JASONLEE@PRINCETON.EDU
Princeton University

Kazusato Oko OKO-KAZUSATO@G.ECC.U-TOKYO.AC.JP
University of Tokyo and RIKEN AIP

Taiji Suzuki TAIJI@MIST.I.U-TOKYO.AC.JP
University of Tokyo and RIKEN AIP

Denny Wu DENNYWU@NYU.EDU
New York University and Flatiron Institute

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 link function $\sigma_* : \mathbb{R} \to \mathbb{R}$ is an unknown degree q polynomial with 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 statistical 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 learns f_* of arbitrary polynomial link function with a sample and runtime complexity of $n \asymp T \asymp C(q) \cdot d\mathrm{polylog}d$, where constant $C(q)$ only depends on the degree of σ_*, regardless of information exponent; this dimension dependence matches the information theoretic limit up to polylogarithmic factors. 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 [3, 5, 9, 25, 27, 34], 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 with polynomial link 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 $\sigma_* : \mathbb{R} \to \mathbb{R}$ is a degree-q polynomial with information exponent p defined as the index of the first non-zero coefficient in the Hermite expansion (see Definition 1).

(1.1) requires the estimation of the link function σ_* and the relevant direction $\boldsymbol{\theta}$ with d parameters; it is known that learning is information theoretically possible with $n \gtrsim d$ samples [4, 16].

Indeed, when σ_* is polynomial, such complexity can be achieved up to logarithmic factors by a tailored algorithm that exploit the specific structure of the low-dimensional target function [13]. On the other hand, for gradient-based training of two-layer NNs, existing works established a sample complexity of $n \gtrsim d^{\Theta(p)}$ [7, 9, 15], which presents a gap between the information theoretic limit and what is computationally achievable by (S)GD. Such a gap is also predicted by the correlational statistical query (SQ) lower bound [2, 17], which states that for a CSQ algorithm to learn 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 loss), strictly speaking, the CSQ model does not include empirical risk minimization with gradient descent, due to non-adversarial noise and existence of non-correlational terms in the gradient computation. Recently, [19] exploited higher-order terms in the gradient update arising from the reuse of the same minibatch, 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 $O(d)$ batch size. While this presents evidence that GD-trained NNs can learn f_* beyond the CSQ complexity, the weak recovery statement in [19] may not translate to statistical guarantees; moreover, the class of functions where SGD can achieve vanishing generalization error is not fully characterized, as only specific examples of σ_* are discussed.

Given the existence of (non-NN) algorithms that learn any single-index polynomials with $n = \tilde{O}(d)$ samples [13, 16], it is natural to ask if gradient-based training of NNs can achieve similar statistical efficiency for the same function class. Motivated by observations in [19] that SGD with reused batch may break the "curse of information exponent", we aim to address the question:

Can a two-layer NN optimized by SGD with reused batch learn arbitrary polynomial single-index models near the information-theoretic limit $n = \tilde{O}(d)$, regardless of the information exponent?

1.1. Our Contributions

We answer the above question in the affirmative by showing that for (1.1) with arbitrary polynomial link function, SGD training on a natural class of shallow NNs can achieve small generalization error using polynomial compute and $n = \tilde{O}(d)$ training examples, if we employ a layer-wise optimization procedure (analogous to that in [2, 3, 17]) and reuse of the same minibatch:

Theorem *[informal] A shallow NN with $N = \tilde{O}_d(1)$ neurons can learn arbitrary single-index polynomials up to small population loss: $\mathbb{E}_{\boldsymbol{x}}[(f_{\boldsymbol{\Theta}}(\boldsymbol{x}) - f_*(\boldsymbol{x}))^2] = o_{d,\mathbb{P}}(1)$, using $n = \tilde{O}_d(d)$ samples, and an SGD-based algorithm (with reused training data) minimizing the squared loss objective in $T = \tilde{O}_d(d)$ gradient steps.*

- The theorem suggests that NN + SGD with reused batch can match the statistical efficiency of SQ algorithms tailored for low-dimensional polynomial regression [13]. Furthermore, the sample complexity is information theoretically optimal up to polylogarithmic factors, and surpasses the CSQ lower bound for $p > 2$ (see Figure 1); this disproves a conjecture in [2] stating that $n \asymp d^{p/2}$ is the optimal sample complexity for empirical risk minimization with SGD.
- A key observation in our analysis is that 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 to (at most) 2, and hence optimization can escape the high-entropy "equator" at initialization in polylogarithmic time.

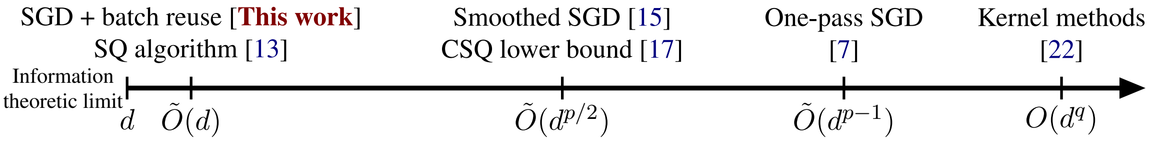

Figure 1: Sample complexity of learning single-index model where the link function σ_* has degree q and information exponent p. For the CSQ lower bound, we translate the tolerance to sample complexity using the i.i.d. concentration heuristic $\tau \approx n^{-1/2}$. We restrict ourselves to algorithms using polynomial compute.

2. Problem Setting and Prior Works

2.1. Complexity of Learning Single-index Models

We aim to learn a single-index polynomial (1.1) where σ_* has information exponent p defined below.

Definition 1 (Information exponent [7]) *Let $\{\mathsf{He}_j\}_{j=0}^{\infty}$ denote the normalized Hermite polynomials. Given $g \in L^2(\gamma)$ and its Hermite expansion $g(z) = \sum_{j=0}^{\infty} \alpha_j \mathsf{He}_j(z)$, the information exponent is defined as* $\mathrm{IE}(g) = p := \min\{j > 0 : \alpha_j \neq 0\}$.

Note that f_* contains $\Theta(d)$ parameters to be estimated, and hence *information theoretically* $n \gtrsim d$ samples are both sufficient and necessary for learning [6, 16, 26]; however the sample complexity achieved by different learning algorithms depends on structure of the link function.

- **Gradient-based Training of NNs.** While NNs can easily approximate a single-index model [4], existing statistical complexity of gradient-based learning scales as $n \gtrsim d^{\Theta(p)}$: in the well-specified setting, [7] proved a complexity of $n = \tilde{\Theta}(d^{p-1})$ for online SGD, which is improved to $\tilde{\Theta}(d^{p/2})$ by smoothing [15]; as for the misspecified setting, [9, 18] showed that $n \gtrsim d^p$ samples suffice. Note that this exponential dependence on p also appears in the CSQ lower bound [1, 17].
- **Statistical Query Learners.** If we do not restrict ourselves to correlational queries, then (1.1) can be efficiently solved near the information-theoretic limit. Specifically, [13] proposed an SQ algorithm that learns any single-index polynomials using $n = \tilde{O}(d)$ samples; the key ingredient is to construct nonlinear transformations to the labels that lowers the information exponent to 2. This is consistent with the recently established SQ lower bound [16].

2.2. Can Gradient Descent Go Beyond Correlational Queries?

Correlational statistical query. A statistical query (SQ) learner [24, 32] 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) [11] 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, we replace the population gradient with the empirical average, and heuristically equate the tolerance τ with the scale of i.i.d. concentration error $n^{-1/2}$.

For Gaussian single-index model with information exponent p, [17] 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. This suggests a sample complexity lower bound $n \gtrsim d^{p/2}$ for poly-time CSQ algorithm, which is conjectured to be optimal for empirical risk minimization with SGD [2].

SGD with reused data. 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, [19] argued that they may arise from two consecutive gradient updates using the same minibatch. For illustrative purposes, consider an example where one neuron $f_{\boldsymbol{w}}(\boldsymbol{x}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$ is updated by two GD steps using the same training example $(\boldsymbol{x}, y)$, starting from zero initialization $\boldsymbol{w}^0 = \mathbf{0}$ (we focus on the correlational term in the loss for simplicity):

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

One can see that in the second gradient step, the label y is transformed by the nonlinearity σ'. Based on this observation, [19] showed that if the non-CSQ term in (2.1) reduces the information exponent to 1, then *weak recovery* can be achieved after two GD steps with $n = O(d)$ samples.

2.3. Challenges in Establishing Statistical Guarantees

Importantly, the analysis in [19] does not lead to concrete learnability guarantees for the class of single-index polynomials, due to the following technical challenges.

SGD decreases information exponent. To show weak recovery, [19, 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 (with batch reuse) is not guaranteed:

- The label transformation employed in prior SQ algorithms [13] is based on the thresholding function, but extracting such transformation from SGD updates on the squared loss is challenging. Instead, we show in Proposition 6 that bounded-degree monomial transformation suffices.

- If the link function σ_* is even, then 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 [19]. To handle this setting, we analyze the SGD update up to $\mathrm{polylog}(d)$ time, at which a nontrivial overlap can be established by a Grönwall-type argument similar to [7].

From weak recovery to sample complexity. Note that weak recovery (i.e., $|\langle \boldsymbol{w}, \boldsymbol{\theta}\rangle| > \varepsilon$ for small constant $\varepsilon > 0$) is 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.

3. SGD Achieves Almost-linear Sample Complexity

We consider the learning of single-index polynomials with degree q and information exponent p; hence the link function admits the Hermite decomposition $\sigma_*(z) = \sum_{j=p}^{q} \alpha_j \mathsf{He}_j(z)$.

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

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

where $\boldsymbol{\Theta} = (\boldsymbol{w}_j, a_j, b_j)_{j=1}^N$ are trainable parameters, and $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function defined as the sum of Hermite polynomials up to degree C_σ given as $\sigma(z) := \sum_{j=0}^{C_\sigma} \beta_j \mathsf{He}_j(z)$, where C_σ only depends on the degree of link function σ_*. Our SGD training procedure is described in Algorithm 1 in the Appendix, 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 [2, 3, 9, 17, 28], 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 parameters $\{a_j\}_{j=1}^N$ to fit the link function σ_*.
- The most important part in Phase I of Algorithm 1 is the reuse of 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).

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 1 *For all $1 \le i \le C_\sigma$ and $k = 0, 1$, we assume that $H\big(\sigma^{(i)}(\sigma^{(1)})^{i-1}; k\big) \neq 0$.*

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

Theorem 2 *Under Assumption 1, for suitable hyperparameters $\eta^t = \tilde{O}(Nd^{-1})$ and $\xi_j^t = 1 - \tilde{O}(1)$, after Phase I of Algorithm 1 is run for $T_{1,1} = d\mathrm{polylog}(d)$ steps, there exists a subset of neurons $\boldsymbol{w}_j^{2T_1} \in \mathcal{W}$ with $|\mathcal{W}| = \tilde{\Theta}(N)$ such that $\big|\langle \boldsymbol{w}_j^{2T_1}, \boldsymbol{\theta}\rangle\big| > c$ for some $c \gtrsim 1/\mathrm{polylog}(d)$.*

The theorem implies that after seeing $n = \tilde{O}(d)$ samples, the parameters escape from the high-entropy equator around initialization, analogous to the information exponent $p = 2$ setting in [7].

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 2 *Recall the Hermite expansions $\sigma_*(z) = \sum_{j=p}^{q} \alpha_j \mathsf{He}_j(z)$, $\sigma(z) = \sum_{j=0}^{C_\sigma} \beta_j \mathsf{He}_j(z)$, we assume the coefficients satisfy $\alpha_j \beta_j \ge 0$ for $p \le j \le q$.*

Lemma 3 *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 some constant r_i, then Assumption 1 and 2 are satisfied in $\exp(-\Theta(q))$-fraction of neurons.*

Note that in our construction of activation for Assumptions 1 and 2, we do not exploit knowledge of the link function σ_* other than its degree q which decides the constant C_σ. The next theorem shows that by running Phase I for $\tilde{O}(d)$ more steps, a subset of neurons can achieve strong recovery.

Theorem 4 *Under Assumptions 1 and 2, given parameter $\boldsymbol{w}_j$ starting from a nontrivial overlap, for suitable hyperparameters $\eta^t = \tilde{O}(Nd^{-1})$ and $\xi_j^t = 1 - \tilde{O}(\varepsilon)$, if we continue to run Phase I of Algorithm 1 for $T_{1,2} = \tilde{O}(d\varepsilon^{-2})$ steps, then $\langle \boldsymbol{w}_j^{2(T_{1,1}+T_{1,2})}, \boldsymbol{\theta}\rangle > 1 - \varepsilon$ with high probability.*

Combining the first two theorems, we know that after $T_1 = 2(T_{1,1} + T_{1,2})$ steps, at least $1/\mathrm{polylog}(d)$ fraction of neurons become ε-close to $\boldsymbol{\theta}$. The following proposition shows that after strong recovery, training the second-layer parameters in Phase II achieves small generalization error.

