Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations

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

Yujin Song　　　　　　　　　　　　　　　　　　YUJIN139@G.ECC.U-TOKYO.AC.JP
University of Tokyo

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

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study the computational and sample complexity of learning a target function $f_* : \mathbb{R}^d \to \mathbb{R}$ with *additive structure*, that is, $f_*(x) = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f_m(\langle x, v_m \rangle)$, where $f_1, f_2, ..., f_M : \mathbb{R} \to \mathbb{R}$ are nonlinear link functions of single-index models (ridge functions) with diverse and near-orthogonal index features $\{v_m\}_{m=1}^{M}$, and the number of additive tasks M grows with the dimensionality $M \asymp d^\gamma$ for $\gamma \geq 0$. This problem setting is motivated by the classical additive model literature, the recent representation learning theory of two-layer neural network, and large-scale pretraining where the model simultaneously acquires a large number of "skills" that are often *localized* in distinct parts of the trained network. We prove that a large subset of polynomial f_* can be efficiently learned by gradient descent training of a two-layer neural network, with a polynomial statistical and computational complexity that depends on the number of tasks M and the *information exponent* of f_m, despite the unknown link function and M growing with the dimensionality. We complement this learnability guarantee with computational hardness result by establishing statistical query (SQ) lower bounds for both the correlational SQ and full SQ algorithms.

1. Introduction

We study the problem of learning a polynomial function $f_* : \mathbb{R}^d \to \mathbb{R}$ on isotropic Gaussian data. Specifically, given pairs of training example $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ is drawn from some distribution P_x and $y_i = f_*(x_i) + \varepsilon_i$, we aim to construct an estimator $\hat{f}$ that achieves small population error: $\mathbb{E}_{x \sim P_x} |f_*(x) - \hat{f}(x)| \leq \varepsilon$. Since the space of degree-q polynomials in $\mathbb{R}^d$ is of dimension $\Theta(d^q)$, without additional structural assumptions on f_*, the sample complexity required to achieve small error should scale as $n \gtrsim d^q$, which is computationally prohibitive for learning high-degree polynomials in high dimensions. Many prior works have therefore imposed the constraint that f_* exhibits certain low-dimensional latent structure [APVZ14; CM20; DLS22], as in the multi-index model: $f_*(x) = g(Vx)$ where $V \in \mathbb{R}^{k \times d}$ for $k = \mathcal{O}_d(1)$. However, such a low-dimensional restriction may rule out many interesting classes of target functions; for instance, when specialized to the learning of two-layer neural network, these prior results only apply to the setting where the network width k is much smaller than the ambient dimensionality d.

In this work, we investigate the efficient learning of f_* under a different kind of structural assumption: we allow the target function to depend on a large number of directions which grows

with d, but we impose an *additive structure*; specifically, we consider f_* to be the sum of $M \asymp d^\gamma$ single-index models (also known as ridge functions) as follows:

$$f_*(x) = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f_m(\langle x, v_m \rangle), \tag{1.1}$$

where $v_1, v_2, ..., v_M \in \mathbb{R}^d$ are the index features, and $f_1, f_2, ..., f_M : \mathbb{R} \to \mathbb{R}$ are the unknown link functions; the $\sqrt{1/M}$ prefactor ensures that $\|f_*\|_{L^2} = \Theta_d(1)$ when the set of directions $\{v_m\}_{m=1}^{M}$ is diverse, i.e., $\langle v_i, v_j \rangle = o_d(1)$ for $i \neq j$. Our setting is motivated by the following lines of literature:

- **Estimation of additive model.** In statistical learning theory, additive model is a classical method employed in high-dimensional nonparametric regression [Sto85; HT87; RLLW09]; especially, when the individual functions take the form of single-index model (ridge function), the estimation of *ridge combinations* has been extensively studied [FS81; P$^+$97; KB16]. While efficient algorithms have been proposed when the basis is given *a priori* [Bac08; RJWY12; SS12], when the index features $\{v_m\}_{m=1}^{M}$ are unknown, most existing approaches involve non-convex optimization, which has been treated as a black box without convergence guarantees [KSP15; AMF$^+$21], or solved with computationally inefficient algorithm [Bac17].

- **Learning two-layer neural network.** An important example in the model class (1.1) is a two-layer neural network with M neurons, and it is natural to ask whether such a network can be efficiently learned via standard gradient-based training. Prior works have shown that in the "narrow width" setting $M = \Theta_d(1)$, gradient descent can learn f_* with polynomial sample complexity depending on the *information exponent* of the target function [AAM22; BES$^+$22; DLS22; BBSS22]. On the other hand, the regime where M, d jointly diverge is not well-understood, and most existing analyses on the complexity of gradient-based training require significant simplification such as quadratic activation [GKZ19; MVEZ20; MBB23].

- **Skill localization & fine-tuning.** Pretrained large neural networks (e.g., language models) can efficiently adapt to diverse downstream tasks by fine-tuning a small set of trainable parameters [DCLT18; LL21; HSW$^+$21]. Recent works have shown that "skills" for each individual task are often *localized* in a subset of neurons [DDH$^+$21; WWZ$^+$22; PSZA23; TLS$^+$23; AG23]. Moreover, high-level features of the input can be expressed by a *linear function* of the neural network's internal representation [MYZ13; AB16; EHO$^+$22]. The additive model (1.1) gives an idealized setting where learning exhibits such skill localization (we interpret each f_m as corresponding to one task) and linear representation property. As we will see, gradient descent training of the first-layer localizes neurons to the target index features, and as a result, the network can linearly represent the downstream tasks on top of the intermediate ("hidden") representation.

Our goal is to characterize the statistical and computational complexity of learning the additive model (1.1), when (i) the number of single-index tasks is large, that is, M grows with the ambient dimensionality d, and (ii), the tasks are *diverse*, i.e., the index features of f_m do not significantly overlap with one another (see Section 2 for precise definition). We ask the following questions:

1. Can we efficiently learn (1.1) via *gradient descent* (GD) training of a neural network?

2. What is the hardness of learning (1.1) as measured by *statistical query* (SQ) lower bounds?
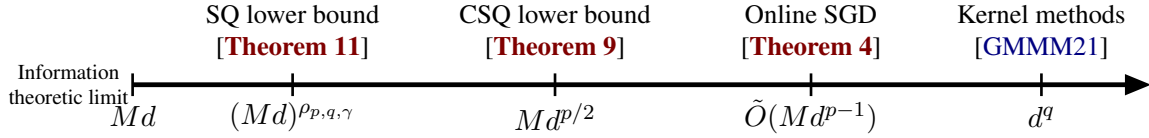
Information theoretic limit	SQ lower bound [**Theorem 11**]	CSQ lower bound [**Theorem 9**]	Online SGD [**Theorem 4**]	Kernel methods [GMMM21]
Md	$(Md)^{\rho_{p,q,\gamma}}$	$Md^{p/2}$	$\tilde{O}(Md^{p-1})$	d^q

Table 1: Complexity upper bound for gradient-based learning and (C)SQ lower bounds for the additive model (1.1), where the single-index tasks have degree q and information exponent p. Our upper bound applies to a subclass of (1.1) specified in Section 2. For the full SQ lower bound we take $d \asymp M^\gamma$, and for any fixed $\gamma > 0$ we may set $p, q > 0$ such that the exponent is arbitrarily large, that is, $\rho_{p,q,\gamma} \overset{p,q \to \infty}{\to} \infty$. We translate the tolerance in the (C)SQ lower bounds to sample complexity using the concentration heuristic $\tau \approx n^{-1/2}$.

1.1. Our Contributions

We address the two questions by providing complexity upper and lower bounds for learning the additive model class (1.1), both of which depend on the number of tasks M, and the *information exponent* $p \in \mathbb{N}$ of the link functions f_m defined as the lowest degree in the Hermite expansion [BAGJ21]. Our findings are summarized as follows (see Table 1 for details).

- In Section 3 we show that a representative subclass of (1.1) can be efficiently learned by gradient-based training (using correlational information) of two-layer neural network, although the number of single-index tasks M is large and the link functions f_m are unknown. Specifically, for $M = \tilde{\Omega}(\sqrt{d})$ tasks with information exponent p, we prove that a layer-wise SGD algorithm (similar to [BBSS22; AAM23]) achieves small population loss in $n = \tilde{\Theta}(Md^{p-1})$ samples. To establish this learning guarantee, we show that neurons *localize* into the task directions during SGD training.

- In Section 4 we establish computational lower bounds for learning (1.1). For *correlational* SQ algorithms, we prove that a tolerance of $\tau^{-2} \gtrsim Md^{p/2}$ is required when link functions f_m have degree q and information exponent $p \le q$. We also provide a full SQ lower bound in the form of $\tau^{-2} \gtrsim (Md)^{\rho_{p,q}}$, where $\rho_{p,q}$ can be made arbitrarily large by varying p and q; under the standard $\tau \approx n^{-1/2}$ heuristic for concentration error, this suggests that prior SQ algorithms that achieve linear sample complexity in the finite-M regime [CM20] cannot attain the same statistical efficiency in our additive model setting with large M.

1.2. Related Works

Gradient-based learning. Recent works have shown that neural network trained by gradient descent can learn useful representation and adapt to low-dimensional target functions such as single-index [BES+22; BBSS22; MHPG+23; BMZ23] and multi-index models [DLS22; AAM22; BBPV23]. In this finite-M setting, the complexity of gradient-based learning is governed by the *information exponent* [BAGJ21] or *leap complexity* [AAM23] of the target function. However, these learning guarantees for low-dimensional f_* yield *superpolynomial* dimension dependence when M grows with d. Another line of works goes beyond the low-dimensional assumption by considering the well-specified setting where f_* and the student network have the same architecture and special activation function [GKZ19; ZGJ21; AS21; VSL+22; MBB23]. The setting we consider (1.1) lies between the two regimes: we allow the width to diverge with dimensionality $M = \omega_d(1)$ but do not assume the nonlinear activation is known, and we show that gradient descent can learn f_* with polynomial sample complexity depending on the information exponent when the target weights are "diverse". We also note that beyond gradient descent training, various SQ algorithms have been introduced to solve related polynomial regression tasks [DH18; CM20; GKS20; DKK+23].

Statistical query lower bound. A statistical query learner [Kea98; Rey20] can access the target function via noisy queries $\tilde{\phi}$ with error tolerance τ: $|\tilde{\phi} - \mathbb{E}_{x,y}[\phi(x,y)]| \leq \tau$. Lower bound on the performance of SQ algorithm is a classical measure of computational hardness. An often-studied subclass of SQ is the *correlational* statistical query (CSQ) [BF02] where the query is restricted to (noisy version of) $\mathbb{E}_{x,y}[\phi(x)y]$. Many existing results on the CSQ complexity assume f_* is low-dimensional ($M = \mathcal{O}_d(1)$), in which case the tolerance scales with $\tau \asymp d^{-\Omega(p)}$, where p is the information exponent or leap complexity of f_* [DLS22; ABA22; AAM23]. On the other hand, [VW19; DKKZ20; GGJ$^+$20] established CSQ lower bounds for learning two-layer ReLU network without structural assumption on the weights, where the error tolerance $\tau \asymp d^{-\Omega(M)}$ implies a *superpolynomial* complexity when $M = \omega_d(1)$. Similar superpolynomial lower bound was shown for three-layer networks in the full SQ model [CGKM22]. Our result connects these two lines of analyses: we show that when the weights $\{v_m\}_{m=1}^M$ are diverse, the (C)SQ complexity is polynomial in M, d, where the dimension dependence is specified by the information exponent of f_m.

2. Problem Setting

Notations. Throughout the analysis, $\|\cdot\|$ denotes the ℓ_2 norm for vectors and the $\ell_2 \to \ell_2$ operator norm for matrices. $O_d(\cdot)$ and $o_d(\cdot)$ stand for the big-O and little-o notations, where the subscript highlights the asymptotic variable d and suppresses dependence on p, q; we write $\tilde{O}(\cdot)$ when (poly-)logarithmic factors are ignored. $\Omega(\cdot), \Theta(\cdot)$ are defined analogously. We say an event A happens with high probability when the failure probability is bounded by $\exp(-C \log d)$ where C is a sufficiently large constant; the high probability events are closed under union over sets of size $\mathrm{poly}(d)$.

Assumptions on target function. We focus on learning a sum of single-index polynomials over Gaussian input, the complexity of which depends on the *information exponent* [DH18; BAGJ21] defined as the smallest degree of non-zero coefficients for the Hermite expansion of the link function.

Definition 1 (Information exponent) *Let $\{\mathsf{He}_j\}_{j=0}^\infty$ be the normalized Hermite polynomials. Given square-integrable $g : \mathbb{R} \to \mathbb{R}$ in its Hermite expansion $g(z) = \sum_{j=0}^\infty \alpha_j \mathsf{He}_j(z)$, the information exponent $\mathrm{IE}(g) := p \in \mathbb{N}_+$ is defined as $p := \min\{j > 0 : \alpha_j \neq 0\}$.*

For example, prior works have shown that online SGD can learn a single-index polynomial with information exponent p over d-dimensional Gaussian input with $\mathcal{O}(d^{p-1})$ samples for $p > 2$ [BAGJ21]. With this definition, we state the class of additive models (1.1) analyzed in this work.

Assumption 1 (Additive model) *We consider the following problem class $\mathcal{F}_{d,M,\varsigma}^{p,q}$:*

$$x \sim \mathcal{N}(0, I_d), \quad y = f_*(x) + \nu, \quad \text{where } f_*(x) = \frac{1}{\sqrt{M}} \sum_{m=1}^M f_m(v_m^\top x), \ \nu \sim \mathcal{N}(0, \varsigma^2),$$

where each f_m is a univariate polynomial with information exponent $p > 2$ and degree q, and we assume proper normalization $\mathbb{E}_{t \sim \mathcal{N}(0,1)}[f_m(t)^2] = 1$, and $\|v_m\| = 1$. We write the Hermite expansion of f_m as $f_m = \sum_{i=p}^q \alpha_{m,i} \mathsf{He}_i$, and define constant $C_p = \left(\frac{\max_m |\alpha_{m,p}|}{\min_{m'} |\alpha_{m',p}|}\right)^{\frac{2}{p-2}}$.

Remark 2
- *We restrict ourselves to $\{f_m\}_{m=1}^M$ with the same information exponent p; this simplifies the optimization dynamics in that all directions are learned at roughly the same rate. To handle heterogeneous tasks with different information exponents, one may consider a student model with mixture of different nonlinearities.*

4

- *We focus on the high information exponent setting $p > 2$, which corresponds to target functions that are more "difficult" to learn via gradient descent. To handle lower information exponent f_m, we may employ a pre-processing procedure analogous to that in [DLS22]: we first fit f_* with a quadratic function, and subtract it from the labels before running Algorithm 1.*

We place the following diversity assumption on the index features $\{v_m\}_{m=1}^M$.

Assumption 2 (Task diversity) *We assume the following diversity condition on $\{v_m\}_{m=1}^M$:*

$$M \leq c_v \max\left\{ \big(\max_{m \neq m'} |v_m^\top v_{m'}| \big)^{-1}, \sqrt{d} \right\}, \quad \text{where } c_v \asymp 1/\text{polylog}(d).$$

Remark 3
- *This condition ensures that the single-index tasks are diverse, in that the index feature directions do not significantly overlap (similar assumption also appeared in [WNL23]). When each v_m is an independent sample from the $(d-1)$-dimensional unit sphere $\mathbb{S}^{d-1}$, via a standard concentration argument, Assumption 2 is satisfied with high probability for $M \asymp d^\gamma, \gamma \in [0, 1/2)$.*
- *The assumption justifies the prefactor of $\frac{1}{\sqrt{M}}$ instead of $\frac{1}{M}$: since v_m are almost orthogonal and the first Hermite coefficient of $f_m(t)$ is zero, $f_m(v_m^\top x)$ are weakly dependent mean-zero variables. Thus the scaling prefactor should be $\frac{1}{\sqrt{M}}$ for the output scaling to be $\Theta(1)$ due to CLT.*

3. Complexity of Gradient-based Training

Neural Network Architecture. In this section we show that a representative subset of our additive model class can be learned via gradient-based training of a neural network. Specifically, we consider the following two-layer network with trainable parameters $\Theta = (a_j, w_j, b_j)_{j=1}^J \in \mathbb{R}^{(1+d+1) \times J}$:

$$f_\Theta = \frac{1}{J} \sum_{j=1}^J a_j \sigma_j(w_j^\top x + b_j). \tag{3.1}$$

For each neuron, we define the Hermite expansion of $a_j \sigma_j(\cdot + b_j)$ as $a_j \sigma_j(\cdot + b_j) = \sum_{i=0}^\infty \beta_{j,i} \mathsf{He}_i(\cdot)$. Note that the Hermite coefficient $\beta_{j,i}$ may differ across neurons. To ensure a descent path from weak recovery to strong recovery, we further require for each task f_i, there exist some neurons j such that

$$\alpha_{m,i}\beta_{j,i} > 0 \text{ for } i = p, \quad \text{and} \quad \alpha_{m,i}\beta_{j,i} \geq 0 \text{ for } p < i \leq q. \tag{3.2}$$

Note that (3.2) is automatically satisfied in the well-specified setting, i.e., the student and teacher models share the same nonlinearity as in [BAGJ21]. In the misspecified setting, such condition has been directly assumed in [MHWSE24]. We defer further discussions to Appendix B.2. Below, we give two concrete settings where (3.2) hold despite the link mismatch.

Assumption 3 (Activation function) *We consider the nonlinearity of the student model (3.1) and the link functions $\{f_i\}_{i=1}^M$ that satisfy one of the following:*
1. *σ_j is a random polynomial activation defined in Appendix B.2.2, and f_i satisfies Assumption 1.*

2. *σ_j is the ReLU activation function, and we additionally require that for each f_i, all the non-zero Hermite coefficients $\alpha_{m,i}$ have the same sign.*

In both settings, we utilize the "diversity" of student nonlinearities to deduce that when the network width J is sufficiently large (quantified in Theorem 4), a subset of neurons can achieve alignment with each target task even though the link functions are unknown.

5

Optimization procedure. The training procedure is presented in Algorithm 1. First, the first-layer parameters are trained to minimize the correlation loss $\mathcal{L} = -y f_\Theta(x)$. We use the correlation loss to ignore the interaction between neurons; analogous strategies appeared in prior analyses of feature learning, for example, by the "one-step" analysis [BES$^+$22; DLS22] or by considering the squared loss with sufficiently small second-layer initialization [AAM23]. Similar to [BAGJ21; DNGL23], we use the *spherical gradient* $\tilde{\nabla}_w f(w) := (I - w w^\top) \nabla_w f(w)$, where ∇_w is the Euclidean gradient.

After T_1 steps, we can show that for each sub-problem there exist some students neurons aligning with the class direction v_m. Then, we train the second-layer parameters (which is a convex problem) with either of L^2 or L^1 regularization. While ridge regression is easier to implement, L^1-regularization gives the better generalization error due to the induced sparsity; more specifically, because not all neurons are aligned with f_i after first-layer training, we need L^1 regularization to efficiently single out neurons that succeeded in aligning with one of the single-index tasks.

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

> **Input** : Learning rates η^t, regularization λ, sample size T_1, T_2, initialization scale C_b.
> **Initialize** $w_j^0 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$, $a_j \sim \mathrm{Unif}\{\pm 1\}$.
> **Phase I: normalized SGD on first-layer parameters**
> > **for** $t = 0$ *to* $T_1 - 1$ **do**
> > > Draw new sample (x^t, y^t).
> > > $w_j^{t+1} \leftarrow w_j^t + \eta^t y^t \tilde{\nabla}_w f_{(a_j, b_j, w_j^t)_{j=1}^J}(x^t)$,
> > > $w_j^{t+1} \leftarrow w_j^{t+1} / \|w_j^{t+1}\|, \quad (j = 1, \ldots, J)$.
> > **end**
> **Initialize** $b_j \sim \mathrm{Unif}([-C_b, C_b])$ *and let* $\hat{w}_j \leftarrow \delta_j w_j^{T_1}$ ($\delta_j \sim \mathrm{Unif}(\{\pm 1\})$)
> **Phase II: Convex optimization for second-layer parameters**
> > Draw new samples $(x^t, y^t)_{t=T_1}^{T_1+T_2-1}$.
> > $\hat{a} \leftarrow \mathrm{argmin}_{a \in \mathbb{R}^J} \frac{1}{T_2} \sum_{t=T_1}^{T_1+T_2-1} \left(f_{(a_j, b_j, \hat{w}_j)_{j=1}^J}(x^t) - y^t \right)^2 + \bar{\lambda} \|a\|_r^r, \quad (r = 1 \text{ or } 2)$.
> **Output** : Prediction function $x \to f_{\hat{\Theta}}(x)$ with $\hat{\Theta} = (\hat{a}_j, \hat{w}_j, b_j)_{j=1}^J$.

Now we present our main theorem for gradient-based training of neural networks.

Theorem 4 *Under Assumptions 1 2, and 3, take the number of neurons $J = \tilde{\Theta}(M^{C_p + \frac{1}{2}} \varepsilon^{-1})$, the number of steps for first-layer training $T_1 = \tilde{\Theta}(M d^{p-1} \vee M d \varepsilon^{-2} \vee M^{\frac{5}{2}} \varepsilon^{-3})$, and the number of steps for second-layer training (i) when $r = 2$ (ridge) $T_2 = \tilde{\Theta}(M^{C_p} \varepsilon^{-2})$, and (ii) when $r = 1$ (LASSO) $T_2 = \tilde{\Theta}(M^{1+\upsilon} \varepsilon^{-2(1+\upsilon)})$ for an arbitrary fixed $\upsilon > 0$. Then, under appropriate choices of η^t and λ, with probability $1 - o_d(1)$, Algorithm 1 outputs $f_{\hat{\Theta}}(x)$ such that*

$$\mathbb{E}_{x \sim \mathcal{N}(0, I_d)} \left[\left| f_*(x) - f_{\hat{\Theta}}(x) \right| \right] \leq \varepsilon.$$

Due to the online SGD update, the runtime complexity $T_1 + T_2$ directly translates to a sample complexity of $n = \tilde{\mathcal{O}}(M d^{p-1})$ (focusing on the term whose exponent depends on the information exponent p). We make the following remarks on the obtained sample complexity.

- **Comparison with prior results.** For single-index model with known link function, [BAGJ21] proved that the online SGD learns the degree-p Hermite polynomial with $\tilde{\mathcal{O}}(d^{p-1})$ samples;

whereas for the *misspecified* setting (unknown f_m), a sample complexity of $n = \tilde{\mathcal{O}}(d^p)$ has been established in [BBSS22]. Our bound with $M = 1$ matches the $\tilde{\mathcal{O}}(d^{p-1})$ sample complexity in [BAGJ21] despite the link misspecification (under the additional restriction that f_m is polynomial).

- **Superiority over kernel methods.** By a dimension argument [KMS20; HSSVG21; AAM22], we know that kernel methods (including neural networks in the lazy regime [JGH18]) require $n \gtrsim d^q$ samples to learn degree-q polynomials in $\mathbb{R}^d$. Since $M < d$ and $p \leq q$, the sample complexity of Algorithm 1 is always better than that achieved by kernels, due to the presence of *feature learning*.

- **The role of additive structure.** Without structural assumptions, learning an M-index model of degree q requires $\Omega(M^q)$ samples. In contrast, in our bound the exponent in the dimension does not depend on q, and the exponent of M is independent of p and q; hence we achieve better complexity when q is large and M diverging with dimensionality. This illustrates the benefit of the additive structure in Assumption 1.

3.1. Outline of Theoretical Analysis

3.1.1. TRAINING OF THE FIRST LAYER

We provide a proof sketch of the first-layer training, where the goal is to show that starting from random initialization, a subset of neurons will achieve alignment with one of the target directions v_m. The following lemma establishes that while the initial correlation with any target direction is small, student neurons may have a constant-factor difference in the magnitude of alignment.

Lemma 5 *Consider the neural network with hyperparameters specified in Theorem 4. Then with high probability, for each m, there exist $J_{\min} = \tilde{\Omega}(M^{\frac{1}{2}}\varepsilon^{-1})$ neurons at initialization satisfying*

$$w_j^\top v_m \geq \max_{m' \neq m} \left| \frac{\beta_{m,p}}{\beta_{m',p}} \right|^{\frac{1}{p-2}} |w_j^\top v_{m'}| + \tilde{\Omega}(d^{-1/2}).$$

Next we show that this small difference in the initial alignment is amplified during SGD update, so that the student neurons will eventually specialize to the target direction v_m with largest overlap at random initialization. Consider the dynamics of one neuron w_j, the update of which is written as

$$w_j^{t+1} = \frac{w_j^t + \eta_j^t y^t a_j \sigma'(w_j^{t\top} x^t + b)(I - w_j^t w_j^{t\top}) x^t}{\|w_j^t + \eta_j^t y^t a_j \sigma'(w_j^{t\top} x^t + b)(I - w_j^t w_j^{t\top}) x^t\|} \approx w_j^t + \eta_j^t \frac{1}{\sqrt{M}} \sum_{m=1}^M \alpha_{j,p} \beta_{m,p} (v_m^\top w_j^t)^{p-1} v_m + Z^t,$$

where Z^t is a mean-zero random variable corresponding to the SGD noise, and the approximation step is due to Hermite expansion in which we ignored the effect of normalization for simplicity. For projection onto a specific target direction v_m, we have the following estimate,

$$v_m^\top w_j^{t+1} \approx v_m^\top w_j^t + \eta_j^t \frac{\alpha_{j,p} \beta_{m,p}}{\sqrt{M}} (v_m^\top w_j^t)^{p-1} + \eta_j^t v_m^\top Z^t.$$

Following [BAGJ21], by using the optimal step size $\eta_j^t = \tilde{\Theta}(M^{-\frac{1}{2}} d^{-\frac{p}{2}})$, the SGD dynamics approximately follows the ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t} v_m^\top w_j^t \approx \eta \frac{\alpha_{j,p} \beta_{m,p}}{\sqrt{M}} (v_m^\top w_j^t)^{p-1} \quad \Rightarrow \quad v_m^\top w_j^t \approx \frac{v_m^\top w_j^0}{\left(1 - \frac{\eta(p-2)\alpha_{j,p}\beta_{m,p}(v_m^\top w_j^0)^{p-2}}{\sqrt{M}} t\right)}.$$

This means that $v_m^\top w_j^t \approx d^{-1/2}$ for a long period, but the SGD dynamics rapidly escapes the high-entropy equator and achieves nontrivial overlap (i.e., weak recovery) at the critical time $t_j := \sqrt{M}\eta^{-1}(p-2)^{-1}\alpha_{j,p}^{-1}\beta_{m,p}^{-1}(v_m^\top w_j^0)^{-(p-2)} = \tilde{O}(Md^{p-1})$, as specified by the following lemma.

Lemma 6 *Consider the SGD dynamics with $\eta_j^t = \tilde{\Theta}(M^{-\frac{1}{2}}d^{-\frac{p}{2}})$. There exists $t_0 = \tilde{O}(Md^{p-1})$ and small constant $c = \tilde{\Theta}(1)$, such that $\langle w_j^{t_0}, v_m \rangle \geq c$, and $|v_m^\top w_j^{t_0}| \leq cM^{-\frac{1}{2}}$ for all $m = 2, \ldots, M$.*

After a nontrivial overlap with v_m is obtained, we continue to run online SGD to amplify the alignment. A technical challenge here is that as the student neuron develop alignment with one direction v_m, the influence from the other $m - 1$ directions is no longer negligible, while the signal from v_m gets smaller due to the projection $(1 - w^t w^{t\top})$; this complication is addressed in Lemma 30.

Finally, in Lemma 33 we prove a local convergence result (analogous to [ZGJ21; AS21]) which entails that for each task direction v_m, there exist some student neurons that that localize to the task, i.e., $\langle v_m, w_j \rangle > 1 - \varepsilon$. Here we exploit the local convexity (more specifically, Łojasiewicz condition) of the loss landscape due to the small overlap between the single-index tasks as specified in Assumption 2. The following proposition describes the configuration of parameters after first-layer training. Here, $\tilde{\varepsilon}$ indicates the desired level of alignment, which is later set to $\tilde{\varepsilon} = \tilde{\Theta}(M^{-\frac{1}{2}}\varepsilon)$, where ε is the desired final generalization error $\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim \varepsilon$.