Proposition 5 *After Phase I terminates, for suitable $\lambda > 0$, the output of Phase II satisfies*

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

with probability 1 as $d \to \infty$, if we set $T_2 = C(q)N^4\mathrm{polylog}(d)\varepsilon^{-2}$, $N = C(q)\mathrm{polylog}(d)\varepsilon^{-1}$ 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

$$n = T_1 + T_2 \asymp C(d\varepsilon^{-2} \vee \varepsilon^{-4}) \cdot \mathrm{polylog}(d), \quad N \asymp C\varepsilon^{-1}\mathrm{polylog}(d),$$

where constant C only depends on the target degree q. 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. Conclusion and Future Directions

In this work we showed that a two-layer neural network (3.1) trained by SGD with reused batch can learn arbitrary single-index polynomials up to ε population error using $n = \tilde{O}(d\varepsilon^{-2})$ samples and compute. Our analysis is based on the observation that by reusing the same minibatch twice in the gradient computation, a non-correlational term arises in the SGD update that transforms the labels (despite the loss function is not modified). Specifically, following the definition in [16], we know that polynomial σ_* has *generative exponent* at most 2, which implies the existence of nonlinear transformation $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ such that the information exponent p_* becomes at most 2, i.e.,

$$\mathbb{E}[\mathcal{T}(y)\mathsf{He}_i(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle)] \neq 0, \quad \text{for } i = 1 \text{ or } 2.$$

We show that restricting $\mathcal{T}$ to be polynomial is sufficient, and such transformation can be extracted by Taylor-expanding the SGD update. Then we show via careful analysis of the trajectory that strong recovery and low population error can be achieved under suitable activation function.

Future directions. First, our analysis only handles link functions with generative exponent $p_* \leq 2$; while this covers arbitrary polynomial σ_* analogous to [13], it is interesting to examine whether SGD with reused batch can learn targets with $p_* \geq 3$ with a sample complexity matching the SQ lower bound. It is also possible that ERM algorithms on i.i.d. data can achieve a statistical complexity beyond the SQ lower bound due to non-adversarial noise [20, 21]; such mechanism is not exploited in our current analysis. Additional interesting directions include extension to multi-index [8, 10, 14, 23], hierarchical polynomials [29], and additive models [31]. Lastly, 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 also achieves the same statistical efficiency.

Acknowledgements

The authors thank Gerard Ben Arous, Joan Bruna, Alex Damian, Marco Mondelli, and Eshaan Nichani for the discussions and feedback on the manuscript. JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994. KO was partially supported by JST ACT-X (JPMJAX23C4). TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2015).

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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. URL https://openreview.net/forum?id=rkllGyBFPH.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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. URL https://openreview.net/forum?id=73XPopmbXH.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Rishabh Dudeja and Daniel Hsu. Statistical query lower bounds for tensor pca. *Journal of Machine Learning Research*, 22(83):1–51, 2021.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

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

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

[34] 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.

Algorithm 1 Gradient-based training of two-layer neural network

```
Input  : Learning rates $\eta^t$, momentum parameters $\xi_j^t$, number of steps $T_1, T_2$, $\ell_2$ regularization $\lambda$.
Initialize $\boldsymbol{w}_j^0 \sim \mathbb{S}^{d-1}(1)$, $a_j \sim \mathrm{Unif}\{\pm r_a\}$.
Phase I: normalized SGD on first-layer parameters
    for $t = 0$ to $T_1$ do
        if $t > 0$ is even then
            Draw i.i.d. sample $(\boldsymbol{x}, y)$.
            Interpolate $\boldsymbol{w}_j^t \leftarrow \boldsymbol{w}_j^t - \xi_j^t(\boldsymbol{w}_j^t - \boldsymbol{w}_j^{t-2})$.
            Normalize $\boldsymbol{w}_j^t \leftarrow \boldsymbol{w}_j^t / \|\boldsymbol{w}_j^t\|$.
        end if
        $\boldsymbol{w}_j^{t+1} \leftarrow \boldsymbol{w}_j^t - \eta^t \tilde{\nabla}_{\boldsymbol{w}} (f_{\boldsymbol{\Theta}}(\boldsymbol{x}) - y)^2, \quad (j = 1, \ldots, N)$.
    end for
Initialize $b_j \sim \mathrm{Unif}([-C_b, C_b])$.
Phase II: SGD on second-layer parameters
    $\hat{\boldsymbol{a}} \leftarrow \mathrm{argmin}_{\boldsymbol{a} \in \mathbb{R}^N} \frac{1}{T_2} \sum_{i=1}^{T_2} (f_{\boldsymbol{\Theta}}(\boldsymbol{x}_i) - y_i)^2 + \lambda \|\boldsymbol{a}\|^2$.
Output: Prediction function $\boldsymbol{x} \mapsto f_{\hat{\boldsymbol{\Theta}}}(\boldsymbol{x})$ with $\hat{\boldsymbol{\Theta}} = (\hat{a}_j, \boldsymbol{w}_j^{T_1}, b_j)_{j=1}^N$.
```

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; we write $\tilde{O}(\cdot)$ when (poly-)logarithmic factors are ignored. $\Omega(\cdot), \Theta(\cdot)$ are defined analogously. γ is the standard Gaussian distribution in $\mathbb{R}$. We denote its L^2-norm of a function f with respect to the data distribution (which will be specified in the sequel) 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. Finally, we say an event A happens *with high probability* when the failure probability is bounded by $\mathrm{poly}(d)e^{-C_H \log d}$, where C_H is a sufficiently large constant and the hidden constants in $\mathrm{poly}(d)$ do not depend on C_H.

Appendix A. Proof Sketch

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

A.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 [13, Section 2.1] and [16, Corollary 4.4]. However, these prior results are based on the thresholding function with specific offset, 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.

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

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

(ii) Let $f^{\mathrm{odd}} : \mathbb{R} \to \mathbb{R}$ be the odd part of f with $\mathbb{E}_{t \sim \mathcal{N}(0, \boldsymbol{I}_d)}[f^{\mathrm{odd}}(t)^2] \ge \rho > 0$. Then there exists some $i \le C_{q,\rho} \in \mathbb{N}_+$ such that $\mathrm{IE}(f^i) = 1$, where constant $C_{q,\rho}$ only depends on q and ρ.

We make the following remarks.

- The proposition implies that for any polynomial link function that is not even, there exists some power $i \in \mathbb{N}_+$ only depending on the degree of σ_* such that raising the function to the i-the power reduces the information exponent to 1. For even link functions, the information exponent after arbitrary transformation is at least 2 (since the transformed function is necessarily even), and this lowest value can also be achieved by suitable 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 label transformations is that they can be obtained from two gradient steps on the training examples, by Taylor expanding the activation function σ' as seen in (2.1). In Section A.2, we build upon this observation to show that Phase I of Algorithm 1 achieves weak recovery using $n \gtrsim d\,\mathrm{polylog}(d)$ samples.

Intuition behind the analysis. Our proof is inspired by [13] 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 [13]. 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.

A.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 $f_{\boldsymbol{w}}(\boldsymbol{x}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$ is updated by two GD steps using the same sample $(\boldsymbol{x}, y)$, starting from $\boldsymbol{w}^0 := \boldsymbol{\omega}$, the alignment with $\boldsymbol{\theta}$ becomes

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

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

We take $\eta \le c_\eta d^{-1}$ with a small constant c_η so that $\eta\|\boldsymbol{x}\|^2 \ll 1$. Crucially, the strength of each term in (A.1) can vary depending on properties of the link function σ_*, which is unknown. Hence a careful analysis is required to ensure that the suitable monomial transformation is always singled out from the gradient update. We therefore divide our analysis into four cases (for simplicity we present the noiseless setting below).

(I) If $\mathrm{IE}(\sigma_*) = 1$**.** All terms in the summation in (A.1) with $i \ge 2$ decay as fast as $\eta\|\boldsymbol{x}\|^2 \ll 1$. On the other hand, the expectation of $y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{\theta}^\top \boldsymbol{x}$ is roughly $\alpha_1\beta_1 \gtrsim 1$. Therefore we may isolate the informative term $y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{\theta}^\top \boldsymbol{x}$. This case is discussed in Section C.3.3.

(II) Else if $\mathrm{IE}((\sigma_*)^I) = 1$ **for some** $2 \le I \le C_\sigma$**.** Let I be the lowest degree of monomial transformation such that $\mathrm{IE}((\sigma_*)^I) = 1$. Since $\sigma_*, \cdots, \sigma_*^{I-1}$ have information exponent larger than 1, expectations of $y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{\theta}^\top \boldsymbol{x}$ and ψ_i $(i = 2, \cdots, I-2)$ scales as $\boldsymbol{\theta}^\top \boldsymbol{\omega} \asymp d^{-\frac{1}{2}}$. For $i = I-1$, because σ_*^I has information exponent 1,

$$\begin{aligned}\mathbb{E}[\psi_{I-1}] =& \mathbb{E}\big[(\eta\|\boldsymbol{x}\|^2)^{I-1} y^I ((I-1)!)^{-1} (\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x}))^{I-1} \sigma^{(I)}(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{\theta}^\top \boldsymbol{x}\big] \\ \asymp& c_\eta^{I-1} H(\sigma_*^I; 1) H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 0).\end{aligned}$$

For $i \geq I$, ψ_i decays as c_η^I, which is smaller than the scale of $\psi_{I-1} \asymp c_\eta^{I-1}$. Hence the term $\psi_{I-1} \gtrsim c_\eta^{I-1}$ is singled out. This case is discussed in Section C.3.2.

(III) **Else if** $\mathrm{IE}(\sigma_*) = 2$**.** We have $\mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{\theta}^\top \boldsymbol{x}] \approx 2\alpha_2\beta_2\boldsymbol{\theta}^\top\boldsymbol{\omega}$. Also, for $i \geq 2$, since $\sigma_*^2, \cdots, \sigma_*^{C_\sigma}$ have information exponent at least 2, expectation of ψ_i is roughly of order $(\eta\|\boldsymbol{x}\|^2)^i \boldsymbol{\theta}^\top \boldsymbol{\omega}$. Therefore, the term $y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{\theta}^\top \boldsymbol{x}$ is singled out, and the expectation scales as $\alpha_2\beta_2 d^{-1/2}$ at initialization. This case is discussed in Section C.3.3.

(IV) **Else** $\mathrm{IE}((\sigma_*)^I) = 2$ **for some** $2 \leq I \leq C_\sigma$**.** In this case, since $\sigma_*, \cdots, \sigma_*^{I-1}$ have information exponent larger than 2, expectations of $y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{\theta}^\top \boldsymbol{x}$ and ψ_i $(i = 2, \cdots, I-2)$ are at most $(\boldsymbol{\theta}^\top\boldsymbol{\omega})^2 \asymp d^{-1}$. And at $i = I-1$, because σ_*^I has information exponent 2,

$$\begin{aligned}\mathbb{E}[\psi_{I-1}] =& \mathbb{E}\big[(\eta\|\boldsymbol{x}\|^2)^{I-1} y^I ((I-1)!)^{-1} (\sigma^{(1)}(\boldsymbol{\omega}^\top \boldsymbol{x}))^{I-1} \sigma^{(I)}(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{\theta}^\top \boldsymbol{x}\big] \\ \asymp& c_\eta^{I-1} H(\sigma_*^I; 2) H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 1)\boldsymbol{\theta}^\top\boldsymbol{\omega}.\end{aligned}$$

As for $i \geq I$, because σ_*^i has information exponent larger than 1 for $I+1 \leq i \leq C_\sigma$, ψ_i decays as $c_\eta^I \boldsymbol{\theta}^\top\boldsymbol{\omega}$, which is smaller than $\psi_{I-1} \asymp c_\eta^{I-1}\boldsymbol{\theta}^\top\boldsymbol{\omega}$. Thus, the term ψ_{I-1} is dominating, whose scale is roughly $c_\eta^{I-1} d^{-1/2}$ at initialization. This case is discussed in Section C.3.1.

Why interpolation is required. In all the cases above, strength of the signal is at least $\eta c_\eta^{I-1} d^{-1/2}$ at initialization. However, this signal strength may not dominate the error coming from discarding the effect of normalization. Usually, given the gradient $-\boldsymbol{g}$ and projection $P_{\boldsymbol{w}} = \boldsymbol{I}_d - \boldsymbol{w}\boldsymbol{w}^\top$, the spherical gradient affects the alignment as $\langle \boldsymbol{\theta}, \boldsymbol{w}^{t+1}\rangle = \big\langle \boldsymbol{\theta}, \frac{\boldsymbol{w}^t + \eta P_{\boldsymbol{w}}\boldsymbol{g}}{\|\boldsymbol{w}^t + \eta P_{\boldsymbol{w}}\boldsymbol{g}\|}\big\rangle \geq \langle \boldsymbol{\theta}, \boldsymbol{w}^t\rangle + \eta\langle \boldsymbol{\theta}, \boldsymbol{g}\rangle - \frac{1}{2}\eta^2\|\boldsymbol{g}\|^2\langle \boldsymbol{\theta}, \boldsymbol{w}^t\rangle + \text{(negligible terms)}$, see [7] or discussion in [15]. 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 the normalization. Thus, taking η sufficiently small, the normalization term shrinks faster than the signal. However, in our setting, the signal shrinks at the rate of c_η^I, and hence taking a smaller step does not improve the signal-to-noise ratio. The interpolation step in our analysis allows us to reduce the effect of normalization without shrinking the signal too fast, by ensuring that $\boldsymbol{w}^{2(t+1)}$ does not move too far from $\boldsymbol{w}^{2t}$.

Combining the four cases yields the following lemma on the evolution of alignment.

Lemma 7 *Under the assumptions per Theorem 2, one of the following holds:*

(i) $\langle \boldsymbol{\theta}, \boldsymbol{w}_j^{2(t+1)}\rangle \geq \langle \boldsymbol{\theta}, \boldsymbol{w}_j^{2t}\rangle + \eta_j^{2t}(1 - \xi_j^{2(t+1)})\gamma + \eta_j^{2t}(1 - \xi_j^{2(t+1)})\nu_j^{2t}$.

(ii) $\langle \boldsymbol{\theta}, \boldsymbol{w}_j^{2(t+1)}\rangle \geq \langle \boldsymbol{\theta}, \boldsymbol{w}_j^{2t}\rangle + \eta_j^{2t}(1 - \xi_j^{2(t+1)})\gamma\boldsymbol{\theta}^\top\boldsymbol{w}_j^{2t} + \eta_j^{2t}(1 - \xi_j^{2(t+1)})\nu_j^{2t}$.

Here $\gamma \gtrsim c_\eta^{I-1}$ *is a constant that depends on* σ_* *and* ν_j^{2t} *is a mean-zero noise.*

For (i), taking expectation immediately yields that weak recovery is achieved within $(\eta(1-\xi)\gamma)^{-1} = \tilde{O}(d)$ steps. For (ii), $\boldsymbol{\theta}^\top\boldsymbol{w}_j^{2t} =: \kappa^t$ 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 $t \lesssim (\eta(1-\xi)\gamma)^{-1} \cdot \log d = \tilde{O}(d)$ steps, similar to the analysis in [7].

A.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. The learnability guarantee follows from standard analysis analogous to that in [1, 3, 17], where we construct a "certificate" second-layer $\boldsymbol{a}^* \in \mathbb{R}^N$ that achieves small loss and small norm:

$$\mathbb{E}_{\boldsymbol{x}}\Big(f_*(\boldsymbol{x}) - \tfrac{1}{N}\textstyle\sum_{j=1}^N a_j^* \sigma({\boldsymbol{w}_j^{T_1}}^\top \boldsymbol{x} + b_j)\Big)^2 \leq \varepsilon^*, \quad \|\boldsymbol{a}^*\| \lesssim r^*,$$

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 [9, 17, 31].

Appendix B. Polynomial Transformation

Proof of Proposition 6. We use a thresholding and compactness argument inspired by [13].

B.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:

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

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

$$\begin{aligned}
\text{(B.1)} &= \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[|f(t)| \geq 2\tau(f)]\left(\frac{f(t)}{2\tau(f)}\right)^i (t^2-1)\right] \\
&\quad + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[\tau(f) \leq |f(t)| < 2\tau(f)]\left(\frac{f(t)}{2\tau(f)}\right)^i (t^2-1)\right] \\
&\quad + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[|f(t)| < \tau(f)]\left(\frac{f(t)}{2\tau(f)}\right)^i (t^2-1)\right] \\
&\geq \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[|f(t)| \geq 2\tau(f)]\left(\frac{2\tau(f)}{2\tau(f)}\right)^i (2^2-1)\right] \\
&\quad + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[\tau(f) \leq |f(t)| < 2\tau(f)]\left(\frac{f(t)}{2\tau(f)}\right)^i (2^2-1)\right] \\
&\quad + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[|f(t)| < \tau(f)]\left(\frac{\tau(f)}{2\tau(f)}\right)^i (0^2-1)\right] \\
&\geq 3\mathbb{P}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}[|f(t)| \geq 2\tau(f)] - 2^{-i}.
\end{aligned}$$

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 $i \in \mathbb{N}$, (B.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:

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

(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 [30] yields that

$$2H(f^i; 2)^2 \le \mathbb{E}_{t \sim \mathcal{N}(0, \boldsymbol{I}_d)}\big[(f(t))^{2i}\big] \le (2i-1)^{iq} \big(\mathbb{E}_{t \sim \mathcal{N}(0, \boldsymbol{I}_d)}\big[f(t)^2\big]\big)^i = (2i-1)^{iq}.$$

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

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

Combining this with the fact that $\mathscr{H}(f) \ge \mathscr{H}_0 > 0$, we know that there exists some $C_q \le 1 + \log_2(\mathscr{H}_0^{-1})$ 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 $i \le C_q$ such that $H(f^i; 2) \ne 0$.

B.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) $g^{\text{odd}}(t) > 0$ for all $t \ge \upsilon(g)$ and $g^{\text{odd}}(t) < 0$ for all $t \le -\upsilon(g)$.

(b) $g^{\text{even}}(t) > |g^{\text{odd}}(t)|$ for all $t \ge \upsilon(g)$ and $t \le -\upsilon(g)$.

(c) For for all $t \ge \upsilon(g)$ and $t \le -\upsilon(g)$, $g(s) = g(t)$ (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 $\tau(g) = \max_{-\upsilon(g) \le t \le \upsilon(g)} |g(t)|$.

Consider the following expectation:

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

(B.3) is decomposed as

$$
\begin{aligned}
\text{(B.3)} = {} & \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[|g(t)| \geq 3\tau(f)]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right] \\
& + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[2\tau(g) \leq |g(t)| < 3\tau(g)]\left(\frac{g(t)}{3\tau(f)}\right)^i t\right] \\
& + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[|g(t)| < 2\tau(g)]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right]. \qquad \text{(B.4)}
\end{aligned}
$$

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), $g(\beta) = g^{\text{even}}(\beta) + g^{\text{odd}}(\beta) = 3\tau(f) > g^{\text{even}}(-\beta) + g^{\text{odd}}(-\beta) = g^{\text{odd}}(-\beta)$. Because $\lim_{t\to-\infty} g^{\text{odd}}(t) = +\infty$, and α is the only solution in $t < 0$, we have $\alpha < -\beta$. Moreover, for all $t > \beta$, we have $g(t) = g^{\text{even}}(t) + g^{\text{odd}}(t) > g^{\text{even}}(-t) + g^{\text{odd}}(-t) = g^{\text{odd}}(-t)$. Combining the above, the first term of (B.4) is bounded as

$$
\begin{aligned}
& \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[|g(t)| \geq 3\tau(f)]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right] \\
& = \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[\beta \leq t \leq -\alpha]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right] + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[t \geq -\alpha]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right] \\
& \quad + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[t \leq \alpha]\left(\frac{g(t)}{3\tau(g)}\right)^i t\right] \\
& = \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[\beta \leq t \leq -\alpha] t\right] + \mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\mathbb{1}[t \geq -\alpha]\left(\left(\frac{g(t)}{3\tau(g)}\right)^i - \left(\frac{g(-t)}{3\tau(g)}\right)^i\right) t\right] \\
& > \beta\mathbb{P}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\beta \leq t \leq -\alpha\right].
\end{aligned}
$$

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

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

Putting things together,

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

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 $\mathbb{E}_{t\sim\mathcal{N}(0,\boldsymbol{I}_d)}[f^{\text{odd}}(t)^2] \geq c$. 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. ■

Appendix C. SGD with Reused Batch

In this section we establish the statistical and computational complexity of Algorithm 1. Recall that the algorithm first trains the first-layer parameters with T_1 steps of SGD update, where we reuse the same sample for two consecutive steps. The analysis of first-layer training is divided into two phases: (i) weak recovery ($\boldsymbol{w}^\top\boldsymbol{\theta}\geq\varepsilon$), and (ii) strong recovery ($\boldsymbol{w}^\top\boldsymbol{\theta}\geq 1-\varepsilon$). We then train the second-layer parameters after strong recovery is achieved.

The section is organized as follows.

- Section C.1 verifies the conditions on the activation function σ to guarantee weak and strong recovery.
- Section C.2 isolates a (nearly) constant fraction of neurons at initialization with an alignment $\boldsymbol{w}^\top\boldsymbol{\theta}$ above a certain threshold. We focus on such neurons in Phase I of first-layer training.
- Section C.3 lower bounds the expected update of alignment $\boldsymbol{w}^\top\boldsymbol{\theta}$ of two gradient steps, and Section C.4 shows that neurons yield weak recovery within $2T_{1,1}=\tilde{O}(d)$ steps.
- Section C.5 discusses how to convert weak recovery to strong recovery using $2T_{1,2}=\tilde{O}(d\varepsilon^{-2})$ SGD steps.
- Finally, Section C.6 analyzes the second-layer training and concludes the proof.

In the following proofs, we introduce constants c_i and C_i, which depends on d at most polylogarithmically. Specifically, the asymptotic strength of the constants is ordered as follows.

$$1\simeq C_1\lesssim c_1^{-1}\lesssim C_2\lesssim C_3\lesssim c_2^{-1}$$

c_η in the main text can be taken as $c_\eta=c_1$, where c_1 should satisfy $\lim_{d\to\infty}c_1=0$, but the convergence can be arbitrarily slow. This requirement comes from the fact that we do not know the exact value of $H(\sigma_*^I;k_*)$, which might be very small. To ensure that the signal is isolated, taking $\eta\asymp c_1d^{-1}$ with arbitrarily slow c_1 suffices. C_2 can also be arbitrarily slow, as long as it satisfies $C_2=\mathrm{poly}(c_1^{-1})$. $C_3=\mathrm{polylog}(d)$ will be used to represent polylogarithmic factor that comes from high probability bounds.

C.1. Conditions on the Activation Function

C.1.1. VERIFYING ASSUMPTION 1

In the following, we focus on the activation function of a single neuron and omit the subscript that distinguishes different neurons. Recall that we consider polynomial activation functions written as

$$\sigma(z) := \sum_{j=0}^{C_\sigma} \beta_j \mathsf{He}_j(z).$$

For weak recovery, we can use any polynomial that has degree $C_\sigma \geq C_q$ as long as the following condition holds: If $\mathrm{IE}(\sigma_*) \geq 2$ and there exists some $i \leq C_\sigma$ such that $\mathrm{IE}(\sigma_*^i) = 1$, σ should satisfy

$$H\left(\frac{1}{(I-1)!}\sigma^{(I)}(\sigma^{(1)})^{I-1}; 0\right) \neq 0. \tag{C.1}$$

If $\mathrm{IE}(\sigma_*) \geq 3$ and there does not exist any $i \leq C_\sigma$ such that $\mathrm{IE}(\sigma_*^i) = 1$ (in this case there exists some $i \leq C_q$ such that $\mathrm{IE}(\sigma_*^i) = 2$), σ should satisfy

$$H\left(\frac{1}{(I-1)!}\sigma^{(I)}(\sigma^{(1)})^{I-1}; 1\right) \neq 0. \tag{C.2}$$

Below we prove Lemma 1 which shows that the above conditions are met with probability 1 for randomly drawn the Hermite coefficients.

Proof of Lemma 1. 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.

Now, the assertion follows from the Schwartz–Zippel Lemma [33], or the fact that zeros of a non-zero polynomial form a measure-zero set. ■

C.1.2. VERIFYING ASSUMPTION 2

On the other hand, for the strong recovery we require an additional condition on the activation function due to link misspecification, which is also introduced in [7, 28]:

$$\sum_{j=p}^{p} j!\alpha_j \beta_j s^j > 0 \quad \text{for all } s > 0.$$

In order to meet Assumption 1 and (2) simultaneously, we follow [31] and randomize the activation function. Specifically, the activation function should satisfy

(I) **If** $\mathrm{IE}(\sigma_*) = 1$**.** We require $\beta_1 > 0$ and $\sum_{j=1}^{q} j!\alpha_j \beta_j s^{j-1} > 0$ for all $s > 0$.

- **(II)** **Else if** $\mathrm{IE}((\sigma_*)^I) = 1$ **for some** $2 \leq I \leq C_\sigma$**.** We require $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 0)$ is not 0 and has the same sign as $H((\sigma_*)^I; 1)$. Also, $\sum_{j=2}^q j!\alpha_j\beta_j s^{j-1} > 0$ for all $s > 0$.
- **(III)** **Else if** $\mathrm{IE}(\sigma_*) = 2$**.** We require $\beta_2 > 0$ and $\sum_{j=2}^q j!\alpha_j\beta_j s^{j-1} > 0$ for all $s > 0$.
- **(IV)** **Else** $\mathrm{IE}((\sigma_*)^I) = 2$ **for some** $2 \leq I \leq C_q$**.** We require $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 1)$ is not 0 and has the same sign as $H((\sigma_*)^I; 2)$. Also, $\sum_{j=3}^q j!\alpha_j\beta_j s^{j-1} > 0$ for all $s > 0$.

Now we prove Lemma 3 which verifies the existence of an activation function that satisfies the assumptions above with non-zero probability. The construction does not depend on the link function itself, but only its degree q.

Proof of Lemma 3. Let c be a sufficiently small constant, and C_σ be the minimum odd integer with $C_\sigma \geq \max\{C_q+1, q+2, 3\}$. With probability $1/2$, we choose the coefficients as $\beta_1 \sim \mathrm{Unif}(\{\pm 1\})$, and $\beta_j \sim \mathrm{Unif}(\{-c, c\})$ for $2 \leq j \leq C_\sigma$. Then, it is easy to see (I) and (III) are met with probability at least 2^{-q}, because they hold when $\mathrm{sign}(\alpha_j) = \mathrm{sign}(\beta_j)$ holds in the summation.