Lemma 7 (Informal) *Take $T_{1,1} = \tilde{\Theta}(Md^{p-1})$, $T_{1,2} = \tilde{\Theta}(Md^{\frac{p}{2}})$, $T_{1,3} = \tilde{\Theta}(\tilde{\varepsilon}^{-2}Md \vee M\tilde{\varepsilon}^{-3})$, and the number of neurons as $J \gtrsim J_{\min}M^{C_p}\log d$. Suppose that $|v_{m'}^\top v_m| = \tilde{O}(M^{-1})$ for all $m' \neq m$, and $M = \tilde{O}(d^{1/2})$. Then under appropriate learning rate η^t, with high probability, for each class m there exist at least $J_{\min}$ neurons that achieves $v_m^\top w_j^{T_1} \geq 1 - \tilde{\varepsilon}$.*

3.1.2. TRAINING OF THE SECOND LAYER

Once localization to each task is achieved, we optimize the second-layer with L^2- or L^1-regularization. Because the objective is convex, gradient-based optimization can efficiently minimize the empirical loss. The learnability guarantee follows from standard analysis analogous to that in [AAM22; DLS22; BES$^+$22], where we construct a "certificate" $a^* \in \mathbb{R}^J$ that achieves small loss

$$\mathbb{E}_{x \sim \mathcal{N}(0,I_d)}\left(f_*(x) - \tfrac{1}{J}\sum_{j=1}^J a_j^* \sigma(w_j^\top x + b_j)\right)^2 \leq \varepsilon$$

for $r = 1, 2$. Then, the population loss of the regularized empirical risk minimizer can be bounded via standard Rademacher complexity argument, using the norm $\|a^*\|_r^r$. Observe that in Theorem 4, by using L^1-regularization, we can avoid the dependency on C_p determined by the Hermite coefficients of f_i. This is because if some f_i is too large, more students neurons are required to satisfy the initial alignment condition (Lemma 5). Since not all neurons are guaranteed to align with the target directions, the sparsity-inducing L^1-regularization allows us to ignore the redundant neurons and obtain better generalization error rate.

Fine-tuning & linear representation property. L^1-regularization is also useful for efficient fine-tuning on a different downstream task, in which a subset of learned features of size $\tilde{M}$ is used to define the target function: $\tilde{f}_*(x) = \tilde{M}^{-1/2}\sum_{m=1}^{\tilde{M}} g_m(v_m^\top x)$, where $g_1, g_2, ..., g_m$ are link functions that may differ from the training observations. In this setting, retraining the second-layer with L^1-regularization requires $\tilde{O}(\tilde{M}\varepsilon^{-2})$ sample to achieve an L^1-error of ε, which is especially useful when the number of the downstream features is small $\tilde{M} \ll M$. This indeed has connections to
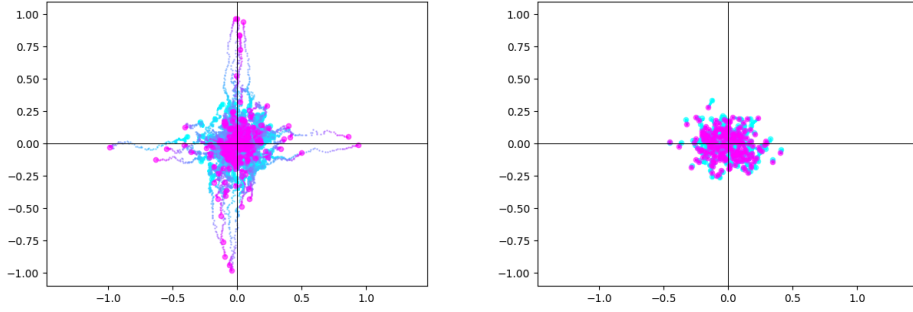
Figure 1: Alignment between neurons and true signals v_1 and v_2, before (blue) and after (purple) SGD training. Left: neural network optimized by online SGD (Algorithm 1), Right: NTK regime.

practical fine-tuning where sparsity is induced to extract relevant features, as seen in localization of skills [DDH$^+$21; WWZ$^+$22; PSZA23] and LoRA [HSW$^+$21].

The existence of second-layer parameters for efficient adaptation also implies a *linear representation property* [MYZ13; AB16; EHO$^+$22; BTB$^+$23], which roughly states that the trained neural network representation can linearly represent a large set of functions. While most theoretical results established this property for representing low-dimensional functions (see [BA24, Appendix A] for discussions), we consider a function class that is not low-dimensional but exhibits additive structure.

Remark 8 *Prior theoretical results on fine-tuning & transfer learning for downstream tasks typically showed that gradient-based pretraining "localizes" the neural network parameters to a low-dimensional subspace spanned by the target functions (tasks), which enables efficient adaptation [DHK$^+$20; DLS22; CHS$^+$23]; for instance, fine-tuning for a degree-q polynomial task in M-dimensional subspace requires $n \gtrsim M^q$ samples [DLS22]. However, when the collection of tasks becomes sufficiently diverse (i.e. M diverges with dimensionality), which is the relevant regime for large-scale pretraining, the benefit of low-dimensional representation diminishes. In contrast, in our setting we prove a different kind of localization, where each neuron aligns with a different v_m depending on the initialization; this leads to efficient fine-tuning despite M being large.*

3.2. Numerical Illustration

We conduct numerical experiments to illustrate the feature learning process for additive models via gradient descent on a two-layer network. The target function is an additive model (1.1), where we set $M = 16$ and $d = 64$, $f_m(x) = \mathsf{He}_3(x)$, and $\{v_m\}_{m=1}^M$ are the canonical basis vectors. The student network is a two-layer ReLU network (3.1), where $J = 8192$ and weights are initialized as $w_j^0 \sim \mathbb{S}^{d-1}$, $a_j^0 \sim \mathrm{Unif}\{\pm 1\}$, and $b_j^0 \sim \mathrm{Unif}([-1, 1])$, respectively. First, we train the first-layer parameters via online SGD for $T = 10^6$ steps. The initial step size is chosen as $\eta^0 = 0.3$, and from $T' = T/2$, we anneal the step size as $\eta^t = \frac{\eta^0}{(t/T')^2}$.

In Figure 1 we plot the alignment between the student neurons and the true signal: the horizontal axis represents the alignment between w_j and v_1, that is, $\langle w_j, v_1 \rangle / \|w_j\|$, and the vertical axis $\langle w_j, v_2 \rangle / \|w_j\|$. For comparison, we also train a network in the NTK regime ([JGH18]), where the magnitude of second layer is $\Theta(1/\sqrt{J})$ and feature learning is suppressed [YH20]. We observe that, via Algorithm 1, a subset of student neurons almost perfectly aligned with one of the true signal directions (Left figure), in contrast to the NTK model (Right figure) where no alignment is observed.

4. Statistical Query Lower Bounds

In Section 3 we showed that two-layer ReLU network optimized via online SGD can learn a representative subset of the additive model class (1.1) with polynomial sample complexity. This being said, due to the introduced diversity condition (Assumption 2), it is not clear if the resulting function class is still intrinsically hard to learn. To understand the fundamental complexity of the learning problem we are addressing, in this section we present several computational lower bounds. Specifically, we consider the statistical query learner [Kea98; BF02], which submits a function of x and y, and receives its expectation within a tolerance τ. The question we ask is how many accurate queries are needed to learn the target function up to certain population L^2-loss.

4.1. Correlational Statistical Query

First, we derive lower bounds for the correlational statistical query (CSQ) learner, which is an important subclass of SQ learners often discussed in the context of feature learning with neural networks [DLS22; AAM23]. For a function $g \colon \mathcal{X} \to \mathbb{R}$ with $\|g\|_{L^2} = 1$ and parameter τ, a correlational SQ oracle $\mathrm{CSQ}(g, \tau)$ returns $\mathbb{E}_{x,y}[yg(x)] + \varepsilon$, where ε is an arbitrary (potentially adversarial) noise that takes value in $\varepsilon \in [-\tau, \tau]$.

The CSQ learner can model gradient-based training as follows: for a neural network f_Θ, the SGD update of the parameters with the minibatch size b using the squared loss is written as

$$\Theta^{t+1} \leftarrow \Theta^t - \frac{2\eta}{b} \sum_{i=1}^{b} y_i^t \nabla_\Theta f_{\Theta^t}(x_i^t) + \frac{2\eta}{b} \sum_{i=1}^{b} f_{\Theta^t}(x_i^t) \nabla_\Theta f_{\Theta^t}(x_i^t).$$

Here, the network gains information of the true function from the correlation $\frac{2\eta}{b} \sum_{i=1}^{b} y_i^t \nabla_\Theta f_{\Theta^t}(x_i^t)$. Roughly speaking, since each coordinate of the gradient concentrates around the expectation with $O(\frac{1}{\sqrt{b}})$ fluctuation, the noisy correlational query can be connected to gradient-based training by matching the tolerance τ with the concentration error $\frac{1}{\sqrt{b}}$. We note that there is a gap between CSQ and SGD update due to the different noise structure, see [AAM23, Remark 6]; nevertheless, a lower bound on CSQ learner serves as a baseline comparison for the sample complexity of SGD learning in many recent works [DLS22; ABA22; DNGL23].

We obtain the following CSQ lower bounds with different dependencies on the target L^2 error and the number of queries. Note that both bounds imply the sample complexity of $\tilde{\Omega}(Md^{\frac{p}{2}})$ under the standard $\tau \approx n^{-1/2}$ concentration heuristic.

Theorem 9 (CSQ lower bound) *For any $p \geq 1$, $\varsigma > 0$ and $C > 0$, there exists a problem class $\mathcal{F} \subset \mathcal{F}_{d,M,\varsigma}^{p,p}$ satisfying Assumptions 1 and 2 such that a CSQ learner using Q queries (outputting $\hat{f}$) requires the following to learn a random choice $f_* \sim \mathcal{F}$:*

(a) tolerance of

$$\tau \lesssim \frac{(\log Qd)^{\frac{p}{4}}}{M^{\frac{1}{2}} d^{\frac{p}{4}}},$$

otherwise, we have $\|f_* - \hat{f}\|_{L^2}^2 \gtrsim 1/M$ with probability at least $1 - O(d^{-C})$.

(b) tolerance of

$$\tau \lesssim \frac{Q^{\frac{1}{2}} (\log dQ)^{\frac{p}{4}+1}}{M^{\frac{1}{2}} d^{\frac{p}{4}}},$$

otherwise, we have $\|f_* - \hat{f}\|_{L^2}^2 \gtrsim 1$ with probability at least $1 - O(d^{-C})$.

Recall that the CSQ lower bound for single-index model translates to a statistical complexity of $n \gtrsim d^{\frac{p}{2}}$; intuitively, our theorem hence implies that for a CSQ algorithm, the difficulty of learning one additive model with M tasks is the same as learning M single-index models separately. To establish this lower bound, we construct a function class $\mathcal{F}$ by choosing $f_m = \mathsf{He}_p$ and randomly sampling the directions v_m from some set $S \in \mathbb{S}^{d-1}$. Here we briefly explain the difference between the two statements, both implying the same dependence on the number of tasks M and ambient dimensionality d, but the target L^2 error and dependence on the number of queries Q differ.

Remark 10 *The target error of the first lower bound is $\Omega(\frac{1}{M})$, and the query dependence is logarithmic; this is obtained by a straightforward extension of the single-index analysis in [DLS22].*

On the other hand, establishing the latter bound with $\Omega(1)$ error, which implies the failure to identify a constant fraction of $\{v_1, \cdots, v_M\}$, is nontrivial. This is because in the additive setting $\frac{1}{\sqrt{M}} \sum_{m=1}^{M} \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_m^\top x)$, the query is not known beforehand, and hence the (adversarial) oracle need to simultaneously prevent the learning of as many directions from $v_1, v_2, ..., v_M$ as possible; whereas in the single-index setting $\frac{1}{\sqrt{p!}} \mathsf{He}_p(v_1^\top x)$, the oracle only need to "hide" one direction. Consequently, we cannot directly connect the identification of $\Omega(1)$-fraction of target directions in the additive setting to the CSQ lower bound for single-index model. To overcome this issue, we introduce a sub-class of CSQ termed noisy correlational statistical query, which adds a random (instead of adversarial) noise to the expected query. For this query model, proving the lower bound for learning single-index models can indeed be translated to our additive setting. However, because the noise is random, the dependency on the number of queries no longer logarithmic but $Q^{\frac{1}{2}}$.

4.2. Full Statistical Query

In this subsection we present a lower bound for full SQ algorithms. For a function $g \colon \mathcal{X} \times \mathcal{Y} \to [-1, 1]$ and parameter τ, a full statistical query oracle $\mathrm{SQ}(g, \tau)$ returns $\mathbb{E}_{x,y}[g(x, y)] + \varepsilon$, where ε is an arbitrary noise that is bounded by the tolerance τ. Note that unlike the CSQ learner, here the query can apply transformations to the target label y and hence reduce the computational complexity.

The full SQ setting is partially motivated by the fact that for sparse polynomials with a constant number of relevant dimensions, a more efficient SQ algorithm that departs from gradient-based training is available: [CM20] showed that general multi-index polynomials of degree q can be learned with $\tilde{O}(d)$ samples. Core to their analysis is a transformation of the label y that reduces the leap complexity [AAM23] to at most 2 (similar transformations also appeared in phase retrieval [MM18; BKM$^+$19]); this enables a warm-start from the direction that overlaps with the relevant dimensions of f_*, from which point projected gradient descent can efficiently learn the target.

One might therefore wonder if the polynomial dimension dependence in Theorem 9 is merely a result of restriction to correlational queries. However, we show that for general SQ algorithms, (1.1) is also challenging to learn despite the additive structure and near-orthogonality constraint. We consider the scaling where M increases with d, i.e., $M \asymp d^\gamma$ with some $\gamma > 0$, and prove that the dimension dependence in the SQ lower bound can be arbitrarily large with $p, q = O_d(1)$.