By taking c sufficiently small, we have

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

When I is even, by adjusting the sign of β_1, $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; 0)$ is not 0 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 even I.

In the same vein, we can verify (IV) for even $I \leq C_q \leq C_\sigma - 1$. We have

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

and (IV) can be verified using the same argument.

Otherwise (also with probability $1/2$), we choose the coefficients as $\beta_j \sim \mathrm{Unif}(\{-c, c\})$ for $1 \leq j \leq C_\sigma - 2$ and $\beta_{C_\sigma - 1} = \beta_{C_\sigma} = \pm 1$ to verify (II) and (IV) for odd I. It is easy to see that $\sum_{j=2}^q j!\alpha_j\beta_j s^{j-1} > 0$ for all $s > 0$ holds for (I) and $\sum_{j=3}^q j!\alpha_j\beta_j s^{j-1} > 0$ for all $s > 0$ for (III). In addition, $H((\mathsf{He}_{C_\sigma} + \mathsf{He}_{C_\sigma - 1})^{(I)}((\mathsf{He}_{C_\sigma} + \mathsf{He}_{C_\sigma - 1})^{(1)})^{I-1}; 0) > 0$ and $H((\mathsf{He}_{C_\sigma} + \mathsf{He}_{C_\sigma - 1})^{(I)}((\mathsf{He}_{C_\sigma} + \mathsf{He}_{C_\sigma - 1})^{(1)})^{I-1}; 1) > 0$. Therefore, by taking c sufficiently small, flipping the sign of $\mathsf{He}_{C_\sigma} + \mathsf{He}_{C_\sigma - 1}$ can change the sign of $H(\sigma^{(I)}(\sigma^{(1)})^{I-1}; k)$ for both (II) and (IV) with odd I. Combining all cases yields the desired claim. ■

C.2. Random Initialization

In Section C.3.1 we focus on the neurons with slightly larger initial alignment that satisfy $\kappa_j^0 = \boldsymbol{\theta}^\top \boldsymbol{w}_j^0 \geq 2C_2 d^{-\frac{1}{2}}$ at initialization, where constant C_2 grows at most polylogarithmically in d. The following lemma states that roughly a constant portion of the neurons satisfies this initial alignment condition.

Lemma 8 *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 9 (Theorem 2 of [12]) *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 8. Because $\kappa^0 = v^\top w \overset{\mathrm{d}}{=} \frac{\mathsf{e}_1^\top g}{g}$, where $g \sim \mathcal{N}(0, \boldsymbol{I}_d)$,

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

By letting $\beta = 2$, we have that $\mathbb{P}[\kappa_j^0 \geq C_2 d^{-\frac{1}{2}}] \gtrsim e^{-16C_2^2}$. Due to symmetry, $\mathbb{P}[\kappa_j^0 \leq 2C_2 d^{-\frac{1}{2}}] = \mathbb{P}[\kappa_j^0 \geq 2C_2 d^{-\frac{1}{2}}]$. ■

C.3. Population Update

We first analyze the training of first-layer parameters by evaluating the expected (population) update of two gradient steps with the same training example. At each step, the parameters are updated as

$$\begin{aligned}
\boldsymbol{w}_j^{t+1} &\leftarrow \boldsymbol{w}_j^t - \eta^t \tilde{\nabla}_{\boldsymbol{w}}\big((f_{\boldsymbol{\Theta}}(\boldsymbol{x}) - y)^2\big) \\
&= \boldsymbol{w}_j^t - \eta^t \tilde{\nabla}_{\boldsymbol{w}}\left(\frac{1}{N}\sum_{j=1}^N a_j \sigma({\boldsymbol{w}_j^t}^\top \boldsymbol{x})\right)^2 + 2\eta^t \tilde{\nabla}_{\boldsymbol{w}}\left(y\frac{1}{N}\sum_{j=1}^N a_j \sigma({\boldsymbol{w}_j^t}^\top \boldsymbol{x})\right).
\end{aligned}$$

While the second term scales with $\eta^t a_j^2 = \eta^t c_a^2$, the third term scales with $\eta^t a_j = \eta^t c_a$. Thus, by setting the second-layer scale c_a sufficiently small, we can ignore the interaction of neurons; similar mechanism also appeared in [2, 3]. Specifically, in the following, we show that the strength of the signal is $\boldsymbol{\theta}^\top \boldsymbol{w}_j^t \gtrsim d^{-\frac{1}{2}}$. Thus, by simply letting $c_a \lesssim C_3^{-1} d^{-\frac{1}{2}}$, we can ignore the effect of the squared term. Thus we may focus on the following correlational update:

$$\boldsymbol{w}_j^{t+1} \leftarrow \boldsymbol{w}_j^t + \eta^t \tilde{\nabla}_{\boldsymbol{w}}\left(y\frac{1}{N}\sum_{j=1}^N a_j \sigma({\boldsymbol{w}_j^t}^\top \boldsymbol{x})\right).$$

Due to the absence of interaction between neurons, we omit the subscript j for the index of neurons and ignore the prefactor of N (which can be absorbed into the learning rate); multiplying N to η^t specified below recovers the scaling of η^t presented in the main text.

C.3.1. WHEN $\mathrm{IE}(\sigma_*^I) = 2$ WITH $I \geq 2$

First we consider the most technically difficult case, when $\mathrm{IE}[\sigma_*] \geq 3$ and the information exponent cannot be lowered to 1 for $i \leq C_\sigma$; in this case, from Proposition 6 we know that there exists some $2 \leq i \leq C_\sigma$ such that $\mathrm{IE}[\sigma_*^i] = 2$ and we let I be the first such i.

Without loss of generality, we assume (LHS of (C.2)) > 0; the same result holds for the case of $H(\sigma_*^I; 2) < 0$ except for the opposite sign for the second term in (C.3), by simply setting $\xi = 1 + \tilde{\eta}$ in the following.

Lemma 10 *Starting from $\boldsymbol{w} = \boldsymbol{\omega}$, if we choose step size $\eta = c_a \eta^t = c_1 d^{-1}$ and negative momentum $\xi = 1 - \tilde{\eta}$, and assume that $C_2 d^{-\frac{1}{2}} \leq \kappa = \boldsymbol{\theta}^\top \boldsymbol{\omega} \leq c_2$ and $\tilde{\eta} \leq c_2$, then the expected change in the alignment after two gradient steps on the same sample $(\boldsymbol{x}, y)$ in Algorithm 1 is as follows:*

$$\boldsymbol{\theta}^\top \boldsymbol{w} = \boldsymbol{\theta}^\top \boldsymbol{\omega} + (1 + O(c_1)) \cdot \eta\tilde{\eta} c_1^{I-1} H\Big(\frac{1}{(I-1)!}\sigma^{(I)}(\sigma^{(1)})^{I-1}; 1\Big) H\big((\sigma_*)^I; 2\big)\kappa + \eta\tilde{\eta}\nu,$$

where ν is a mean-zero random variable that satisfies $\mathbb{P}[|\nu| > s] \leq \exp(-s^{1/C_1}/C_1)$ for all $s > 0$.

Proof. We first compute one gradient step from $\boldsymbol{w} = \boldsymbol{\omega}$ with a fresh sample $(\boldsymbol{x}, y)$.

$$\tilde{\nabla}_{\boldsymbol{w}} y\sigma(\boldsymbol{w}^\top \boldsymbol{x}) = y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{x}.$$

Then, with a projection matrix $\boldsymbol{P}_{\boldsymbol{\omega}} = I - \boldsymbol{\omega}\boldsymbol{\omega}^\top$, the updated parameter becomes

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{P}_{\boldsymbol{\omega}} \eta y \sigma'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x} = \boldsymbol{\omega} + \eta y \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}, \tag{C.3}$$

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

$$\begin{aligned}\tilde{\nabla}_{\boldsymbol{w}} y\sigma(\boldsymbol{w}^\top \boldsymbol{x}) &= \eta y \sigma'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x} \\ &= y\sigma'\big((\boldsymbol{\omega} + \eta y \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x})^\top \boldsymbol{x}\big)\boldsymbol{x} \\ &= y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2 \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{x}, \end{aligned} \tag{C.4}$$

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

$$\begin{aligned}&\boldsymbol{w} \leftarrow \boldsymbol{w} + \tilde{\nabla}_{\boldsymbol{w}} y\sigma(\boldsymbol{w}^\top \boldsymbol{x}) \\ &\quad = \boldsymbol{\omega} + \eta y \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x} + \eta y \sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2 \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x} \\ &\quad = \boldsymbol{\omega} + \eta\boldsymbol{g}, \end{aligned}$$

where we defined

$$\boldsymbol{g} := y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x} + y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2 \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}. \tag{C.5}$$

Finally, normalization yields

$$\boldsymbol{w} \leftarrow \frac{\boldsymbol{w} - \xi(\boldsymbol{w} - \boldsymbol{\omega})}{\|\boldsymbol{w} - \xi(\boldsymbol{w} - \boldsymbol{\omega})\|} = \frac{\boldsymbol{\omega} + \eta\tilde{\eta}\boldsymbol{g}}{\|\boldsymbol{\omega} + \eta\tilde{\eta}\boldsymbol{g}\|}.$$

Therefore, the update of the alignment is

$$\begin{aligned}\boldsymbol{\theta}^\top \boldsymbol{w} &= \frac{\kappa + \eta\tilde{\eta}\boldsymbol{\theta}^\top \boldsymbol{g}}{\|\boldsymbol{\omega} + \eta\tilde{\eta}\boldsymbol{g}\|} = \frac{\kappa + \eta\tilde{\eta}\boldsymbol{\theta}^\top \boldsymbol{g}}{(1+\eta^2\tilde{\eta}^2\|\boldsymbol{g}\|^2)^{\frac{1}{2}}} \\ &\geq \kappa + \eta\tilde{\eta}\boldsymbol{\theta}^\top \boldsymbol{g} - \frac{1}{2}\kappa\eta^2\tilde{\eta}^2\|\boldsymbol{g}\|^2 - \frac{1}{2}\eta^3\tilde{\eta}^3|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2. \end{aligned} \tag{C.6}$$

On the other hand, we have

$$\boldsymbol{\theta}^\top \boldsymbol{w} \leq \kappa + \eta\tilde{\eta}\boldsymbol{\theta}^\top \boldsymbol{g} + \frac{1}{2}\kappa\eta^2\tilde{\eta}^2\|\boldsymbol{g}\|^2 + \frac{1}{2}\eta^3\tilde{\eta}^3|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2. \tag{C.7}$$

We evaluate the expectation of (C.6). For the j-th Hermite polynomial He_j and $\boldsymbol{u} \in \mathbb{S}^{d-1}$, we have that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\mathsf{He}_i(\mathsf{e}_1^\top \boldsymbol{x})f(\boldsymbol{u}^\top \boldsymbol{x})\mathsf{e}_1^\top \boldsymbol{x}] \\ &= j(\mathsf{e}_1^\top x)^{j-1}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[f^{(j-1)}(\boldsymbol{u}^\top \boldsymbol{x})] + (\mathsf{e}_1^\top \boldsymbol{x})^{j+1}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[f^{(j+1)}(\boldsymbol{u}^\top \boldsymbol{x})], \\ &\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\mathsf{He}_j(x_1)f(\boldsymbol{u}^\top \boldsymbol{x})\mathsf{e}_2^\top \boldsymbol{x}] = (\mathsf{e}_1^\top \boldsymbol{x})^j(\mathsf{e}_2^\top \boldsymbol{x})\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[f^{(j+1)}(\boldsymbol{u}^\top \boldsymbol{x})].\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\mathsf{He}_j(\mathsf{e}_1^\top \boldsymbol{x})f(\boldsymbol{u}^\top \boldsymbol{x})\boldsymbol{x}] \\ &= \begin{pmatrix} j(\mathsf{e}_1^\top \boldsymbol{x})^{j-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[f^{(j-1)}(\boldsymbol{u}^\top \boldsymbol{x})]\mathsf{e}_1 + (\mathsf{e}_1^\top \boldsymbol{x})^j\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[f^{(j+1)}(\boldsymbol{u}^\top \boldsymbol{x})]\boldsymbol{u}.\end{aligned}$$

Hence the first term of $\boldsymbol{g}$ (C.5) can be exapanded as

$$\begin{aligned}\mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] &= \boldsymbol{P}_{\boldsymbol{\omega}}\mathbb{E}\Bigg[\bigg(\sum_{j=p}^{q}\alpha_j\mathsf{He}_j(\boldsymbol{\theta}^\top \boldsymbol{x})\bigg)\bigg(j\sum_{j=0}^{C_q}\beta_j\mathsf{He}_j\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\bigg)\boldsymbol{x}\Bigg] \\ &= \boldsymbol{P}_{\boldsymbol{\omega}}\sum_{j=p}^{q}\Big[j!\alpha_j\beta_j\big(\boldsymbol{\theta}^\top \boldsymbol{\omega}\big)^{j-1}\boldsymbol{\theta} + (j+2)!\alpha_j\beta_{j+2}\big(\boldsymbol{\theta}^\top \boldsymbol{\omega}\big)^{j}\boldsymbol{\omega}\Big] \\ &= \sum_{j=p}^{q} j!\alpha_j\beta_j\big(\boldsymbol{\theta}^\top \boldsymbol{\omega}\big)^{j-1}\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{\theta}. \end{aligned} \tag{C.8}$$

The coefficient is evaluated as

$$\left|\sum_{j=p}^{q} j!\alpha_j\beta_j\big(\boldsymbol{\theta}^\top \boldsymbol{\omega}\big)^{j-1}\right| \lesssim \kappa^{p-1} \leq \kappa^2. \tag{C.9}$$

For the second term of $\boldsymbol{g}$ (C.5), we first bound the difference in replacing $\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}}$ with d,

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

$$= 2\Big|\mathbb{E}[(\eta\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}} - \eta d)h(\eta\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}}, \boldsymbol{\omega}^\top\boldsymbol{x}, \boldsymbol{\theta}^\top\boldsymbol{x}, \boldsymbol{\theta}^\top\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x})]\Big|, \tag{C.10}$$

where h is a polynomial with degree at most $(C_q+q)^{C_q-1}+q+1$ and coefficients are all $O(1)$. (C.10) is further upper bounded as

$$(\text{C.10}) \le 2\mathbb{E}[(\eta\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}} - \eta d)^2]^{\frac{1}{2}}\mathbb{E}[(h(\eta\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}}, \boldsymbol{\omega}^\top\boldsymbol{x}, \boldsymbol{\theta}^\top\boldsymbol{x}, \boldsymbol{\theta}^\top\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}))^2]^{\frac{1}{2}},$$

by Cauchy-Schwarz inequality. $\mathbb{E}[(\eta\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}} - \eta d)^2]^{\frac{1}{2}} = \eta(2d-1)^{\frac{1}{2}}$, and the expectation of $\mathbb{E}[h^2]$ is $O(1)$ when $\eta \le d^{-1}$. Therefore, (C.10) is bounded by $C_1\eta d^{\frac{1}{2}}$.

Now, we consider $\mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top\boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}]$. The following decomposition can be made.

$$\begin{aligned}
&\mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top\boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}]\\
&= \sum_{j=1}^{C_\sigma} j\beta_j\mathbb{E}\Big[y\mathsf{He}_{j-1}\big(\boldsymbol{\omega}^\top\boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}\Big]\\
&= \sum_{k=0}^{j-1}\sum_{j=1}^{C_\sigma} j\beta_j\binom{j-1}{k}\mathbb{E}\Big[y\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top\boldsymbol{x}\big)\big(\eta d\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})y\big)^k\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}\Big].
\end{aligned} \tag{C.11}$$

We evaluate each term of (C.11) except for $k = I-1$. Each term of (C.11) is a constant multiple of $\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{\theta}$. and we can evaluate the constant by

$$\begin{aligned}
&\left|\boldsymbol{\theta}^\top\mathbb{E}\Big[y\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top\boldsymbol{x}\big)\big(\eta d\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})y\big)^k\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}\Big]\right|\\
&= (\eta d)^k\left|\mathbb{E}\Big[\big(\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})\big)^k\big(\sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x}) + \upsilon\big)^{k+1}\mathsf{He}_{j-k-1}\big(\boldsymbol{\omega}^\top\boldsymbol{x}\big)\boldsymbol{\theta}^\top\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}\Big]\right|.
\end{aligned} \tag{C.12}$$

When $k \le I-2$, $\sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x}), \cdots, \sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x})^{k+1}$ has information exponent larger than 2. Therefore, we have

(C.12)

$$\begin{aligned}
&= (\eta d)^k\left|\sum_{l=0}^{k+1}\binom{k+1}{l}\mathbb{E}[\upsilon^l]\mathbb{E}\Big[\big(\sigma'(\boldsymbol{\omega}^\top\boldsymbol{x})\big)^k\big(\sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x})\big)^{k-l+1}\mathsf{He}_{j-k-1}\big(\boldsymbol{\omega}^\top\boldsymbol{x}\big)\boldsymbol{\theta}^\top\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}\Big]\right|\\
&= (\eta d)^k\left|\sum_{l=0}^{k+1}\binom{k+1}{l}\mathbb{E}[\upsilon^l]\sum_{m=3}^{\infty} m!\kappa^{q(k-l+1)}H\big((\sigma_*)^{k-l+1};m\big)H\big((\sigma')^k\mathsf{He}_{j-k-1};m-1\big)\boldsymbol{\theta}^\top\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{v}\right|
\end{aligned}$$

$$\lesssim (\eta d)^k\kappa^2.$$

When $k \ge I$, we know that $(\sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x}))^{k+1}, \cdots, (\sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x}))^I$ have information exponent larger than 1, and $(\sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x}))^{I-1}, \cdots, (\sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x}))$ have information exponent larger than 2. Thus, the expectation in (C.12) is bounded by

(C.12)

$$= (\eta d)^k \bigg| \sum_{l=0}^{k+1} \binom{k+1}{l} \mathbb{E}[\upsilon^l] \sum_{m=2}^{\infty} m!\kappa^{q(k-l+1)} H\big((\sigma_*)^{k-l+1}; m-1\big) H\big((\sigma')^k \mathsf{He}_{j-k-1}; m\big) \boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{v} \bigg|$$

$$\lesssim (\eta d)^k \kappa.$$

Now, consider the case when $k = I - 1$:

$$\begin{aligned}
&\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1} \mathbb{E}\Big[y\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\big(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}) + \upsilon\big)\big)^I \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big] \\
&= \sum_{l=0}^{I} \sum_{j=1}^{C_\sigma} (\eta d)^k j\beta_j \binom{j-1}{I-1} \binom{I}{l} \mathbb{E}[\upsilon^l] \mathbb{E}\Big[\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^{I-1}\big(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x})\big)^{I-l} \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big].
\end{aligned}$$

Note that $\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}), \cdots, (\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^{I-1}$ has information exponent larger than 2. Therefore, for $l \geq 1$, we have

$$\begin{aligned}
&\bigg|\boldsymbol{\theta}^\top (\eta d)^{I-1} j\beta_j \binom{j-1}{I-1} \binom{I}{l} \mathbb{E}[\upsilon^l] \mathbb{E}\Big[\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^{I-1}\big(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x})\big)^{I-l} \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big]\bigg| \\
&\lesssim (\eta d)^{I-1} \bigg| \mathbb{E}\bigg[\sum_{m=3}^{\infty} m!\kappa^{m-1} H\big(\mathsf{He}_{j-I}(\sigma')^{I-1}; m-1\big) H\big((\sigma_*)^{I-l}; m\big) \boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\bigg]\bigg| \\
&\lesssim (\eta d)^{I-1} \kappa^2.
\end{aligned}$$

And for $l = 0$, $(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^I$ has information exponent 2 and we have

$$\begin{aligned}
&\sum_{j=1}^{C_\sigma} (\eta d)^{I-1} j\beta_j \binom{j-1}{I-1} \mathbb{E}\Big[\mathsf{He}_{j-I}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^{I-1}\big(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x})\big)^{I-l} \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big] \\
&= (\eta d)^{I-1} \sum_{m=2}^{\infty} m!\kappa^{m-1} H\bigg(\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1} (\sigma')^{I-1} \mathsf{He}_{j-I}; m-1\bigg) H\big((\sigma_*)^I; m\big) \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{v}
\end{aligned} \tag{C.13}$$

If $H\bigg(\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1} (\sigma')^{I-1} \mathsf{He}_{j-I}; 1\bigg) \neq 0$, we have

$$\text{(C.13)} = (1 + O(\kappa)) \cdot (\eta d)^{I-1} \underbrace{2H\bigg(\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1} (\sigma')^{I-1} \mathsf{He}_{j-I}; 1\bigg) H\big((\sigma_*)^I; 2\big)}_{=:\gamma} \kappa \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{v}.$$

We have that

$$\begin{aligned}
&\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big) \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}] \\
&= (1 + O(\kappa)) \cdot (\eta d)^{I-1} \gamma\kappa \boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x} + O(\kappa^2 + C_1 \eta d^{\frac{1}{2}} + (\eta d)^{I-2}\kappa^2 + (\eta d)^I \kappa + (\eta d)^{I-1}\kappa^2)
\end{aligned}$$

When $d^{-\frac{1}{2}} C_3 \leq \kappa \leq c_2$ and $\eta = c_1 d^{-1}$, we have

$$\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big) \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}] = (1 + O(c_1)) \cdot c_1^{I-1} \gamma\kappa$$

Together with the bound on the first term ((C.8) and (C.9)), the expectation of $\boldsymbol{g}$ (C.5) is evaluated as

$$\boldsymbol{\theta}^\top \mathbb{E}[\boldsymbol{g}] = (1 + O(c_1)) \cdot c_1^{I-1} \gamma\kappa.$$

By using this, the expected update of the alignment becomes

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}] \geq \kappa + \eta\tilde{\eta}(1 + O(c_1)) \cdot c_1^{I-1}\gamma\kappa - \frac{1}{2}\kappa\eta^2\tilde{\eta}^2\mathbb{E}[\|\boldsymbol{g}\|^2] - \frac{1}{2}\eta^3\tilde{\eta}^3\mathbb{E}[|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2].$$

Note that $\mathbb{E}[\|\boldsymbol{g}\|^2], \mathbb{E}[|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2] \lesssim d$ and $\kappa \lesssim c_1$. Thus,

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}] \geq \kappa + \eta\tilde{\eta}(1 + O(c_1)) \cdot c_1^{I-1}\gamma\kappa - C_1\eta\tilde{\eta}^2(\kappa\eta d + \eta^2\tilde{\eta}d).$$

When $\tilde{\eta} \leq c_2$, $\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}]$ is evaluated as

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}] \geq \kappa + \eta\tilde{\eta}(1 + O(c_1)) \cdot c_1^{I-1}\gamma\kappa.$$

In the same way, using (C.7), we also have the opposite bound:

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}] \leq \kappa + \eta\tilde{\eta}(1 + O(c_1)) \cdot c_1^{I-1}\gamma\kappa$$

Regarding the noise, recall that

$$\boldsymbol{\theta}^\top \boldsymbol{w} = \frac{\kappa + \eta\tilde{\eta}\boldsymbol{\theta}^\top \boldsymbol{g}}{\|\boldsymbol{\omega} + \eta\tilde{\eta}\boldsymbol{g}\|}.$$

$\eta^{-1}\tilde{\eta}^{-1}\big(\kappa + \eta\tilde{\eta}\boldsymbol{\theta}^\top \boldsymbol{g} - \mathbb{E}[\kappa + \eta\tilde{\eta}\boldsymbol{\theta}^\top \boldsymbol{g}]\big) = \boldsymbol{\theta}^\top \boldsymbol{g} - \mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}]$ is a mean-zero polynomial of Gaussian inputs, with all coefficients and variances of inputs bounded by $O(1)$. Notice that normalization does not increase the absolute value of the noise. Thus, regarding $\nu = \eta^{-1}\tilde{\eta}^{-1}\big(\boldsymbol{\theta}^\top \boldsymbol{w} - \mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}]\big)$, we have that

$$\mathbb{P}\big[\big|\eta^{-1}\tilde{\eta}^{-1}\boldsymbol{\theta}^\top \boldsymbol{w} - \mathbb{E}[\eta^{-1}\tilde{\eta}^{-1}\boldsymbol{\theta}^\top \boldsymbol{w}]\big| > t\big] \leq \mathbb{P}\left[\left|\boldsymbol{\theta}^\top \boldsymbol{g} - \mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{g}\right]\right| > t\right] \leq \exp(-t^{1/C_1}/C_1).$$

This completes the proof. ■

C.3.2. WHEN $\mathrm{IE}(\sigma_*^I) = 1$ WITH $I \geq 2$

Next we consider the case when $\mathrm{IE}(\sigma_*) \geq 2$ and there exists some $i \leq C_\sigma$ such that $\mathrm{IE}(\sigma_*^i) = 1$. Let $I (\geq 2)$ be the first such i. Without loss of generality, we assume (LHS of (C.1)) > 0.

Lemma 11 *For the case of* $\mathrm{IE}(\sigma_*^I) = 1$*, starting from* $\boldsymbol{w} = \boldsymbol{\omega}$*, if we choose step size* $\eta = c_a\eta^t = c_1 d^{-1}$ *and negative momentum* $\xi = 1 - \tilde{\eta}$*, and assume that* $d^{-\frac{1}{2}} \leq \kappa = \boldsymbol{\theta}^\top\boldsymbol{\omega} \leq c_2$ *and* $\tilde{\eta} \leq c_2$*, then the expected change in the alignment after two gradient steps on the same sample* $(\boldsymbol{x}, y)$ *in Algorithm 1 is as follows:*

$$\boldsymbol{\theta}^\top \boldsymbol{w} = \boldsymbol{\theta}^\top \boldsymbol{\omega} + (1 + O(c_1)) \cdot \eta\tilde{\eta}c_1^{I-1} H\bigg(\frac{1}{(I-1)!}\sigma^{(I)}(\sigma^{(1)})^{I-1}; 0\bigg) H\big((\sigma_*)^I; 1\big) + \eta\tilde{\eta}\nu,$$

where ν *is a mean-zero random variable that satisfies* $\mathbb{P}[|\nu| > s] \leq \exp(-s^{1/C_1}/C_1)$ *for all* $s > 0$.