Theorem 11 (SQ lower bound) *Fix $0 < \gamma \le p$ and $\varsigma > 0$ arbitrarily, and consider the number of tasks $M = \Theta(d^\gamma)$. For any $\rho > 0$, there exist constants $p, q = O_d(1)$ depending only on γ and ρ, and a problem class $\mathcal{F} \subset \mathcal{F}_{d,M,\varsigma}^{p,q}$ satisfying Assumptions 1 and 2 for which an SQ learner outputting $\hat{f}$ from Q queries requires the following tolerance to learn a random choice of $f_* \sim \mathcal{F}$,*

$$\tau \lesssim d^{-\rho},$$

otherwise, we have $\|f_* - \hat{f}\|_{L^2}^2 \gtrsim 1$ *with probability at least* $1 - Qe^{-\Omega(\sqrt{d})}$.

This lower bound illustrates the intrinsic difficulty of our large-M setting compared to the well-studied low-dimensional polynomial regression. For the latter, gradient-based training of polynomial-width network can achieve low loss using $n = \Omega(d^{p-1})$ samples [AAM23], and the CSQ lower bound gives a $\Omega(d^{p/2})$ complexity [DLS22]; however, by applying nonlinear transformation to the labels, an efficient SQ algorithm can achieve $\tilde{O}(d)$ complexity [CM20]. In contrast, while our additive model has $O(Md)$ total parameters to be estimated, the dimension dependence in the SQ lower bound can be arbitrarily larger than M times the single-index complexity. Therefore, unlike the CSQ lower bound (Theorem 9), Theorem 11 suggests that for an SQ algorithm, learning one additive model with M tasks is more difficult than learning M single-index models separately. See Section 4.2.2 for more discussions.

Remark 12 *A concurrent work [DPVLB24] established an SQ lower bound for single-index model that implies a sample complexity of* $n \gtrsim d^{k^*/2}$*, where* k^* *is a prescribed generative exponent that can be made arbitrarily large. While this also entails the existence of link functions not learnable by SQ algorithms with* $\tilde{O}(d)$ *samples, the source of computational hardness is fundamentally different than our setting: [DPVLB24] constructed "hard" non-polynomial link functions that preserve high information exponent after arbitrary transformations; in contrast, we consider polynomial link functions which have generative exponent* $k^* \leq 2$*, so the lower bound relies on the additive structure with a large number of tasks* $M = \Omega(1)$.

4.2.1. Outline of Theoretical Analysis

In the proof of the CSQ lower bounds, Hermite polynomials were used to construct the "worst-case" target functions [DLS22]. This is because higher-order Hermite polynomials have small L^2-correlation under Gaussian input, and hence correlational queries are ineffective. For the SQ learner, which goes beyond correlational queries, a hard function class should also "hide" information from nonlinear transformations to the target labels. As we will see, for our additive model, the class of target functions should maintain orthogonality after polynomial transformations. We refer to this property as *superorthogonality*, and the following proposition shows the existence of such functions.

Proposition 13 (Superorthogonal polynomials) *For any K and L, there exists a polynomial f :* $\mathbb{R} \to \mathbb{R}$ *that is not identically zero and satisfies the following:*

$$\int (f(x))^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x = 0,$$

for every $1 \leq k \leq K$ and $1 \leq l \leq L$.

Because the proof is not constructive, it is difficult to determine the specific form of f for general K and K. Below, we present specific examples for small K, L:

Example 1 *Examples of link function f in Proposition 13 are given as follows.*

(i) *For $K = 1$ and $L \in \mathbb{N}$, $f(x) = \mathsf{He}_{L+1}(x)$.*

(ii) *For $K = L = 2$, $f(x) = \mathsf{He}_4(x) - \frac{4}{15}\mathsf{He}_6(x) + \frac{11}{280}\mathsf{He}_8(x) - \frac{19}{4725}\mathsf{He}_{10}(x) + \frac{311}{997920}\mathsf{He}_{12}(x) - \frac{719}{37837800}\mathsf{He}_{14}(x) +$*
$\frac{14297}{15567552000}\mathsf{He}_{16}(x) - \frac{35369}{1042053012000}\mathsf{He}_{18}(x) + (\frac{35369}{41682120480000} - \frac{1}{83364240960000}\sqrt{\frac{11163552839}{38}})\mathsf{He}_{20}(x).$

Remark 14 *While the $K = 1$ case follows from the orthogonality of Hermite polynomials, the $K = L = 2$ case is already nontrivial — it implies the existence of link functions with information exponent $p > 2$, and furthermore, squaring the function cannot reduce the information exponent to 1 or 2; note that for $f(x) = \mathsf{He}_k(x)$, the information exponent of f^2 is at most 2.*

In the lower bound construction we use the polynomial f from the above lemma with suitably large K and L. Specifically, we let $f_m = f$ for all m and sample v_m from a set $S \subset \mathbb{S}^{d-1}$ with almost orthogonal elements. We prove that the two target functions $\frac{1}{\sqrt{M}} \sum_{m=1}^{M} f(v_m^\top x)$ and $\frac{1}{\sqrt{M}} \sum_{m=1}^{M} f(v_m'^\top x)$, where $\{v_m'\}_{m=1}^{M}$ is an independent copy of $\{v_m\}_{m=1}^{M}$, cannot be distinguished by SQ unless the tolerance is below $d^{-\rho}$. To show this, we Taylor expand the target function and perform an entry-wise swapping of $f(v_m^\top x)$ with $f(v_m'^\top x)$, in which smoothness of the function is entailed by additive Gaussian noise. Specifically, for $\delta \ll 1$, we have

$$\mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \varsigma^2)}[g(x, z + \delta + \varepsilon)] = \sum_{k=0}^{K} a_k(x, z) \delta^k + O(\delta^{K+1}),$$

where $a_k(x, z) = \frac{1}{k!} \int g(x, w) \left(\frac{\mathrm{d}^k}{\mathrm{d}z^k} e^{-(w-z)^2/2\varsigma^2} \right) \mathrm{d}w = O(1)$ (see Lemma 46 for derivation). Below we heuristically demonstrate how swapping is conducted for $m = 1$:

$$\mathbb{E}\left[g\left(x, \tfrac{1}{\sqrt{M}} \sum_{m=1}^{M} f(v_m^\top x) + \varepsilon \right) \right]$$

$$\overset{(i)}{\approx} \mathbb{E}\left[g\left(x, \tfrac{1}{\sqrt{M}} \sum_{m=2}^{M} f(v_m^\top x) + \varepsilon \right) \right] + \sum_{k=1}^{K} \mathbb{E}\left[a_k\left(x, \tfrac{1}{\sqrt{M}} \sum_{m=2}^{M} f(v_m^\top x) \right) \frac{f^k(v_1^\top x)}{M^{\frac{i}{2}}} \right]$$

$$\overset{(ii)}{\approx} \mathbb{E}\left[g\left(x, \tfrac{1}{\sqrt{M}} \sum_{m=2}^{M} f(v_m^\top x) + \varepsilon \right) \right] + \sum_{k=1}^{K} \mathbb{E}\left[a_k\left(x, \tfrac{1}{\sqrt{M}} \sum_{m=2}^{M} f(v_m^\top x) \right) \right] \frac{\mathbb{E}\left[f^k(v_1^\top x) \right]}{M^{\frac{i}{2}}}$$

$$\overset{(iii)}{\approx} \mathbb{E}\left[g\left(x, \tfrac{1}{\sqrt{M}} \sum_{m=1}^{M-1} f(v_m^\top x) + \tfrac{1}{\sqrt{M}} f(v_1'^\top x) + \varepsilon \right) \right],$$

where (i) follows from Taylor expansion, (ii) is due to f^k being orthogonal to $\mathsf{He}_1, \ldots, \mathsf{He}_L$ and hence its correlation to a_k (which does not contain information of v_1) is approximated by its He_0 component, and (iii) is obtained by swapping $\mathbb{E}[f^k(v_1^\top x)]$ and $\mathbb{E}[f^k(v_1'^\top x)]$.

4.2.2. STATISTICAL-TO-COMPUTATIONAL GAP

The additive model (1.1) contains M directions $v_1, v_2, \ldots v_M \in \mathbb{R}^d$ and M univariate link functions $f_1, f_2, \ldots f_M : \mathbb{R} \to \mathbb{R}$ to be estimated; therefore, we intuitively expect a sample size of $n \gtrsim Md$ to be information-theoretically sufficient to learn this function class. Indeed, following [Suz18], it is easy to check that the covering number of the width-J neural network is $\log \mathcal{N}(\delta, \{\frac{1}{J} \sum_{j=1}^{M} a_j \sigma_j(w_j^\top \cdot + b_j)\}, \|\cdot\|_\infty) \lesssim Jd(\log J + \log d)$. If we take $J = Mq$ and let $\sigma_j = \mathsf{He}_i(\cdot) \, ((q-1)M < j \leq qM)$, the network $\frac{1}{J} \sum_{j=1}^{M} a_j \sigma_j(w_j^\top \cdot + b_j)$ can perfectly approximate the additive model (1.1). Therefore, by applying a standard generalization error bound (e.g., see [SH20, Lemma 4]), we can upper bound the squared loss by $\log \mathcal{N}(\delta, \{\frac{1}{J} \sum_{j=1}^{M} a_j \sigma_j(w_j^\top \cdot + b_j)\}, \|\cdot\|_\infty)/n$, which yields the following proposition (the detailed derivation of which we omit due to the standard proof).

Proposition 15 *For the additive model* (1.1) *with any p and q, there exists an (computationally inefficient) algorithm that returns a function $\hat{f}$ with L^2-error of ε using $n = \tilde{O}(Md\varepsilon^{-2})$ samples.*

The procedure in Proposition 15 involves finding the empirical risk minimizer of a neural network, which can be computationally infeasible in polynomial time [BR88; ABMM16]. Nevertheless, the existence of a statistically efficient algorithm that learns (1.1) in $n \gtrsim Md$ samples suggests a statistical-to-computational gap under the SQ class (with polynomial compute; see [DH22; DPVLB24] for related discussions), since the lower bound in Theorem 11 implies a worse sample complexity of $n \gtrsim (Md)^\rho$ for SQ learners, where ρ can be made arbitrarily large by varying p and q.

Importantly, the same statistical-computational gap is *not present* in the finite-M setting due to the restriction of f_* being polynomial; specifically, when $M = O_d(1)$, there exists an SQ algorithm that learns multi-index polynomials using $n = \tilde{O}_d(d)$ samples [CM20], and in the single-index setting, polynomial link functions have *generative exponent* at most 2 [DPVLB24], and hence the SQ lower bound only implies that a sample size of $n \gtrsim d$ is necessary. In contrast, in our large-M setting, the sample complexity of SQ algorithms may have large polynomial dimension dependence, despite the link functions being polynomial. The key observation is that when the number of tasks M is diverging, the nonlinear label transformation cannot obtain higher-order exponentiation of the individual single-index tasks, which is employed by SQ learners such as [MM18; CM20]; hence the information exponent of the link function may still be large after the nonlinear transformation.

5. Conclusion and Future Directions

We studied the learning of additive models, where the number of (diverse) single-index tasks M grows with the dimensionality d. We showed that a layer-wise SGD algorithm achieves $\tilde{O}(Md^{p-1})$ sample complexity, and a subset of first-layer neurons *localize* into each additive task by achieving significant alignment. We also investigated the computational barrier in learning the additive model class by establishing lower bounds for both Correlational SQ and Full SQ learners. Our lower bound suggests a computational-to-statistical gap under the SQ class, which highlights the fundamental difference between our large M setting and the previously studied finite-M multi-index regression.

We highlight a few future directions. First observe that there is a gap between our complexity upper bound for gradient-based training and the computational lower bounds, and one would hope to design more efficient algorithms that close such gap. It is also worth noting that in certain regimes, SGD on the squared loss may outperform the CSQ complexity due to the presence of higher-order (non-correlational) information, as demonstrated in recent works [DTA+24; LOSW24; ADK+24]; an interesting direction is to examine if similar mechanisms can improve the statistical efficiency of SGD in our setting. It is also important to relax the near-orthogonality condition for the true signals, and explore how our derived upper and lower bounds are affected. Last but not least, we may theoretically analyze task localization in more complicated architectures beyond two-layer neural network, such as multi-head attention [CSWY24].

Acknowledgement

References

[AAM22] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

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

[AB16] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[ABA22] Emmanuel Abbe and Enric Boix-Adsera. On the non-universality of deep learning: quantifying the cost of symmetry. *Advances in Neural Information Processing Systems*, 35:17188–17201, 2022.

[ABMM16] Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

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

[AG23] Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

[Akh90] Naum Il'ich Akhiezer. *Elements of the theory of elliptic functions*, volume 79. American Mathematical Soc., 1990.

[AMF+21] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34:4699–4711, 2021.

[APVZ14] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning sparse polynomial functions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 500–510. SIAM, 2014.

[AS21] Shunta Akiyama and Taiji Suzuki. On learnability via gradient method for two-layer relu neural networks in teacher-student setting. In *International Conference on Machine Learning*, pages 152–162. PMLR, 2021.

[BA24] Enric Boix-Adsera. Towards a theory of model distillation. *arXiv preprint arXiv:2403.09053*, 2024.

[Bac08] Francis R Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(6), 2008.

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

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

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

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

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

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

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

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

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

[BR88] Avrim Blum and Ronald Rivest. Training a 3-node neural network is np-complete. *Advances in neural information processing systems*, 1, 1988.

[BTB+23] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, page 2, 2023.

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

[CGKM22] Sitan Chen, Aravind Gollakota, Adam Klivans, and Raghu Meka. Hardness of noise-free learning for two-hidden-layer neural networks. *Advances in Neural Information Processing Systems*, 35:10709–10724, 2022.

[CHS+23] Liam Collins, Hamed Hassani, Mahdi Soltanolkotabi, Aryan Mokhtari, and Sanjay Shakkottai. Provable multi-task representation learning by two-layer relu neural networks. *arXiv preprint arXiv:2307.06887*, 2023.