Proof. Similarly to Lemma 10, update of the alignment is evaluated as

$$\boldsymbol{\theta}^\top \boldsymbol{w} \geq \kappa + \eta c_2 \boldsymbol{\theta}^\top \boldsymbol{g} - \frac{1}{2}\kappa\eta^2\tilde{\eta}^2\|\boldsymbol{g}\|^2 - \frac{1}{2}\eta^3\tilde{\eta}^3|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2 \tag{C.14}$$

and

$$\boldsymbol{\theta}^\top \boldsymbol{w} \leq \kappa + \eta c_2 \boldsymbol{\theta}^\top \boldsymbol{g} + \frac{1}{2}\kappa\eta^2\tilde{\eta}^2\|\boldsymbol{g}\|^2 + \frac{1}{2}\eta^3\tilde{\eta}^3|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2,$$

where

$$\boldsymbol{g} = y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x} + y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}}\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}. \tag{C.15}$$

From (C.8) and (C.9) we know that the first term of $\boldsymbol{g}$ (C.15) is evaluated as

$$\mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] = \sum_{j=p}^{q} j!\alpha_j\beta_j\big(\boldsymbol{\theta}^\top\boldsymbol{\omega}\big)^{j-1}\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{\theta}, \tag{C.16}$$

where the sum of coefficients are bounded by $\lesssim \kappa$.

We then consider the second term of $\boldsymbol{g}$ (C.15). We can replace $\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}}$ by d with the following bound similarly to (C.10):

$$\big|\boldsymbol{\theta}^\top\mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|^2_{\boldsymbol{P}_{\boldsymbol{\omega}}}\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] - \boldsymbol{\theta}^\top\mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}]\big| \leq C_1\eta d^{\frac{1}{2}}.$$

For the term $\mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}]$. the following decomposition can be made.

$$\begin{aligned}
&\mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] \\
&= \sum_{j=1}^{C_\sigma} j\beta_j\mathbb{E}\Big[y\mathsf{He}_{j-1}\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}\Big] \\
&= \sum_{k=0}^{j-1}\sum_{j=1}^{C_\sigma} j\beta_j\binom{j-1}{k}\mathbb{E}\Big[y\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)^k\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}\Big].
\end{aligned} \tag{C.17}$$

We evaluate each term of (C.17) except for $k = I-1$. Similarly to the bounds on (C.12), each term is a constant multiple of $\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{v}$, and we want to bound

$$\left|\boldsymbol{\theta}\mathbb{E}\Big[y\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)^k\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}\Big]\right|. \tag{C.18}$$

When $k \leq I-2$, $\sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x}), \cdots, \sigma_*(\boldsymbol{\theta}^\top\boldsymbol{x})^{k+1}$ has information exponent larger than 1. Therefore, we have

$$\text{(C.18)} \lesssim (\eta d)^k\kappa.$$

When $k \geq I$, we have

$$\text{(C.18)} \lesssim (\eta d)^k \leq c_1^I.$$

Now we consider the case when $k = I - 1$:

$$\begin{aligned}
&\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1} \mathbb{E}\Big[y \mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\big(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}) + \upsilon\big)\big)^I \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big] \\
&= \sum_{l=0}^{I} \sum_{j=1}^{C_\sigma} (\eta d)^k j\beta_j \binom{j-1}{I-1}\binom{I}{l} \mathbb{E}[\upsilon^l] \mathbb{E}\Big[\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^{I-1}\big(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x})\big)^{I-l} \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big].
\end{aligned}$$

Note that $\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}), \cdots, (\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^{I-1}$ has information exponent larger than 1. Therefore, for $l \geq 1$, we have

$$\begin{aligned}
&\left|\boldsymbol{\theta}^\top (\eta d)^{I-1} j\beta_j \binom{j-1}{I-1}\binom{I}{l} \mathbb{E}[\upsilon^l] \mathbb{E}\Big[\mathsf{He}_{j-1-k}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^{I-1}\big(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x})\big)^{I-l} \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big]\right| \\
&\lesssim (\eta d)^{I-1} \left|\mathbb{E}\Big[\sum_{m=2}^{\infty} m! \kappa^{m-1} H\big(\mathsf{He}_{j-I}(\sigma')^{I-1}; m-1\big) H\big((\sigma_*)^{I-l}; m\big) \boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big]\right| \\
&\lesssim (\eta d)^{I-1} \kappa.
\end{aligned}$$

And for $l = 0$, $(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x}))^I$ has information exponent 1 and we have

$$\begin{aligned}
&\sum_{j=1}^{C_\sigma} (\eta d)^{I-1} j\beta_j \binom{j-1}{I-1} \mathbb{E}\Big[\mathsf{He}_{j-I}\big(\boldsymbol{\omega}^\top \boldsymbol{x}\big)\big(\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\big)^{I-1}\big(\sigma_*(\boldsymbol{\theta}^\top \boldsymbol{x})\big)^{I-l} \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}\Big] \\
&= (\eta d)^{I-1} \sum_{m=1}^{\infty} m! \kappa^{m-1} H\bigg(\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1} (\sigma')^{I-1} \mathsf{He}_{j-I}; m-1\bigg) H\big((\sigma_*)^I; m\big) \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{v} \qquad \text{(C.19)}
\end{aligned}$$

If $H\bigg(\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1}(\sigma')^{I-1}\mathsf{He}_{j-I}; 0\bigg) \neq 0$, we have

$$\text{(C.19)} = (1 + O(\kappa)) \cdot (\eta d)^{I-1} \underbrace{H\bigg(\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1} (\sigma')^{I-1} \mathsf{He}_{j-I}; 0\bigg) H\big((\sigma_*)^I; 1\big)}_{=:\gamma} \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{v}.$$

Note that

$$H\bigg(\sum_{j=1}^{C_\sigma} j\beta_j \binom{j-1}{I-1} (\sigma')^{I-1} \mathsf{He}_{j-I}; 0\bigg) = H\bigg(\frac{1}{(I-1)!} \sigma^{(I)} (\sigma^{(1)})^{I-1}; 0\bigg).$$

Now we have that

$$\begin{aligned}
&\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big) \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}] \\
&= (1 + O(\kappa)) \cdot (\eta d)^{I-1} \gamma \boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x} + O(\kappa + C_1 \eta d^{\frac{1}{2}} + (\eta d)^{I-2} \kappa + (\eta d)^I + (\eta d)^{I-1} \kappa).
\end{aligned}$$

When $d^{-\frac{1}{2}} C_3 \leq \kappa \leq c_2$ and $\eta = c_1 d^{-1}$, we have

$$\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta d\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big) \boldsymbol{P}_{\boldsymbol{\omega}} \boldsymbol{x}] = (1 + O(c_1)) \cdot c_1^{I-1} \gamma$$

Together with the bound on the first term ((C.16)), the expectation of $\boldsymbol{g}$ is evaluated as

$$\boldsymbol{\theta}^\top \mathbb{E}[\boldsymbol{g}] = (1 + O(c_1)) \cdot c_1^{I-1}.$$

By using this and (C.14), the expected update of the alignment becomes

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}] \geq \kappa + \eta\tilde{\eta}(1 + O(c_1)) \cdot c_1^{I-1}\gamma - \frac{1}{2}\kappa\eta^2\tilde{\eta}^2 \mathbb{E}[\|\boldsymbol{g}\|^2] - \frac{1}{2}\eta^3\tilde{\eta}^3 \mathbb{E}[|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2].$$

Note that $\mathbb{E}[\|\boldsymbol{g}\|^2], \mathbb{E}[|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2] \lesssim d$ and $\kappa \lesssim c_1$. Taking $\tilde{\eta} \leq c_2$ yields that

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}] \geq \kappa + \eta\tilde{\eta}(1 + O(c_1)) \cdot c_1^{I-1}\gamma.$$

The upper bound can be obtained in similar fashion,

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{w}] \leq \kappa + \eta\tilde{\eta}(1 + O(c_1)) \cdot c_1^{I-1}\gamma.$$

Finally, the noise can be handled in the exact same way as that of Lemma 10, the details of which we omit. ■

C.3.3. WHEN $\mathrm{IE}(\sigma_*^I) = 1, 2$ WITH $I = 1$

We finally consider the case when $\mathrm{IE}[\sigma_*] = 1$ or when $\mathrm{IE}[\sigma_*] = 2$ and $\mathrm{IE}[\sigma_*^i] \geq 2$ for $i = 2, \cdots, C_\sigma$.

Lemma 12 *Starting from $\boldsymbol{w} = \boldsymbol{\omega}$, if we choose step size $\eta = c_a\eta^t = c_1 d^{-1}$ and negative momentum $\xi = 1 - \tilde{\eta}$, and assume that $d^{-\frac{1}{2}} \leq \kappa = \boldsymbol{\theta}^\top \boldsymbol{\omega} \leq c_2$ and $\tilde{\eta} \leq c_2$, then for $\mathrm{IE}[\sigma_*] = 1$, the expected change in the alignment after two gradient steps on the same sample $(\boldsymbol{x}, y)$ in Algorithm 1 is as follows:*

$$\boldsymbol{\theta}^\top \boldsymbol{w} = \boldsymbol{\theta}^\top \boldsymbol{\omega} + (1 + O(c_1)) \cdot 2\eta\tilde{\eta}\alpha_1\beta_1 + \eta\tilde{\eta}\nu,$$

and when $\mathrm{IE}[\sigma_*] = 2$ and $\mathrm{IE}[\sigma_*^i] \geq 2$ for all $i = 2, \cdots, C_q$,

$$\boldsymbol{\theta}^\top \boldsymbol{w} = \boldsymbol{\theta}^\top \boldsymbol{\omega} + (1 + O(c_1)) \cdot 4\eta\tilde{\eta}\alpha_2\beta_2\kappa + \eta\tilde{\eta}\nu,$$

where ν is a mean-zero random variable that satisfies $\mathbb{P}[|\nu| > s] \leq \exp(-s^{1/C_1}/C_1)$ for all $s > 0$.

Proof. Similar to Lemma 10, the update of the alignment is

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

and

$$\boldsymbol{\theta}^\top \boldsymbol{w} \leq \kappa + \eta\tilde{\eta}\boldsymbol{\theta}^\top \boldsymbol{g} + \frac{1}{2}\kappa\eta^2\tilde{\eta}^2\|\boldsymbol{g}\|^2 + \frac{1}{2}\eta^3\tilde{\eta}^3|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2$$

where

$$\boldsymbol{g} = y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x} + y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2 \sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}.$$

We first consider the case when $\mathrm{IE}[\sigma_*] = 1$. We have

$$\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] = \alpha_1\beta_1\boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{\theta} = \alpha_1\beta_1(1-\kappa^2).$$

Because we take $\eta = c_1 d^{-1}$ with a vanishing constant c_1 and assume that $\kappa \le c_2$, we have

$$\big|\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] - \boldsymbol{\theta}^\top \mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}]\big| \lesssim \eta d\beta_1^2 \lesssim c_1.$$

and

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}] = 2(1+O(c_1)) \cdot \alpha_1\beta_1.$$

The first claim follows from the fact that $\mathbb{E}[\|\boldsymbol{g}\|^2], \mathbb{E}[|\boldsymbol{\theta}^\top \boldsymbol{g}|\|\boldsymbol{g}\|^2] \lesssim d$ and $\eta = c_1 d^{-1}$.

Next we consider the case when $\mathrm{IE}[\sigma_*^i] = 2$ for $i = 1, 2, \cdots, C_q$. We have

$$\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] = 2\alpha_2\beta_2\kappa\boldsymbol{\theta}^\top \boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{\theta} = 2\alpha_2\beta_2(1-\kappa^2)\kappa.$$

On the other hand, because $y, \cdots, y^{C_\sigma}$ has information exponent larger than 1, we have

$$\big|\boldsymbol{\theta}^\top \mathbb{E}[y\sigma'\big(\boldsymbol{\omega}^\top \boldsymbol{x} + \eta\|\boldsymbol{x}\|_{\boldsymbol{P}_{\boldsymbol{\omega}}}^2\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})y\big)\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}] - \boldsymbol{\theta}^\top \mathbb{E}[y\sigma'(\boldsymbol{\omega}^\top \boldsymbol{x})\boldsymbol{P}_{\boldsymbol{\omega}}\boldsymbol{x}]\big| \lesssim \eta d\kappa = c_1\kappa.$$

Thus, when $\kappa \le c_1$,

$$\mathbb{E}[\boldsymbol{\theta}^\top \boldsymbol{g}] = (1+O(c_1)) \cdot 4\alpha_2\beta_2\kappa,$$

which establishes the second claim. ■

C.4. Stochastic Update

As a result of the previous subsection, by choosing either of $\xi_j = 1 - \tilde{\eta}_j$ or $\xi_j = 1 + \tilde{\eta}_j$, we obtained that

(i) When $\mathrm{IE}[\sigma_*^i] = 1$ for some $i \le C_\sigma$, the update can be written as

$$\kappa_j^{2t} + \eta_j^{2t}\tilde{\eta}_j\gamma + \eta_j^{2t}\tilde{\eta}_j\nu_j^{2t} \le \kappa_j^{2(t+1)} \le \kappa_j^{2t} + 2\eta_j^{2t}\tilde{\eta}_j\gamma + \eta_j^{2t}\tilde{\eta}_j\nu_j^{2t}.$$

(ii) Otherwise, the update can be written as

$$\kappa_j^{2t} + \eta_j^{2t}\tilde{\eta}_j\gamma\kappa_j^{2t} + \eta_j^{2t}\tilde{\eta}_j\nu_j^{2t} \le \kappa_j^{2(t+1)} \le \kappa_j^{2t} + 2\eta_j^{2t}\tilde{\eta}_j\gamma\kappa_j^{2t} + \eta_j^{2t}\tilde{\eta}_j\nu_j^{2t}.$$

Here $\eta_j^{2t} = \eta_j^{2t+1} = c_1 d^{-1}$, $\tilde{\eta}_j \le c_2$, and $\gamma \ge c_2$ that depends on σ_*. ν_j^{2t} is a mean-zero random variable that satisfies $\mathbb{P}[|\nu_j^{2t}| > s] \le \exp(-s^{1/C_1}/C_1)$ for all $s > 0$. We assumed that $d^{-\frac{1}{2}} \le \kappa_j^{2t} \le c_2$ for the former and $C_2 d^{-\frac{1}{2}} \le \kappa_j^{2t} \le c_2$ for the latter. For each neuron j, we sample $T_{1,1,j} \sim \mathrm{Unif}(\{1, \cdots, T_{1,1}\})$ and let $\eta^t = c_1 d^{-1}$. For $t = 0, \cdots, 2(T_{1,1,j} - 1)$, we let $\xi_j^t \equiv 1 - \tilde{\eta}_j$ or $\xi_j^t \equiv 1 + \tilde{\eta}_j$ with equiprobability, where $\tilde{\eta}_j = c_2$, and $\xi_j^t = 1$ for $t = 2T_{1,1,j}, \cdots, 2(T_{1,1} - 1)$.

The goal of this subsection is to prove the following lemma.

Lemma 13 *Let $T_1 = \tilde{\Theta}(d)$. If the initial alignment satisfies $\kappa_j^0 \ge C_2 d^{-\frac{1}{2}}$ (for (i)) or $\kappa_j^0 \ge 2C_2 d^{-\frac{1}{2}}$ (for (ii)), then we have $\frac{1}{4}c_2 \le \kappa_j^{2T_{1,1}} \le c_2$ for at least $1/\mathrm{polylog}(d)$ fraction of neurons, with high probability.*

Proof.

(i) the case of $\mathrm{IE} = 1$. If $d^{-\frac{1}{2}} \leq \kappa_j^{2t} \leq c_2$ for all $t = 0, 1, \cdots, \tau$ $(0 \leq \tau \leq T_{1,j} - 1)$, we have

$$\begin{aligned}
\kappa_j^{2(\tau+1)} &\geq \kappa_j^{2\tau} + \eta^\tau \tilde{\eta}_j \gamma + \eta^\tau \tilde{\eta}_j \nu_j^{2\tau} && \text{(C.20)} \\
&= \kappa_j^0 + \sum_{s=0}^{\tau} \eta^s \tilde{\eta}_j \gamma + \sum_{s=0}^{\tau} \eta^s \tilde{\eta}_j \nu_j^{2s} \\
&\geq 2d^{-\frac{1}{2}} + c_1 c_2 d^{-1} \tilde{\varepsilon} \gamma (s+1) - c_1 c_2^2 d^{-1} \left| \sum_{s=0}^{\tau} \nu_j^{2s} \right|. && \text{(C.21)}
\end{aligned}$$

With high probability, $\sum_{s=0}^{\tau} \nu_j^{2s}$ is bounded by $C_3\sqrt{\tau+1}$. If $\tau + 1 \leq 4\gamma^{-2}C_3^2$, by letting $c_2^2 \leq \frac{1}{2}\gamma c_1^{-1} C_3^{-2} d^{\frac{1}{2}}$, we have $c_1 c_2 d^{-1} C_3 \sqrt{\tau+1} \leq d^{-\frac{1}{2}}$. If $\tau + 1 \geq 4\gamma^{-2}C_3^2$, we have $C_3\sqrt{\tau+1} \leq \frac{1}{2}\gamma(\tau+1)$. Thus, in either case,

$$\text{(C.21)} \geq d^{-\frac{1}{2}} + \frac{1}{2} c_1 c_2^2 d^{-1} \gamma (\tau + 1).$$

This verifies that $\kappa_j^{2t} \geq C_1 d^{-\frac{1}{2}}$ holds for $t = \tau + 1$. By induction,

$$\kappa_j^{2t} \geq d^{-\frac{1}{2}} + \frac{1}{2} c_1 c_2^2 d^{-1} \gamma (\tau + 1)$$

holds for t until κ^{2t} gets larger than c_2. By letting $T_{1,1} \geq c_1^{-1} c_2^{-2} \gamma^{-1} d$, we have $\kappa^{2t} \geq \frac{1}{2} c_2$ for some $t \leq T_{1,1}$, with high probability.

Now together with (C.20), we have

$$\kappa_j^{2(t+1)} \leq \kappa_j^{2t} + \eta^t \tilde{\eta}_j \cdot 2\gamma + \eta^t \tilde{\eta}_j \nu_j^{2t}.$$

Hence we obtain the following bound with high probability,

$$|\kappa_j^{2(t+1)} - \kappa_j^{2t}| \leq \eta_j^t \tilde{\eta}_j (2\gamma + C_3) \lesssim c_1 c_2^2 C_3 d^{-1} =: \Delta_1.$$

When $\kappa^{2\tau} \geq \frac{1}{2} c_1$ holds for some τ, with high probability, we have

$$\frac{1}{4} c_2 \leq \kappa^{2\tau + s} \leq c_2$$

for all $0 \leq s \leq \Delta_1 / 4c_2$. Because $1/4\Delta_1 c_2 = \tilde{\Theta}(d)$ and $T_{1,1}$ is also $\tilde{\Theta}(d)$, with probability $\tilde{\Theta}(1)$, $2T_{1,1,j}$ satisfies $2\tau \leq 2T_{1,1,j} \leq 2\tau + 1/4\Delta_1 c_2$. This establishes the first assertion.

(ii) the case of $\mathrm{IE} = 2$: If $C_2 d^{-\frac{1}{2}} \leq \kappa_j^{2t} \leq c_2$ for all $t = 0, 1, \cdots, \tau$ $(0 \leq \tau \leq T_{1,j} - 1)$, we have

$$\begin{aligned}
\kappa_j^{2(\tau+1)} &\geq \kappa_j^{2\tau} + \eta^\tau \tilde{\eta}_j \kappa_j^{2\tau} \gamma + \eta^\tau \tilde{\eta}_j \nu_j^{2s} \\
&= \kappa_j^0 + c_1 c_4 d^{-1} \gamma \sum_{s=0}^{\tau} \kappa_j^{2s} + c_1 c_2 d^{-1} \sum_{s=0}^{\tau} \nu_j^{2s}. && \text{(C.22)}
\end{aligned}$$

With high probability, $\sum_{s=0}^{\tau} \nu_j^{2s}$ is bounded by $C_3\sqrt{\tau+1}$. If $\tau+1 \le 4\gamma^{-2}C_2^{-2}C_3^2 d$, by letting $c_2 \le \frac{1}{2}\gamma c_1^{-1}C_2^2 C_3^{-2}$, we have $c_1 c_2 d^{-1} C_3\sqrt{\tau+1} \le C_2 d^{-\frac{1}{2}}$. If $\tau+1 \ge 4\gamma^{-2}C_2^{-2}C_3^2 d$, we have $C_3\sqrt{\tau+1} \le \frac{1}{2}\gamma(\tau+1)C_2 d^{-\frac{1}{2}}$. Thus, in either case,

$$(\text{C.22}) \ge \kappa_j^0/2 + \frac{1}{2}c_1 c_4 d^{-1}\tilde{\varepsilon}\gamma\sum_{s=0}^{\tau}\kappa_j^{2s}.$$

This verifies that $\kappa_j^{2t} \ge C_1 d^{-\frac{1}{2}}$ holds for $t = \tau+1$. By induction,

$$\kappa^{2t} \ge \kappa_j^0/2 + \frac{1}{2}c_1 c_2 d^{-1}\gamma\sum_{s=0}^{\tau}\kappa_j^{2s}$$

holds for t until κ^{2t} gets larger than c_2. By letting $T_{1,1} \ge \log_{(1+\frac{1}{2}c_1c_2d^{-1})}\frac{c_2}{C_1 d^{-\frac{1}{2}}}$, we have $\kappa^{2t} \ge \frac{1}{2}c_2$ for some $t \le T_{1,1}$, with high probability. Similarly to the case of (i), we can verify that $\frac{1}{4}c_2 \le \kappa^{2t} \le c_2$ for $\tilde{\Theta}(d)$ steps. Therefore, we obtain the second assertion. ■

C.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, $T_{1,1,j}$, and $\tilde{\eta}_j$, neurons achieve small alignment with the target direction $\frac{1}{4}c_2 \le \kappa_j^{2T_{1,1}} \le c_2$. This subsection discusses how to convert the weak recovery into the strong recovery. We focus on the neurons that satisfy $\alpha_j\beta_j \ge 0$ for all j as specified in Assumption 2.

For each neuron j, we let $c_a\eta^t = \eta = c_1 d^{-1}$ if t is even and $\eta^t = 0$ if t is odd, for $t = 2T_{1,1}, \cdots, 2(T_{1,1}+T_{1,2}-1)$. The momentum is defined as $\xi_j^t = 1-\tilde{\eta}$, where $\tilde{\eta} = c_2\varepsilon$.

In the following, we show that the strength of the signal is greater than some constant ε. Thus, for second-layer initialization $c_a \lesssim \varepsilon$, the effect of the interaction term $\nabla_{\boldsymbol{w}}\left(\frac{1}{N}\sum_{j=1}^{N} a_j\sigma(\boldsymbol{w}_j^{t\top}\boldsymbol{x})\right)^2$ can be ignored, and we drop the subscript that distinguishes N neurons.

Lemma 14 *Consider the neuron that satisfies $\frac{1}{4}c_2 \le \kappa_j^{2T_{1,1}} \le c_2$. We have*

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

with high probability, where $T_{1,2} = \Theta O_d(d\varepsilon^{-2})$.

Proof. The expected gradient (of the correlation term) can be computed as

$$\begin{aligned}\mathbb{E}\big[\tilde{\nabla}_{\boldsymbol{w}}y\sigma(\boldsymbol{w}^{2t\top}\boldsymbol{x})\big] &= \mathbb{E}\left[\tilde{\nabla}_{\boldsymbol{w}}\left(\sum_{j=p}^{q}\alpha_j\mathsf{He}_j(\boldsymbol{\theta}^\top\boldsymbol{x})\right)\left(\sum_{j=0}^{C_q}\beta_j\mathsf{He}_j(\boldsymbol{w}^{2t\top}\boldsymbol{x})\right)\right] \\ &= \sum_{j=p}^{q}\big[j!\alpha_j\beta_j(\boldsymbol{\theta}^\top\boldsymbol{w}^{2t})^{j-1}\boldsymbol{\theta} + (j+2)!\alpha_j\beta_{j+2}(\boldsymbol{\theta}^\top\boldsymbol{w}^{2t})^{j}\boldsymbol{w}^{2t}\big].\end{aligned}$$

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

$$\mathbb{E}\big[P_{\boldsymbol{w}^{2t}}\tilde{\nabla}_{\boldsymbol{w}} y\sigma({\boldsymbol{w}^{2t}}^\top \boldsymbol{x})\big] = (\boldsymbol{\theta} - ({\boldsymbol{w}^{2t}}^\top \boldsymbol{\theta})\boldsymbol{w}^{2t}) \sum_{j=p}^{q} j!\alpha_j\beta_j(\boldsymbol{\theta}^\top \boldsymbol{w}^{2t})^{j-1}. \tag{C.23}$$

The update of the alignment is

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

where

$$\boldsymbol{g} = P_{\boldsymbol{w}^{2t}} y c_a \sigma'({\boldsymbol{w}^{2t}}^\top \boldsymbol{x})\boldsymbol{x}. \tag{C.24}$$

From (C.23), the expectation of (C.24) is bounded by

$$\begin{aligned}\mathbb{E}[\kappa^{2(t+1)}] &\geq \kappa^{2t} + \eta\tilde{\eta}(1-(\kappa^{2t})^2)\sum_{j=p}^{q} j!\alpha_j\beta_j(\boldsymbol{\theta}^\top \boldsymbol{w}^{2t})^{j-1} - C_3\eta^2\tilde{\eta}^2 d(\kappa^{2t}+\eta\tilde{\eta}). \\ &\geq \kappa^{2t} + \eta\tilde{\eta}\tilde{\varepsilon}p!\alpha_p\beta_p(\kappa^{2t})^{p-1} - C_3\eta^2\tilde{\eta}^2 d(\kappa^{2t}+\eta\tilde{\eta}).\end{aligned}$$

By letting $\eta \leq c_1 d^{-1}$ and $\tilde{\eta} \leq c_2\varepsilon$, we have

$$\mathbb{E}[\kappa^{2(t+1)}] \geq \kappa^{2t} + \frac{1}{2}\eta\tilde{\eta}\varepsilon p!\alpha_p\beta_p(\kappa^{2t})^{p-1}.$$

Because the noise $\nu^{2t} := \eta^{-1}\tilde{\eta}^{-1}(\kappa^{2(t+1)} - \mathbb{E}[\kappa^{2(t+1)}])$ satisfies $\mathbb{P}[|\nu^{2t}| > s] \leq \exp(-s^{1/C_1}/C_1)$ for all $s > 0$,

$$\kappa^{2(T1,1+t)} \geq \kappa^{2T_{1,1}} + \frac{1}{2}\eta\tilde{\eta}\varepsilon p!\alpha_p\beta_p \sum_{s=T_{1,1}}^{T_{1,1}+t-1} (\kappa^{2s})^{p-1} + \eta\tilde{\eta}\sum_{s=T_{1,1}}^{T_{1,1}+t-1} \nu^{2s},$$

with high probability. The third term is bounded by $\eta\tilde{\eta}C_3 t \leq \frac{1}{8}c_2$ when $t \leq \frac{1}{8}\eta^{-1}\tilde{\eta}^{-1}C_3^{-1}c_2$ and by $\eta\tilde{\eta}C_3\sqrt{t} \leq 4\eta^{\frac{3}{2}}\tilde{\eta}^{\frac{3}{2}}c_2^{-\frac{1}{2}}C_3^{\frac{3}{2}}t \leq \frac{1}{2}\eta\tilde{\eta}\varepsilon p!\alpha_p\beta_p(c_2/8)^{p-1}$ when $t \leq \frac{1}{8}\eta^{-1}\tilde{\eta}^{-1}C_3^{-1}c_2$ and $\varepsilon = \tilde{O}(d^{-1})$.