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

[CSWY24] Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DDH+21] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

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

[DH22] Rishabh Dudeja and Daniel Hsu. Statistical-computational trade-offs in tensor pca and related problems via communication complexity. *arXiv preprint arXiv:2204.07526*, 2022.

[DHK+20] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[DKK+23] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostically learning multi-index models with queries. *arXiv preprint arXiv:2312.16616*, 2023.

[DKKZ20] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539. PMLR, 2020.

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

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

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

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

[EHO+22] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[EY07] Alexandre Eremenko and Peter Yuditskii. Uniform approximation of $\text{sgn}(x)$ by polynomials and entire functions. *Journal d Analyse Mathématique*, 101:313–324, 05 2007.

[FS81] Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

[GGJ+20] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020.

[GKS20] Ankit Garg, Neeraj Kayal, and Chandan Saha. Learning sums of powers of low-degree polynomials in the non-degenerate case. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 889–899. IEEE, 2020.

[GKZ19] David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Stationary points of shallow neural networks with quadratic activation function. *arXiv preprint arXiv:1912.01599*, 2019.

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

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

[HSW+21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[HT87] Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[KB16] Jason M Klusowski and Andrew R Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

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

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

[KSP15] Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. High dimensional bayesian optimisation and bandits via additive models. In *International conference on machine learning*, pages 295–304. PMLR, 2015.

[LL21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[LOSW24] Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

[Mau16] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 3–17, Cham, 2016. Springer International Publishing.

[MBB23] Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. *arXiv preprint arXiv:2311.03794*, 2023.

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

[MHWSE24] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[Muc70] Benjamin Muckenhoupt. Mean convergence of Hermite and Laguerre series. II. *Transactions of the American Mathematical Society*, 147(2):433–460, 1970.

[MVEZ20] Sarao Stefano Mannelli, Eric Vanden-Eijnden, and Lenka Zdeborová. Optimization and generalization of shallow neural networks with quadratic activation functions. *Advances in Neural Information Processing Systems*, 33:13445–13455, 2020.

[MYZ13] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[P+97] Allan Pinkus et al. Approximating by ridge functions. *Surface fitting and multiresolution methods*, pages 279–292, 1997.

[PSZA23] Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. *arXiv preprint arXiv:2302.06600*, 2023.

[Qi10] Feng Qi. Bounds for the ratio of two gamma functions. *Journal of Inequalities and Applications*, 2010:1–84, 2010.

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

[RJWY12] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of machine learning research*, 13(2), 2012.

[RLLW09] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

[SH20] Anselm Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Annals of statistics*, 48(4):1875–1897, 2020.

[SS12] Taiji Suzuki and Masashi Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. In *Artificial Intelligence and Statistics*, pages 1152–1183. PMLR, 2012.

[Sto85] Charles J Stone. Additive regression and other nonparametric models. *The annals of Statistics*, 13(2):689–705, 1985.

[Suz18] Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

[Szö09] Balázs Szörényi. Characterizing statistical query learning: Simplified notions and proofs. In *Algorithmic Learning Theory*, pages 186–200, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[TLS+23] Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023.

[VSL+22] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. *Advances in Neural Information Processing Systems*, 35:23244–23255, 2022.

[VW19] Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. In *Conference on Learning Theory*, pages 3115–3117. PMLR, 2019.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Number 48 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[WNL23] Zihao Wang, Eshaan Nichani, and Jason D Lee. Learning hierarchical polynomials with three-layer neural networks. *arXiv preprint arXiv:2311.13774*, 2023.

[WWZ+22] Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*, 2022.

[YH20] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

[ZGJ21] Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pages 4577–4632. PMLR, 2021.

[Zol77] E Zolotarev. Application of elliptic functions to questions concerning functions more or less deviating from zero. *Notes of the Russian Science Academy*, 30(5):1–71, 1877.

Table of Contents

Appendix A. Preliminaries

A.1. Hermite Polynomials

For $k \in \mathbb{Z}_+$, the k-th Hermite polynomial $\mathsf{He}_k \colon \mathbb{R} \to \mathbb{R}$ is a univariate function defined as $\mathsf{He}_k(t) = e^{\frac{t^2}{2}} \frac{\mathrm{d}^k}{\mathrm{d}t^k} e^{-\frac{t^2}{2}}$. The Hermite polynomials form an orthogonal basis for the Hilbert space of square-integrable functions. We provide several basic properties of the Hermite polynomials.

Lemma 16 *The Hermite polynomials satisfy the following properties:*

- **Derivatives:**

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathsf{He}_k(t) = k \mathsf{He}_{k-1}(t).$$

- **Integration by parts (I):**

$$\int \mathsf{He}_k(t) f(t) \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, \mathrm{d}t = \int \mathsf{He}_{k-1}(t) f'(t) \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, \mathrm{d}t.$$

- **Integration by parts (II):** *For $u \in \mathbb{S}^{d-1}(1)$ and $v \in \mathbb{R}^d$,*

$$\int \mathsf{He}_k(u^\top x) f(v^\top x) \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{\|x\|^2}{2}} \, \mathrm{d}x = (u^\top v) \int \mathsf{He}_{k-1}(u^\top t) f'(v^\top x) \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{\|x\|^2}{2}} \, \mathrm{d}x.$$

- **Orthogonality (I):**

$$\int \mathsf{He}_k(t) \mathsf{He}_l(t) \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, \mathrm{d}t = k! \delta_{k,l}.$$

- **Orthogonality (II):** *For $u, v \in \mathbb{S}^{d-1}(1)$,*

$$\int \mathsf{He}_k(u^\top x) \mathsf{He}_l(v^\top x) \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{\|x\|^2}{2}} \, \mathrm{d}x = k! (u^\top v)^k \delta_{k,l}.$$

- **Hermite expansion:** *if $f \in \mathbb{R} \to \mathbb{R}$ is square-integrable with respect to the standard Gaussian,*

$$f(t) \overset{L^2}{=} \sum_{k=0}^{\infty} \frac{\alpha_k}{k!} \mathsf{He}_k(t), \quad \alpha_k = \int f(t) \mathsf{He}_k(t) \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, \mathrm{d}t.$$

A.2. High Probability Bound on y

Recall that in the definition of f_*, we used a scaling factor of $\frac{1}{\sqrt{M}}$ instead of $\frac{1}{M}$ in order to ensure that the output is of order $\Theta(1)$ with high probability. This subsection proves the $\Theta(1)$ output scale rigorously. Since v_m are almost orthogonal and $f_m(t)$ are mean-zero, $f_m(v_m^\top x)$ are almost independent mean-zero variables. Thus we expect $\frac{1}{\sqrt{M}} \sum_{m=1}^{M} f_m(v_m^\top x) \to \mathcal{N}(0,1)$ by central limit theorem.

23

Lemma 17 *Let* $\|v_m\| = 1$ *for all* $m = 1, 2, \cdots, M$. *If* $M \leq \max_{m \neq m'} |v_m^\top v_{m'}|^{-p}$ *and* S *is even, we have*

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M} f_m(v_m^\top x)\right\|_S\right] \leq \mathbb{E}\left[\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M} f_m(v_m^\top x)\right\|_S^S\right]^{1/S} \lesssim S^{q+1}. \qquad \text{(A.1)}$$

Thus, by applying Lemma 18, we have

$$\left|\frac{1}{\sqrt{M}}\sum_{m=1}^{M} f_m(v_m^\top x)\right| \lesssim (\log 1/\delta)^{q+1} = \tilde{O}(1), \qquad \text{(A.2)}$$

with probability at least $1 - \delta$.

Proof. Recall the Hermite expansion of f_m: $f_m(t) = \sum_{i=p}^{q} \alpha_{m,i}\mathsf{He}_i(t)$. We decompose the LHS of (A.1) as

$$\mathbb{E}\left[\left|\frac{1}{\sqrt{M}}\sum_{m=1}^{M} f_m(v_m^\top x)\right|^S\right]$$

$$\leq \frac{1}{\sqrt{M}^S}\sum_{(m_1,\cdots,m_S)\in[M]^S} \mathbb{E}\left[f_{m_1}(v_{m_1}^\top x)\cdots f_{m_S}(v_{m_S}^\top x)\right].$$

$$\leq \frac{1}{\sqrt{M}^S}\sum_{(m_1,\cdots,m_S)\in[M]^S}\sum_{(i_1,\cdots,i_S)\in[q-p+1]^S} (\alpha_{m_1,i_1}\cdots\alpha_{m_S,i_S})\mathbb{E}\left[\mathsf{He}_{i_1}(v_{m_1}^\top x)\cdots\mathsf{He}_{i_S}(v_{m_S}^\top x)\right].$$

$$\text{(A.3)}$$

We evaluate each term of (A.3). If $i_1 + \cdots + i_S$ is odd, the term is 0. Otherwise, we bound the value of $\mathbb{E}[\mathsf{He}_{i_1}(v_{m_1}^\top x)\cdots\mathsf{He}_{i_S}(v_{m_S}^\top x)]$ recursively.

For $T \in \mathbb{N}$, $\boldsymbol{i} = (i_1, \cdots, i_T)$, and $\boldsymbol{m} = (m_1, \cdots, m_T)$, let us define

$$A(T, \boldsymbol{i}, \boldsymbol{m}) := \mathbb{E}[\mathsf{He}_{i_1}(v_{m_1}^\top x)\cdots\mathsf{He}_{i_T}(v_{m_T}^\top x)],$$

and

$$B(T-1, \boldsymbol{i}) = \left\{\boldsymbol{j} = (j_1, \cdots, j_{T-1}) \in \mathbb{Z}_+^{T-1} \middle| \sum_{t=1}^{T-1} j_t = i_T, \ j_t \leq i_t \ (t = 1, \cdots, T-1)\right\}.$$

For $\boldsymbol{j} \in B(T-1, \boldsymbol{i})$, the multinomial coefficient for permuting a multiset of $\sum_{t=1}^{T-1} j_t$ elements (where j_t is the multiplicity of each of the t-th element) is denoted by

$$c_{\boldsymbol{j}} := \frac{(\sum_{t=1}^{T-1} j_t)!}{(j_1)!\cdots(j_{T-1})!}.$$

By the basic property of Hermite polynomials, we have

$$A(T, \boldsymbol{i}, \boldsymbol{m}) = \mathbb{E}\left[\frac{\partial^{i_T}}{\partial(v_{m_T}^\top x)^{i_T}}\left(\mathsf{He}_{i_1}(v_{m_1}^\top x)\cdots\mathsf{He}_{i_{T-1}}(v_{m_{T-1}}^\top x)\right)\right]$$

$$= \sum_{\boldsymbol{j} \in B(T-1, \boldsymbol{i})} c_{\boldsymbol{j}} \prod_{t=1}^{T-1} \left[(v_{m_T}^\top v_{m_t}^\top)^{j_t}{}_{i_t} P_{j_t} \right] A(T-1, \boldsymbol{i}_{1:T-1} - \boldsymbol{j}, \boldsymbol{m}_{1:T-1}).$$

We can bound the coefficients as $c_{\boldsymbol{j}} \le i_T! \le q!$ and $\prod_{t=1}^{T-1}{}_{i_t} P_{j_t} \le q^q$. We also bound the size of $B(T-1, \boldsymbol{i})$ by S^q. We define $I(m_1, \cdots, m_S)$ as the number of m distinct in $(m_1, \cdots, m_S)$. Then, by recursively bounding A, we have

$$A(S, (i_1, \cdots, i_S), (m_1, \cdots, m_S)) \le (q! q^q S^q)^S \left(\max_{m \ne m'} |v_m^\top v_{m'}| \right)^{\frac{p I(m_1, \cdots, m_S)}{2}}. \qquad (A.4)$$

Moreover,

$$\sum_{(m_1, \cdots, m_S) \in [M]^S} \left(\max_{m \ne m'} |v_m^\top v_{m'}| \right)^{\frac{p I(m_1, \cdots, m_S)}{2}}$$

$$= \sum_{i=0}^{S} \sum_{(m_1, \cdots, m_S) \in [M]^S} \mathbb{1}[I(m_1, \cdots, m_S) = i] \left(\max_{m \ne m'} |v_m^\top v_{m'}| \right)^{\frac{pi}{2}}$$

$$\le \sum_{i=0}^{S} \mathbb{1} \left[i \ge \frac{S-i}{2}, \frac{S-i}{2} \in \mathbb{Z}_+ \right] {}_M P_{M-i} \cdot {}_i P_{\frac{S-i}{2}} \left(\frac{S-i}{2} \right)^{\frac{S-i}{2}} \left(\max_{m \ne m'} |v_m^\top v_{m'}| \right)^{\frac{pi}{2}}$$

$$\le \max_{0 \le i \le S} M^i S^S \left(\max_{m \ne m'} |v_m^\top v_{m'}| \right)^{\frac{pi}{2}}. \qquad (A.5)$$

If $M \le (\max_{m \ne m'} |v_m^\top v_{m'}|)^{-p}$, applying (A.4) and (A.5) to (A.3), we obtain

$$\mathbb{E} \left[\left| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f_m(v_m^\top x) \right|^S \right]$$

$$\le \frac{1}{\sqrt{M}^S} \sum_{(m_1, \cdots, m_S) \in [M]^S} \sum_{(i_1, \cdots, i_S) \in [q-p+1]^S} |\alpha_{m_1, i_1} \cdots \alpha_{m_S, i_S}| (q! q^q S^q)^S \left(\max_{m \ne m'} |v_m^\top v_{m'}| \right)^{\frac{p I(m_1, \cdots, m_S)}{2}}$$

$$\le \frac{((q-p+1) q! q^q S^q \max |\alpha_{m_s, i_s}|)^S}{\sqrt{M}^S} \sum_{(m_1, \cdots, m_S) \in [M]^S} \left(\max_{m \ne m'} |v_m^\top v_{m'}| \right)^{\frac{p I(m_1, \cdots, m_S)}{2}}$$

$$\le \frac{((q-p+1) q! q^q S^q \max |\alpha_{m_s, i_s}|)^S}{\sqrt{M}^S} \max_{0 \le i \le S} M^i S^S \left(\max_{m \ne m'} |v_m^\top v_{m'}| \right)^{\frac{pi}{2}}$$

$$\le \left((q-p+1) q! q^q S^{q+1} \max |\alpha_{m_s, i_s}| \right)^S,$$

which yields (A.1).

By applying Lemma 18 with $c_1 = q + 1$, we have

$$\left| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f_m(v_m^\top x) \right| \lesssim (\log 1/\delta)^{q+1},$$

with probability $1 - \delta$, which yields (A.3). ∎

The above proof makes use of the following classical inequality.

Lemma 18 *Let $\delta > 0$ and X be a mean-zero random variable satisfying*

$$\mathbb{E}[|X|^S] \leq C_1 S^{c_1} \quad \text{for } S = \frac{\log 1/\delta}{c_1}$$

for $C_1, c_1 > 0$. Then, with probability at least $1 - \delta$, we have

$$|X| \leq C_1(eS)^{c_1}.$$

Proof. The proof follows from [DNGL23].

$$\mathbb{P}[|X| \geq C_1(eS)^{c_1}] = \mathbb{P}[|X|^S \geq C_1^S(eS)^{c_1 S}] \leq \frac{\mathbb{E}[|X|^S]}{C_1^S(eS)^{c_1 S}} \leq \frac{C_1^S S^{c_1 S}}{C_1^S(eS)^{c_1 S}} \leq e^{-c_1 S} = \delta,$$

which concludes the proof. ∎

A.3. Bihari–LaSalle Inequality and Gronwall Inequality

For later use, we provide the proofs of the Bihari–LaSalle inequality and the Grönwall's inequality for completeness. The proof of the Bihari–LaSalle inequality is borrowed from [BAGJ21].

Lemma 19 (Bihari–LaSalle inequality and Gronwall inequality) *For $p \geq 3$ and $c > 0$, consider a positive sequence $(a^t)_{t \geq 0}$ such that*

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

Then, we have

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

Moreover, when $a^t \leq 1$ holds for all $t \leq T - 1$, we have

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

for all $t \leq T$.

Proof. From definition, we have

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

Taking the summation and re-arranging the terms yield

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

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

which gives the lower bound.

On the other hand, when $a^t \leq 1$, we have $a^{t+1} \leq (1+c)a^t$, and therefore

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

By taking the summation and re-arranging the terms yield

$$(a^t)^{-(p-2)} \leq (a^0)^{-(p-2)} - c(1+c)^{p-1}(k-1)t,$$

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

which gives the upper bound. ∎

A.4. Orthonormal Basis from Nearly Orthogonal Vectors

In case where the feature vectors v_m are not orthogonal, the following lemma shows an orthonormal basis can be constructed as a linear combination of v_m.

Lemma 20 *Let $\{v_m\}_{m=1}^M \subset \mathbb{S}^{d-1}$ be a set of unit vectors in $\mathbb{R}^d$, and suppose that $\max_{m \neq m'} |v_m^\top v_{m'}| \leq \frac{1}{2}M^{-1}$ holds. Then, we can construct an orthonormal basis $\{\tilde{v}_m\}_{m=1}^M \subset \mathbb{S}^{d-1}$ so that $\tilde{v}_m$ is a linear combination of $v_1, \cdots, v_m$. Specifically, we can take $\tilde{v}_1 = v_1$ and*

$$\tilde{v}_m = \sum_{m'=1}^m c_{m,m'} v_{m'}, \tag{A.6}$$

so that $|c_{m,m'}| \leq 4 \max_{m',m''} |v_{m'}^\top v_{m''}|$ for $m' = 1, \cdots, m-1$, $|1 - c_{m,m}| \leq 20M \max_{m' \neq m''} |v_{m'}^\top v_{m''}|$, and $\tilde{v}_m^\top v_{m'} = 0$ for $m' = 1, \cdots, m-1$ hold.

Proof. First, we show that $\dim(\mathrm{span}\{v_1, \cdots, v_m\}) = m$ for all m. Assume the opposite and then we have

$$v_m = \sum_{m'=1}^{m-1} a_{m,m'} v_{m'}$$

for some m and $\{a_{m,m'}\}_{1 \leq m' \leq m-1}$. Then, we have

$$1 = \|v_m\|^2 \geq \sum_{m'=1}^{m-1} a_{m,m'}^2 + \sum_{m' \neq m''} a_{m,m'} a_{m,m''} v_{m'}^\top v_{m''}$$

$$\geq \sum_{m'=1}^{m-1} a_{m,m'}^2 - \sum_{m' \neq m''} \frac{|a_{m,m'}||a_{m,m''}|}{2M}$$

27

$$\geq \sum_{m'=1}^{m-1} a_{m,m'}^2 - \frac{1}{2M}\left(\sum_{m'=1}^{m-1} |a_{m,m'}|\right)^2$$

$$\geq \sum_{m'=1}^{m-1} a_{m,m'}^2 - \frac{1}{2}\sum_{m'=1}^{m-1} a_{m,m'}^2 = \frac{1}{2}\sum_{m'=1}^{m-1} a_{m,m'}^2, \tag{A.7}$$

where we used the Cauchy–Schwarz inequality for the last step. Thus, we have $|a_{m,m'}| \leq \sqrt{2}$ and therefore

$$1 = \|v_m\|^2 = \sum_{m'=1}^{m-1} a_{m,m'} v_{m'}^\top v_m \leq (M-1)\max_{m'\neq m}|a_{m,m'}| \cdot \max_{m'\neq m}|v_{m'}^\top v_m| \leq (M-1)\cdot\sqrt{2}\cdot\frac{1}{2M} < 1,$$

which yields the contradiction.

Therefore, we have an orthogonal basis $\{\tilde{v}_m\}_{m=1}^M$ such that (A.6) and $\tilde{v}_m^\top v_{m'} = 0$ ($1 \leq m \leq M, 1 \leq m' \leq m-1$) hold, with some $\{c_{m,m'}\}_{1\leq m\leq M, 1\leq m'\leq m}$. Similarly to (A.7), we have $|c_{m,m'}| \leq \sqrt{2}$ for all $1 \leq m \leq M$ and $1 \leq m' \leq m$. What remains is to bound the coefficients. Since we have

$$\tilde{v}_m = \sum_{m'=1}^m c_{m,m'} v_{m'}$$

for all m, taking an inner product with $\sum_{m'=1}^{m-1}\text{sign}(c_{m,m'})v_{m'}$, we get

$$0 = \tilde{v}_m^\top \sum_{m'=1}^{m-1}\text{sign}(c_{m,m'})v_{m'} = \left(\sum_{m'=1}^m c_{m,m'}v_{m'}\right)^\top \sum_{m'=1}^{m-1}\text{sign}(c_{m,m'})v_{m'}$$

$$= \sum_{m'=1}^{m-1}|c_{m,m'}| + \sum_{m'=1}^{m-1}\sum_{m''\neq m',m} c_{m,m'}\text{sign}(c_{m,m''})v_{m'}^\top v_{m''} + \sum_{m'=1}^{m-1} c_{m,m}\text{sign}(c_{m,m'})v_m^\top v_{m'}$$

$$\geq \sum_{m'=1}^{m-1}|c_{m,m'}| - \sum_{m'=1}^{m-1} m\sqrt{2}\frac{1}{2M} - \sqrt{2}m\frac{1}{2M}$$

$$\geq \left(1 - \frac{1}{\sqrt{2}}\right)\sum_{m'=1}^{m-1}|c_{m,m'}| - \frac{1}{\sqrt{2}}.$$

Thus, $\sum_{m'=1}^{m-1}|c_{m,m'}|$ is bounded by $\sqrt{2}+1 \leq 3$ and $\sum_{m'=1}^m |c_{m,m'}|$ is bounded by $\sqrt{2}+1+\sqrt{2} \leq 4$.

Also considering an inner product with $\text{sign}(c_{m,m'})v_{m'}$ ($1 \leq m' \leq m-1$), we get

$$0 = \tilde{v}_m^\top \text{sign}(c_{m,m'})v_{m'} = \left(\sum_{m''=1}^m c_{m,m''}v_{m''}\right)^\top \text{sign}(c_{m,m'})v_{m'}$$

$$= |c_{m,m'}| + \sum_{m''\neq m'} c_{m,m''}\text{sign}(c_{m,m'})v_{m'}^\top v_{m''}$$

$$\geq |c_{m,m'}| - \sum_{m'=1}^m |c_{m,m'}| \max_{m'\neq m''}|v_{m'}^\top v_{m''}|.$$

Thus, $|c_{m,m'}|$ $(1 \leq m' \leq m-1)$ is bounded by $\sum_{m'=1}^{m} |c_{m,m'}| \max_{m' \neq m''} |v_{m'}^{\top} v_{m''}| \leq 4 \max_{m' \neq m''} |v_{m'}^{\top} v_{m''}|$. Finally, we get

$$1 = \|\tilde{v}_m\|^2 = \sum_{m'=1}^{m} c_{m,m'}^2 + \sum_{m'=1}^{m} \sum_{m'' \neq m'} c_{m,m'} c_{m,m''} v_{m'}^{\top} v_{m''}$$

$$\therefore |1 - c_{m,m}^2| \leq \sum_{m'=1}^{m-1} |c_{m,m'}| + \sum_{m'=1}^{m} |c_{m,m'}| \sum_{m'' \neq m'} |c_{m,m''}| |v_{m'}^{\top} v_{m''}|$$

$$\leq 4M \max_{m' \neq m''} |v_{m'}^{\top} v_{m''}| + 16 \max_{m' \neq m''} |v_{m'}^{\top} v_{m''}| \leq 20M \max_{m' \neq m''} |v_{m'}^{\top} v_{m''}|,$$

which implies that $|1 - c_{m,m}| \leq 20M \max_{m' \neq m''} |v_{m'}^{\top} v_{m''}|$. ∎

Appendix B. Proof of Gradient-based Training

Overview of analysis. We define polylogarithmic constants with the following order of strength:

$$C_1 \lesssim c_1^{-1} \lesssim c_2^{-1} \lesssim C_2 \lesssim c_3^{-1} \lesssim c_4^{-1} = \tilde{O}(1).$$

Here c_1 and C_1 are different from those in Lemma 18. c_4 is the same as c_v used in Assumption 2. C_1 will be used to represent any polylogarithmic factor that comes from high probability bounds. Also, Section B.1 will introduce another constant C_p.

In Algorithm 1 we first train the first-layer parameters, where we aim to show that a for each class m, there exist sufficiently many neurons that almost align with v_m to approximate each $f_m(v_m^{\top} x)$. We define the alignment for the m-th task at time t as $\kappa_m^t = v_m w_j^t$. The goal of first-layer training is to prove the following. We introduce the error $\tilde{\varepsilon} = \tilde{\Theta}(M^{-\frac{1}{2}}\varepsilon)$, where ε is the desired final generalization error $\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim \varepsilon$.

Lemma 7 (formal) *Let $T_{1,1} = \tilde{\Theta}(Md^{p-1})$, $T_{1,2} = \tilde{\Theta}(Md^{\frac{p}{2}})$, $T_{1,3} = \tilde{\Theta}(\tilde{\varepsilon}^{-2}Md \vee \tilde{\varepsilon}^{-3}M)$, and $T_1 = T_{1,1} + T_{1,2} + T_{1,3}$. Take the step size as $\eta^t = \tilde{\Theta}(M^{-\frac{1}{2}}d^{-\frac{p}{2}})$ for $0 \leq t \leq T_{1,1} + T_{1,2} - 1$ and $\eta^t = \tilde{\Theta}(\tilde{\varepsilon}M^{-\frac{1}{2}}d^{-1} \wedge \tilde{\varepsilon}^2 M^{-\frac{1}{2}})$ for $T_{1,1} + T_{1,2} \leq t \leq T_{1,1} + T_{1,2} + T_{1,3}$, and the number of neurons as $J \gtrsim J_{\min}M^{C_p}\log d$. Suppose that $|v_{m'}^{\top} v_m| = \tilde{O}(M^{-1})$ for all $m' \neq m$, and $M = \tilde{O}(d^{\frac{1}{2}})$. Then, with high probability, for each class m there exist at least $J_{\min}$ neurons that achieves $v_m^{\top} w_j^{T_1} \geq 1 - 3\tilde{\varepsilon}$.*

We let $T_1' = T_1 + T_2$ in Theorem 4. The proof for Lemma B will be divided into the following parts. First we consider the initialization and activation functions.

- In Section B.1 we analyze the random initialization. We show that, at the time of initialization, for each class m, there exist $J_{\min}$ neurons classified into some class J_m (i.e., with slightly higher overlap with v_m).

- Section B.2 discusses the assumptions on the target/activation functions. We show that Assumption 4 is satisfied with a constant probability (i.e., $\Omega(1)$ fraction of neurons). Since neurons do not interact in the correlation loss update, in the subsequent sections of the first-layer training, we focus on the dynamics of one neuron in J_1 that satisfies Assumption 4 and omit the subscript $w^t = w_j^t$ without loss of generality.

Next we consider the training of the first-layer weights via a *power-method like dynamics*.

- In Section B.3, we decompose the gradient update into the population dynamics and noise fluctuation. The training dynamics consist of following three different phases.

- The first phase corresponds to Section B.4. Here, we show that neurons in J_m will obtain a small $\tilde{\Omega}(1)$ alignment to the class v_m after $t_1 (\leq T_{1,1})$ iterations.

- The second phase corresponds to Section B.5, where we show that neurons continue to grow in the direction of v_m and achieves $1 - \tilde{O}(1)$ alignment within $t_2 (\leq T_{1,2})$ iterations, while remaining almost orthogonal to other directions.

- Finally, in Section B.6, we show that, after $(T_{1,1} - t_1) + (T_{1,2} - t_2) + T_{1,3}$ iterations, neurons in J_m will eventually achieve $w_j^\top v_m \geq 1 - \tilde{\varepsilon}$ with high probability.