Therefore, if $\kappa^{2s} \geq \frac{1}{8}c_2$ holds for all $s = T_{1,1}, \cdots, T_{1,1}+t-1$, we have

$$\kappa^{2(T_{1,1}+t)} \geq \frac{1}{8}c_2 + \frac{1}{4}\eta\tilde{\eta}\varepsilon p!\alpha_p\beta_p(c_2/8)^{p-1}t, \tag{C.25}$$

with high probability. Thus, by induction, $\kappa^{2(T1,1+t)} \geq \frac{1}{8}c_2$ holds and (C.25) holds for all t, until we get $\kappa^{2(T1,1+t)} \geq 1-\varepsilon$. Because of (C.25), we have $\kappa^{2(T1,1+t)} \geq 1-\varepsilon$ until $t \leq T_{1,2}$, where $T_{1,2} \geq \left(\frac{1}{4}\eta\tilde{\eta}\varepsilon p!\alpha_p\beta_p(c_2/8)^{p-1}\right)^{-1} = \tilde{O}\big((\eta\tilde{\eta}\varepsilon)^{-1}\big) = \tilde{O}(d\varepsilon^{-2})$.

After we achieve the strong recovery $\kappa^{2(T1,1+t)} \geq 1-\varepsilon$ for some t, $\kappa^{2(T1,1+s)}$ may get smaller than $1-\varepsilon$. However, by letting s' be the first such step, because at each step the alignment only moves at most $\tilde{O}(\eta\tilde{\eta}) = \tilde{O}(d^{-1}\varepsilon)$, s' should still satisfies $\kappa^{2(T1,1+s')} \geq 1-2\varepsilon \geq c_2$. Thus, (C.25) holds again until $\kappa^{2(T1,1+t)} \geq 1-\varepsilon$. Therefore, we can guarantee $\kappa^{2(T1,1+t)} \geq 1-\varepsilon$ after $t \geq T_{1,2}$, with high probability. ■

C.6. Second Layer Training

This subsection proves the generalization error after training the second layer. 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. Here we let $\boldsymbol{a}^* \in \mathbb{R}^N$ be the "certificate" such that $\|\boldsymbol{a}^*\| = \tilde{O}(N)$ that is shown to exist in Lemma 16 (here we suppress dependence on constants p, q). The following lemma is a complete version of Proposition 5.

Lemma 15 *There exists a $4q$-th order polynomial $Q(R_w, b, q')$ of $R_{\boldsymbol{w}} = \max_j \|\boldsymbol{w}_j\|$ and $b = (b_j)_{j=1}^N$ such that, if we set $\lambda = \Theta\Big(\sqrt{\frac{2}{T_2\delta_0}N^2 Q(R_w, b, q')}\Big)$ for some $\delta_0 > 0$, the ridge estimator $\hat{\boldsymbol{a}}$ satisfies*

$$\|f_{\hat{\boldsymbol{a}}} - f_*\|_{L^2}^2 \lesssim (N^{-2} + \varepsilon^2) + \frac{1}{T_2\lambda\delta_0}\big(2N^2 Q(R_w, b, q') + \mathbb{E}_{\boldsymbol{x}}[(f_*)^4]\big) + \frac{3\lambda}{2}\|\boldsymbol{a}^*\|^2, \tag{C.26}$$

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

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

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 $\psi(\boldsymbol{x}) = (\sigma(\langle \boldsymbol{x}, \widehat{\boldsymbol{w}}_j\rangle) + b_j))_{j=1}^N$ so that $f_{\boldsymbol{a}}(\boldsymbol{x}) = \langle \boldsymbol{a}, \psi(\boldsymbol{x})\rangle$.

Part (1). Here, we first bound the second 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 that

$$\begin{aligned} &\|f_{\hat{\boldsymbol{a}}} - f_*\|_{L^2(P_{T_2})}^2 + \lambda\|\hat{\boldsymbol{a}}\|^2 \\ &\le \|f_{\boldsymbol{a}^*} - f_*\|_{L^2(P_{T_2})}^2 + \frac{2}{T_2}\sum_{i=1}^{T_2}(f_{\boldsymbol{a}^*}(\boldsymbol{x}_i) - f_{\hat{\boldsymbol{a}}}(\boldsymbol{x}_i))\varepsilon_i + \lambda\|\boldsymbol{a}^*\|^2. \end{aligned} \tag{C.27}$$

Now, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \frac{2}{T_2}\sum_{i=1}^{T_2}(f_{\boldsymbol{a}^*}(x_i) - f_{\hat{\boldsymbol{a}}}(x_i))\varepsilon_i &= (\boldsymbol{a}^* - \hat{\boldsymbol{a}})^\top \frac{2}{T_2}\sum_{i=1}^{T_2}\psi(\boldsymbol{x}_i)\varepsilon_i \\ &\le 2\|\boldsymbol{a}^* - \hat{\boldsymbol{a}}\|\sqrt{\frac{\sum_{i,j}\varepsilon_i\varepsilon_j\psi(\boldsymbol{x}_i)^\top\psi(\boldsymbol{x}_j)}{T_2^2}}. \end{aligned}$$

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 (C.27), we arrive at

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

Here, by using the evaluation $\|f_{\boldsymbol{a}^*} - f_*\|_{L^2(P_{T_2})} = \tilde{O}(N^{-1} + \varepsilon)$ in Lemma 16, the right hand side can be further bounded by

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

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

$$\begin{aligned}
&\|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_{T_2})} \\
&= \|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_{T_2})} - \|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_x)} + \|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_x)} \\
&= \|f_{\hat{\boldsymbol{a}}}\|^2_{L^2(P_{T_2})} - \|f_{\hat{\boldsymbol{a}}}\|^2_{L^2(P_x)} - 2\left(\frac{1}{T_2}\sum_{i=1}^{T_2} f_{\hat{\boldsymbol{a}}}(\boldsymbol{x}_i)f_*(\boldsymbol{x}_i) - \mathbb{E}[f_{\hat{\boldsymbol{a}}}(\boldsymbol{x}_i)f_*(\boldsymbol{x}_i)]\right) \\
&\quad + \|f_*\|^2_{L^2(P_{T_2})} - \|f_*\|^2_{L^2(P_x)} + \|f_{\hat{\boldsymbol{a}}} - f_*\|^2_{L^2(P_x)}. \qquad \text{(C.28)}
\end{aligned}$$

The first two terms of Eq. (C.28) can be bounded by

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

The standard Rademacher complexity bound yields that

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

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

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

with probabilty $1-\delta_2$.

The third term in Eq. (C.28) can be evaluated as

$$\begin{aligned}
&2\left(\frac{1}{T_2}\sum_{i=1}^{T_2} f_{\hat{\boldsymbol{a}}}(\boldsymbol{x}_i)f_*(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[f_{\hat{\boldsymbol{a}}}(\boldsymbol{x})f_*(\boldsymbol{x})]\right) \\
&= \hat{\boldsymbol{a}}^\top\left(\frac{1}{T_2}\sum_{i=1}^{T_2}(\psi(\boldsymbol{x}_i)f_*(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x})f_*(\boldsymbol{x})])\right)
\end{aligned}$$

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

with probability $1-\delta_3$ where we used Markov's inequality again in the second inequality.

Finally, the fourth and fifth term in Eq. (C.28) can be bounded as

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

with probability $1-\delta_4$ where we used Markov's inequality in the last inequality.

Combining these inequalities, we finally arrive at

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

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

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

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

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

with probability $1 - 4\delta_0$. Thus, since $\|\boldsymbol{a}^*\|^2 = \tilde{O}(N)$, by setting $T_2 = \tilde{\Theta}((N^4 Q(R_w, b, q') + \mathbb{E}[f_*(\boldsymbol{x})^4])\varepsilon^{-2})$, and $N = \tilde{\Theta}(\varepsilon^{-1})$, we obtain that (C.26) $\lesssim \varepsilon$. ■

Finally, we provide the approximation guarantee: If σ is a degree-q polynomial, we have the following result, which follows Lemmas 29 and 30 of [31].

Lemma 16 *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 $a_1, \dots, a_N$ such that*

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

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 $\sum_{j=1}^N a_j^2 = \tilde{O}(N)$.

We rely on the following result.

Lemma 17 *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,*

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

Proof. When $g_q(s) = \sigma(s)$ is a degree-q polynomial,

$$g_q(s) = \int_{b=-q}^{0} \sigma(s + b)\mathrm{d}b$$

is also a degree-q polynomial. Let us repeatedly define

$$g_{q-i}(s) := g_{q-(i-1)}(s + 1) - g_{q-(i-1)}(s) \quad (i = 1, 2, \cdots, q),$$

and let $(c_{i,j})$ be coefficients so that $(s-1)^i = \sum_{j=0}^i c_{i,j} s^j$ holds for all z. Then, by induction, $g_i(s)$ is a degree-i polynomial. Moreover, we have

$$\begin{aligned} g_{q-i}(s) &= \sum_{j=0}^{i} c_{i,j} \int_{b=-q}^{0} \sigma(s + b + j)\mathrm{d}b \\ &= 2C_b \mathbb{E}_{b \sim \mathrm{Unif}([-C_b, C_b])} \left[\left(\sum_{j=0}^{i} c_{i,j} \mathbb{1}[j - q \leq b \leq j] \right) \sigma(s + b) \right], \end{aligned}$$

when $C_b \geq q$. Therefore, 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,

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

■

Proof of Lemma 16. We now discretize Lemma 17. We focus on N' neurons that satisfy $\|\boldsymbol{w}_j^{2T_1} - \boldsymbol{\theta}\| \leq \varepsilon$ (by letting $a_j = 0$ otherwise).

For $A = \tilde\Theta(N') = \tilde\Theta(N)$ (with a small hidden constant), we consider $2A$ intervals $[-C_b, C_b(-1+\frac{1}{A})), [C_b(-1+\frac{1}{A}), C_b(-1+\frac{2}{A})), \cdots, [C_b(1-\frac{1}{A}), C_b]$. By taking the hidden constant sufficiently small, for each interval there exists at least one b_j. Then, for b_j corresponding to $[C_b(-1+\frac{i}{A}), C_b(-1+\frac{i+1}{A}))$, we set $a_j = \frac{N}{2}\int_{C_b(-1+\frac{i}{A})}^{C_b(-1+\frac{i+1}{A}))} \bar{v}(b;h)\mathrm{d}b$. Here we note that $|a_j| = \tilde{O}(1)$ holds for all j. If each interval contains more than one b_j, we ignore all but one by letting $a_j = 0$ except for the one. By doing so, because of Lipschitzness of σ, we have

$$\left|\frac{1}{N}\sum_{j=1}^{N} a_j\sigma(s+b_j) - h(s)\right| = \tilde{O}(N)$$

for all $s = \tilde{O}(1)$. Because $|\boldsymbol{\theta}^\top\boldsymbol{x}^t| = \tilde{O}(1)$ with high probability, we have

$$\left|\frac{1}{N}\sum_{j=1}^{N} a_j\sigma(\boldsymbol{\theta}^\top\boldsymbol{x}+b_j) - h(\boldsymbol{\theta}^\top\boldsymbol{x})\right| = \tilde{O}(N^{-1}) \tag{C.29}$$

with high probability. Finally, because $\|\boldsymbol{w}_j^{2T_1} - \boldsymbol{\theta}\| \leq \varepsilon$, we have

$$\left|\frac{1}{N}\sum_{j=1}^{N} a_j\sigma\big(\boldsymbol{w}_j^{2T_1\top}\boldsymbol{x}+b_j\big) - \frac{1}{N}\sum_{j=1}^{N} a_j\sigma(\boldsymbol{\theta}^\top\boldsymbol{x}+b_j)\right| = \tilde{O}\big((\boldsymbol{w}_j^{2T_1}-\boldsymbol{\theta})^\top\boldsymbol{x}\big) = \tilde{O}(\varepsilon). \tag{C.30}$$

Combining (C.29) and (C.30), we obtain the assertion. ■