After first-layer training is completed, in Section B.7 we prove the existence of suitable second-layer parameters with small norm that can approximate f_* (Lemma 38). Section B.8 concludes the generalization error analysis, which is stated as follows:

Lemma 21 *Suppose that $J = \tilde{\Theta}(J_{\min} M^{C_p})$, and σ be either of the ReLU activation or any univariate polynomial with degree q. There exists $\lambda > 0$ such that for ridge regression ($p = 2$), we have*

$$\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim M^{\frac{1}{2}}(|J_{\min}|^{-1} + \tilde{\varepsilon}) + \sqrt{\frac{dM^{C_p}}{T_2}}$$

with probability $1 - o_d(1)$. Therefore, by taking $T_2 = \tilde{\Theta}(dM^{C_p}\varepsilon^{-2})$, $\tilde{\varepsilon} = \tilde{\Theta}(M^{-\frac{1}{2}}\varepsilon)$, $J_{\min} = \tilde{\Theta}(M^{\frac{1}{2}}\varepsilon^{-1})$, and $J = \tilde{\Theta}(M^{C_p+\frac{1}{2}}\varepsilon^{-1})$, we have $\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim \varepsilon$.

On the other hand, for LASSO ($p = 1$), we have

$$\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim M^{\frac{1}{2}}(|J_{\min}|^{-1} + \tilde{\varepsilon}) + \sqrt{\frac{dM}{T_2}}$$

with probability $1 - o_d(1)$. Therefore, by taking $T_2 = \tilde{\Theta}(dM\varepsilon^{-2})$, $\tilde{\varepsilon} = \tilde{\Theta}(M^{-\frac{1}{2}}\varepsilon)$, $J_{\min} = \tilde{\Theta}(M^{\frac{1}{2}}\varepsilon^{-1})$, and $J = \tilde{\Theta}(M^{C_p+\frac{1}{2}}\varepsilon^{-1})$, we have $\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim \varepsilon$.

Combining Lemma 21 and Lemma B completes the proof of Theorem 4.

B.1. Initialization

To begin with, we need diversity of the neurons at initialization. Note that we do not require neurons to achieve $\Theta(1)$ alignment with v^m, as this requires exponential width. Instead, we prove that, for each class v^m, there exist some neurons w_j^0 that are more aligned to this direction v^m than others — in other words, $v_m^\top w_j^0 \gtrsim \max_{m' \neq m} |v_{m'}^\top w_j^0|$. The statement is formalized as follows. In the proof, the subscript t is omitted because we only consider the initialization $t = 0$.

Lemma 5 (formal) *Suppose that the first layer weight of each neuron is initialized as an independent sample from the uniform distribution over $\mathbb{S}^{d-1}(1)$. Let $p > 2$, $\delta = (\log d)^{-\frac{p-2}{2}}$, and*

$C_p = \left(\frac{\max_m |\beta_{m,p}|}{\min_{m'} |\beta_{m',p}|} \right)^{\frac{2}{p-2}}$. *We define a set of indexes of neurons J_m that have the highest alignment with $f_m(v_m^\top x)$ as*

$$J_m := \left\{ j \in [J] \ \middle| \ w_j^\top v_m \geq \frac{1}{\sqrt{d}}, \ (w_j^\top v_m)^{p-2} \geq \max_{m' \neq m} C_p^{\frac{p-2}{2}} |w_j^\top v_{m'}|^{p-2} + \delta \left(\frac{1}{\sqrt{d}} \right)^{p-2} \right\}.$$

For $J_{\min} > 0$, if

$$J \geq A J_{\min} M^{C_P} (\log d)^{\frac{3}{2}}, \ \text{with } A = \exp \left(O \left(\max_{m \neq m'} |v_m^\top v_{m'}| \log d + 1 \right) \right),$$

then $|J_m| \geq J_{\min}$ for all m with high probability.

Since $w_j^\top v_m = O(\sqrt{\log d})$ with high probability, we have the following corollary, where we use a small constant $c_1 \lesssim (\log d)^{(p-2)}$.

Corollary 22 *When $J \gtrsim J_{\min} M^{C_P} \log d$, for each class m, we have at least $J_{\min}$ neurons w_j such that*

$$w_j^\top v_m \geq \frac{1}{\sqrt{d}}, \ \text{and } |\beta_{m,p}|(w_j^\top v_m)^{p-2} \geq \max_{m'} |\beta_{m',p}| \max_{m'' \neq m} |w_j^\top v_{m''}|^{p-2} + c_1 (w_j^\top v_m)^{p-2}. \quad \text{(B.1)}$$

To prove the lemma, we make use of the following upper and lower bounds.

Lemma 23 (Theorems 1 and 2 of [CCM11]) *For any $\beta > 1$ and $x \in \mathbb{R}$, we have*

$$\frac{\sqrt{2e(\beta - 1)}}{2\beta\sqrt{\pi}} e^{-\frac{\beta x^2}{2}} \leq \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt \leq \frac{1}{2} e^{-\frac{x^2}{2}}$$

Proof of Lemma 5. Note that $w_j \sim \text{Unif}(\mathbb{S}^{d-1}(1))$ can be obtained by sampling $\tilde{w}_j \sim \mathcal{N}(0, \frac{1}{d} I_d)$ and setting $w_j = \frac{\tilde{w}_j}{\|\tilde{w}_j\|}$. With high probability, we have $\|\tilde{w}_j\| \approx 1 \leq 2^{\frac{1}{p-2}}$. Thus we will instead show that

$$\tilde{w}_j^\top v_m \geq \frac{2}{\sqrt{d}}, \ \text{and } (\tilde{w}_j^\top v_m)^{p-2} \geq \max_{m' \neq m} C_p^{\frac{p-2}{2}} |\tilde{w}_j^\top v_{m'}|^{p-2} + \frac{2\delta}{d^{\frac{p-2}{2}}}.$$

Fix $m \in [M]$. For each $m' \neq m$, consider the value of $\tilde{w}_j^\top (I - v_m v_m^\top) v_{m'}$ and $\tilde{w}_j^\top v_m v_m^\top v_{m'}$. The distribution of $\tilde{w}_j^\top (I - v_m v_m^\top) v_{m'}$ follows $\mathcal{N}(0, \|(I - v_m v_m^\top) v_{m'}\|)$, therefore by Lemma 23,

$$\mathbb{P} \left[\text{for all } m' \neq m, |\tilde{w}_j^\top (I - v_m v_m^\top) v_{m'}| \leq t \right]$$
$$\leq 1 - (M - 1) \exp \left(-\frac{dt^2}{2 \max_{m' \neq m} \|(I - v_m v_m^\top) v_{m'}\|} \right). \quad \text{(B.2)}$$

By taking

$$t = t_1 := \left(2d^{-1} \max_{m' \neq m} \|(I - v_m v_m^\top) v_{m'}\| \log 2M \right)^{\frac{1}{2}},$$

(B.2) is bounded by $\frac{M+1}{2M}$.

Note that $\tilde{w}_j^\top v_m v_m^\top v_{m'} = O(d^{-\frac{1}{2}}\sqrt{\log d} \max_{m \neq m'} |v_m^\top v_{m'}|)$ with high probability. When $|\tilde{w}_j^\top (I - v_m v_m^\top)v_{m'}| \leq t_1$ for all $m' \neq m$ and $\tilde{w}_j^\top v_m v_m^\top v_{m'} = O(d^{-\frac{1}{2}}\sqrt{\log d} \max_{m \neq m'} |v_m^\top v_{m'}|)$, we have $|\tilde{w}_j^\top v_{m'}| \leq t_1 + O(d^{-\frac{1}{2}}\sqrt{\log d} \max_{m \neq m'} |v_m^\top v_{m'}|)$. To satisfy $(\tilde{w}_j^\top v_m)^{p-2} \geq C_p^{\frac{p-2}{2}} |\tilde{w}_j^\top v_{m'}|^{p-2} + \frac{2\delta}{d^{\frac{p-2}{2}}}$ and $\tilde{w}_j^\top v_m \geq \frac{2}{\sqrt{d}}$, it suffices for $\tilde{w}_j^\top v_m$ to be $w_j^\top v_m \geq t_2 := C_p^{\frac{1}{2}}t_1 + C_p^{\frac{1}{2}}O(d^{-\frac{1}{2}}\sqrt{\log d} \max_{m \neq m'} |v_m^\top v_{m'}|) + \frac{2^{\frac{1}{p-2}}\delta^{\frac{1}{p-2}}}{\sqrt{d}}$.

We lower bound the probability that $w_j^\top v_m \geq t_2$ holds. For any $\beta > 1$, Lemma 23 implies that

$$
\mathbb{P}\left[\tilde{w}_j^\top v_m \geq t_2\right]
$$
$$
\geq \frac{\sqrt{2e(\beta-1)}}{2\beta\sqrt{\pi}}\exp\left(-\frac{d\beta t_2^2}{2}\right)
$$
$$
\geq \frac{\sqrt{2e(\beta-1)}}{2\beta\sqrt{\pi}}\exp\left(-C_p\beta\left[\max_{m' \neq m}\|(I - v_m v_m^\top)v_{m'}\|\log 2M\right.\right.
$$
$$
\left.\left.+ O(\sqrt{\log d}\max_{m \neq m'}|v_m^\top v_{m'}| + \delta^{\frac{1}{p-2}})\sqrt{\log M} + O(\log d \max_{m \neq m'}|v_m^\top v_{m'}|^2 + \delta^{\frac{2}{p-2}})\right]\right)
$$
$$
\geq \frac{1}{(\sqrt{\log 2M} + \frac{1}{\sqrt{\log 2M}})\sqrt{\pi}}\exp\left(-C_p\left[\log 2M + O\left(\max_{m \neq m'}|v_m^\top v_{m'}|\log d + \delta^{\frac{2}{p-2}}\sqrt{\log d} + 1\right)\right]\right),
$$

where we took $\beta = 1 + \frac{1}{\log 2M}$ and used $\log M \lesssim \log d$, $\delta < 1$, and $\max_{m' \neq m}|v_m^\top v_{m'}| \leq 1$ for the last inequality. To simplify the notation, by letting

$$
A = \exp\left(O\left(\max_{m \neq m'}|v_m^\top v_{m'}|\log d + \delta^{\frac{1}{p-2}}\sqrt{\log d} + 1\right)\right) = \exp\left(O\left(\max_{m \neq m'}|v_m^\top v_{m'}|\log d + 1\right)\right),
$$

we have

$$
\mathbb{P}\left[\tilde{w}_j^\top v_m \geq t_2\right] \geq \frac{2M^{-C_p}}{A\sqrt{\log 2M}}.
$$

Note that this argument is independent from the one for (B.2), because $(I - v_m v_m^\top)v_{m'}$ and v_m are orthogonal.

To sum up, (B.2) ensures that $|\tilde{w}_j^\top (I - v_m v_m^\top)v_{m'}| \leq t_1$ for all $m' \neq m$ with probability at least $\frac{M+1}{2M}$, under which $\tilde{w}_j^\top v_m v_m^\top v_{m'} = O(d^{-\frac{1}{2}}\sqrt{\log d}\max_{m \neq m'}|v_m^\top v_{m'}|)$ for all $m' \neq m$ with high probability and $\tilde{w}_j^\top v_m \geq t_2$ holds with probability at least $\frac{2M^{-C_p}}{A\sqrt{\log M}}$. $\tilde{w}_j^\top v_m \geq t_2$ implies that $(\tilde{w}_j^\top v_m)^{p-2} \geq C|\tilde{w}_j^\top v_{m'}|^{p-2} + \frac{2\delta}{d^{\frac{p-2}{2}}}$ and $\tilde{w}_j^\top v_m \geq \frac{2}{\sqrt{d}}$. Therefore, over the randomness of initialization of w_j, $w_j^\top v_m \geq \frac{1}{\sqrt{d}}$ and $(w_j^\top v_m)^{p-2} \geq C|w_j^\top v_{m'}|^{p-2} + \delta(\frac{1}{\sqrt{d}})^{p-2}$ for all $m' \neq m$ with probability at least $\frac{M^{-C_p}}{A\sqrt{\log M}}$. Taking the uniform bound over all v_m, we know that the number of required neurons to satisfy $|J_m| \geq J_{\min}$ for all m is at most $J \geq AJ_{\min}M^{C_p}(\log d)^{\frac{3}{2}}$ (up to a constant factor, which can be absorbed in the definition of A). $\blacksquare$

B.2. Descent Path for Population Gradient

This section discusses the assumption on the target/activation functions. To translate *weak recovery* to *strong recovery* and establish alignment, we require that at any (positive) level of alignment $u = v_1^\top w^t > 0$, the population correlation loss between the neuron and the target sub-problem $f_1(v_1^\top x)$ has the descent path on the sphere. In other words, the population correlation loss as a function of u should be monotonically decreasing with respect to $u \in (0, 1]$. Observe that without such monotonicity, if the alignment α becomes large, higher-order terms in the Hermite expansion may generate a repulsive force that prevents the neuron from further aligning with the target direction v_m. For the well-specified setting (matching activation), this condition is automatically satisfied as shown in [BAGJ21]; whereas for the misspecified scenario, such a condition appeared in [MHWSE24] as an assumption, which we restate below, and then verify for specific choices of student activation functions.

Recall the Hermite expansion of one neuron $a^0 \sigma(z + b^0)$:

$$a^0 \sigma(z + b^0) = \sum_{i=0}^{\infty} \frac{\alpha_i}{\sqrt{i!}} \mathsf{He}_i(z)$$

and the Hermite expansion of each sub-problem of the target function

$$f_m(z) = \sum_{i=p}^{p} \frac{\beta_{m,i}}{\sqrt{i!}} \mathsf{He}_i(z).$$

Assumption 4 (Descent path) *The neuron satisfies* $\alpha_i \beta_{1,i} > 0$ *for all* $p \le i \le q$.

This assumption ensures that

$$\mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[a^0 \sigma(x^\top w^t + b^0) f_1(v_1^\top x)] = \sum_{i=p}^{q} (\alpha_i \beta_{m,i})^i$$

is monotonically increasing with respect to $u = \alpha_i \beta_{m,i} \in (0, 1]$. We show that for certain choices of (randomized) activation function, a $\tilde{\Omega}(1)$ fraction of student neurons satisfy Assumption 4. This indicates that Assumption 4 only affects the required width up to constant factor. In the subsequent sections, we focus on the training dynamics of individual neurons that satisfy Assumption 4 (without explicitly mentioning so).

B.2.1. RELU ACTIVATION

For the ReLU activation, we verify this condition for target function in which at each task f_m, the non-zero Hermite coefficients have the same sign, i.e., for $\beta_{m,i}, \beta_{m,j} \ne 0$, we have $\mathrm{sign}(\beta_{m,i}) = \mathrm{sign}(\beta_{m,j})$. For example, this condition is met when the link functions are pure Hermite polynomial. Then, the following lemma adapted from [BES$^+$23] ensures that $\alpha_i > 0$ for all i with probability at least $\frac{1}{4}$. Because the distribution of a^0 is symmetric, Assumption 4 holds with probability at least $\frac{1}{8}$.

Lemma 24 *Given degree* $q \ge 0$ *and* $b \sim [-C_b, C_b]$, *the* i-*th Hermite coefficient of* $\mathrm{ReLU}(z + b)$ *is positive with probability* $\frac{1}{4}$ *for all* $p \le i \le q$, *if* C_b *is larger than some constant that only depends on* q.

Proof. First note that for $i = 1$, we have

$$\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\text{ReLU}(z+b)\mathsf{He}_1(z)] = \frac{1}{2} \int_{z=b}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, \mathrm{d}z + \frac{1}{2} > 0.$$

Moreover, for $i \geq 2$, we have

$$\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\text{ReLU}(z+b)\mathsf{He}_i(z)] = \frac{(-1)^i}{\sqrt{2\pi}} e^{-\frac{b^2}{2}} \mathsf{He}_{i-2}(b),$$

Because $\lim_{b \to -\infty} \mathsf{He}_i(b) = \infty$ for even i and $-\infty$ for odd i, there exists some b_0 such that if $b \leq C_b'$ then $(-1)^p e^{\frac{b^2}{2}} \mathsf{He}_i(b) > 0$. By taking $C_b \geq 2C_b'$, $b \leq C_b'$ (and thus the assertion) holds with probability $\frac{1}{4}$. ∎

B.2.2. GENERAL POLYNOMIAL LINK FUNCTIONS

To deal with general polynomial link functions, we randomize the student activations as follows,

$$\sigma_j(z) = \sum_{i=p}^{q} \frac{\varepsilon_{i,j}}{\sqrt{i}} \mathsf{He}_i(z),$$

where ε_i are independent Rademacher variables (taking 1, -1, and 0 with equiprobability).

Lemma 25 *Given degree $q \geq 0$ and $b \sim [-C_b, C_b]$, for each $p_{\min} \leq p' \leq p_{\max}$, the i-th Hermite coefficient of $a^0 \sigma(z + b^0)$ is non-zero with probability $\Omega(C_b^{-1})$, for all $p' \leq i \leq q$. Here Ω hides constant only depending on q.*

Proof. Because of the randomized Hermite coefficient of the activation function, $a^0 \sigma(z)$ has positive coefficients with probability $2^{-(q-p+1)}$. As long as the bias b^0 is small, $a^0 \sigma(z + b^0)$ also have positive coefficients. ∎

B.3. Decomposition of Gradient Update

From now, we discuss the training dynamics of the first layer. We focus on one neuron in J_1 that satisfies Assumption 4. To track the alignment during the dynamics, we define $\kappa_m^t = v_m^\top w_j^t$. We first consider the decomposition of the update into population and stochastic terms:

Lemma 26 *Suppose that $\eta^t = \eta \leq c_4 d^{-1}$ and $\kappa_1^t \geq \frac{1}{2} d^{-\frac{1}{2}}$. With high probability, the update of κ_m^t can be bounded as*

$$\kappa_1^{t+1} \geq \kappa_1^t + \frac{\eta}{\sqrt{M}} \sum_{m=1}^{M} \sum_{i=p}^{q} \left[i\alpha_i \beta_{m,i} (\kappa_m^t)^{i-1} (v_1^\top v_m - \kappa_1^t \kappa_m^t) \right] - \kappa_1^t \eta^2 C_1^2 d + \eta v_1^\top (I - w^t w^{t\top}) Z^t.$$

(B.3)

Moreover, κ_m^{t+1} is evaluated as

$$\kappa_m^t + \frac{\eta}{\sqrt{M}} \sum_{m'=1}^{M} \sum_{i=p}^{q} \left[i\alpha_i \beta_{m',i} (\kappa_{m'}^t)^{i-1} (v_m^\top v_{m'} - \kappa_m^t \kappa_{m'}^t) \right] - \frac{|\kappa_m^t| + \eta C_1 d^{\frac{1}{2}}}{2} \eta^2 C_1^2 d$$

$$+ \eta v_m^\top (I - w^t w^{t\top}) Z^t$$
$$\leq \kappa_m^{t+1}$$
$$\leq \kappa_m^t + \frac{\eta}{\sqrt{M}} \sum_{m'=1}^M \sum_{i=p}^q \left[i\alpha_i \beta_{m',i} (\kappa_{m'}^t)^{i-1} (v_m^\top v_{m'} - \kappa_m^t \kappa_{m'}^t) \right] + \frac{|\kappa_m^t| + \eta C_1 d^{\frac{1}{2}}}{2} \eta^2 C_1^2 d$$
$$+ \eta v_m^\top (I - w^t w^{t\top}) Z^t.$$

Here Z^t is a mean-zero random variable satisfying $\|Z^t\| = \tilde{O}(1)$ with high probability. For any $v \in \mathbb{R}^d$ with $\|v\| = O(1)$ that is independent from Z^t, $|v^\top Z^t| = \tilde{O}(1)$ with high probability. Also, $|\kappa_m^t - \kappa_m^{t+1}| = \tilde{O}(\eta)$ with high probability.

To prove Lemma 26, we first establish the following characterization of the stochastic gradient.

Lemma 27 *The stochastic gradient $-\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)$ is decomposed as*

$$- \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)$$
$$= -\frac{1}{\sqrt{M}} \sum_{m=1}^M \sum_{i=p}^q \left[i\alpha_i \beta_{m,i} (\kappa_m^t)^{i-1} v_m + \sqrt{(i+2)(i+1)} \alpha_{i+2} \beta_{m,i} (\kappa_m^t)^i w^t \right] + Z^t,$$

where Z^t is a mean-zero random variable such that $\|Z^t\| = \tilde{O}(d^{\frac{1}{2}})$ and $|v^\top Z^t| = \tilde{O}(1)$ for any fixed $v \in \mathbb{S}^{d-1}$ with high probability. Also, $\|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\| = \tilde{O}(d^{\frac{1}{2}})$ with high probability, and for any fixed v with $\|v\| = O(1)$, $\left(\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) \right)^\top v = \tilde{O}(1)$ with high probability.

Proof. For i-th Hermite polynomial He_i and $u \in \mathbb{S}^{d-1}$, we have that

$$\mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[\mathsf{He}_i(x_1) f(u^\top x) x_1] = i u_1^{i-1} \mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[f^{(i-1)}(u^\top x)] + u_1^{i+1} \mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[f^{(i+1)}(u^\top x)],$$
$$\mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[\mathsf{He}_i(x_1) f(u^\top x) x_2] = u_1^i u_2 \mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[f^{(i+1)}(u^\top x)].$$

Therefore,

$$\mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[\mathsf{He}_i(x_1) f(u^\top x) x] = \begin{pmatrix} i u_1^{i-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[f^{(i-1)}(u^\top x)] + u_1^i u \mathbb{E}_{x \sim \mathcal{N}(0, I_d)}[f^{(i+1)}(u^\top x)].$$

Using this fact, the population gradient is computed as

$$\nabla_w \mathbb{E}\left[y^t a^t \sigma(w^{t\top} x + b^t) \right]$$
$$= \mathbb{E}\left[\frac{1}{\sqrt{M}} \sum_{m=1}^M f_m(v_m^\top x) a^t \sigma'(w^{t\top} x + b^t) x \right]$$
$$= \mathbb{E}\left[\frac{1}{\sqrt{M}} \left(\sum_{m=1}^M \sum_{i=p}^p \frac{\beta_{m,i}}{\sqrt{i!}} \mathsf{He}_i(z) \right) \left(\sum_{i=0}^\infty \frac{i\alpha_i}{\sqrt{i!}} \mathsf{He}_{i-1}(z) \right) x \right]$$

$$= \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \sum_{i=p}^{q} \left[i^2 \alpha_i \beta_{m,i} (v_m^\top w^t)^{i-1} v_m + \sqrt{(i+2)(i+1)} \alpha_{i+2} \beta_{m,i} (v_m^\top w^t)^i w^t \right]$$

$$= \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \sum_{i=p}^{q} \left[i \alpha_i \beta_{m,i} (\kappa_m^t)^{i-1} v_m + \sqrt{(i+2)(i+1)} \alpha_{i+2} \beta_{m,i} (\kappa_m^t)^i w^t \right].$$

We define Z^t as the difference between the population gradient and the empirical gradient:

$$Z^t = -\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) + \nabla_w \mathbb{E} \left[y^t a^t \sigma(w^{t\top} x + b^t) \right].$$

We have

$$- \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)$$

$$= -\frac{1}{\sqrt{M}} \sum_{m=1}^{M} \sum_{i=p}^{q} \left[i \alpha_i \beta_{m,i} (\kappa_m^t)^{i-1} v_m + \sqrt{(i+2)(i+1)} \alpha_{i+2} \beta_{m,i} (\kappa_m^t)^i w^t \right] + Z^t$$

It is easy to see that $\mathbb{E}_{x^t, y^t}[Z^t] = 0$, and $\|Z^t\| = \tilde{O}(d^{\frac{1}{2}})$ with high probability by (A.2) of Lemma 17. (A.2) also yields that $|v^\top Z^t| = \tilde{O}(1)$ with high probability for any fixed $v \in \mathbb{S}^{d-1}$. Finally, the norm of $\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) = y^t a^t \sigma'(w^{t\top} x^t + b^t) x^t$ is of order $\tilde{O}(d^{\frac{1}{2}})$ with high probability, and for any fixed v with $\|v\| = O(1)$, $\left(\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) \right)^\top v = \tilde{O}(1)$ with high probability. ∎

Proof of Lemma 26. First, we consider κ_1^t. We have

$$\|w^t - \eta(I - w^t w^{t\top}) \nabla_w(-y^t a^t \sigma(w^{t\top} x^t + b^t))\|^{-1}$$

$$= \left(1 + \eta^2 \|(I - w^t w^{t\top}) \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2 \right)^{-\frac{1}{2}}$$

$$\geq 1 - \frac{\eta^2}{2} \|(I - w^t w^{t\top}) \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$\geq 1 - \frac{\eta^2}{2} \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2.$$

By using this, with high probability,

$$v_1^\top w^{t+1} = v_1^\top \frac{w^t - \eta(I - w^t w^{t\top}) \nabla_w(-y^t a^t \sigma(w^{t\top} x^t + b^t))}{\|w^t - \eta(I - w^t w^{t\top}) \nabla_w(-y^t a^t \sigma(w^{t\top} x^t + b^t))\|}$$

$$\geq \kappa_1^t + \eta v_1^\top (I - w^t w^{t\top}) \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) - \frac{\kappa_1^t \eta^2}{2} \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$- \frac{\eta^3}{2} |v_1^\top (I - w^t w^{t\top}) \nabla_w(-y^t a^t \sigma(w^{t\top} x^t + b^t))| \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$\geq \kappa_1^t + \eta v_1^\top \nabla_w \mathbb{E} \left[y^t a^t \sigma(w^{t\top} x^t + b^t) \right] + v_1^\top Z^t - \frac{\kappa_1^t \eta^2}{2} \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$- \frac{\eta^3}{2} \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^3$$

$$\overset{(i)}{\geq} \kappa_1^t + \eta v_1^\top (I - w^t w^{t\top}) \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \sum_{i=p}^{q} \left[i \alpha_i \beta_{m,i} (\kappa_m^t)^{i-1} v_m \right] + \eta v_1^\top (I - w^t w^{t\top}) Z^t$$

$$- \frac{\kappa_1^t \eta^2 C_1^2 d}{2} - \frac{\eta^3 C_1^3 d^{\frac{3}{2}}}{2}$$

$$\overset{(ii)}{\geq} \kappa_1^t + \eta \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \sum_{i=p}^{q} \left[i \alpha_i \beta_{m,i} (\kappa_m^t)^{i-1} (v_1^\top v_m - \kappa_1^t \kappa_m^t) \right] - \kappa_1^t \eta^2 C_1^2 d$$

$$+ \eta v_1^\top (I - w^t w^{t\top}) Z^t,$$

where we used Lemma 27 in (i), and (ii) is due to the fact that we take $\eta \leq c_4 d^{-1}$ and $\kappa_1^t \geq d^{-\frac{1}{2}}$, and hence $\eta^3 C_1^3 d^{\frac{3}{2}} \leq \kappa_1^t \eta^2 C_1^2 d$ and $-\frac{\kappa_1^t \eta^2 C_1^2 d}{2} - \frac{\eta^3 C_1^3 d^{\frac{3}{2}}}{2} \geq -\kappa_1^t \eta^2 C_1^2 d$. Thus we obtained (B.3).

In the same way, for the lower bound on κ_m^t, we have

$$\kappa_m^{t+1} \geq \kappa_m^t + \eta v_m^\top (I - w^t w^{t\top}) \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) - \frac{|\kappa_m^t| \eta^2}{2} \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$- \frac{\eta^3}{2} |v_m^\top (I - w^t w^{t\top}) \nabla_w (-y^t a^t \sigma(w^{t\top} x^t + b^t))| \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$\geq \kappa_m^t + \frac{\eta}{\sqrt{M}} \sum_{m'=1}^{M} \sum_{i=p}^{q} \left[i \alpha_i \beta_{m',i} (\kappa_{m'}^t)^{i-1} (v_m^\top v_{m'} - \kappa_m^t \kappa_{m'}^t) \right] - \frac{|\kappa_m^t| + \eta C_1 d^{\frac{1}{2}}}{2} \eta^2 C_1^2 d$$

$$+ \eta v_m^\top (I - w^t w^{t\top}) Z^t.$$

As for the upper bound,

$$\kappa_m^{t+1} \leq \kappa_m^t + \eta v_m^\top (I - w^t w^{t\top}) \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) + \frac{|\kappa_m^t| \eta^2}{2} \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$+ \frac{\eta^3}{2} |v_m^\top (I - w^t w^{t\top}) \nabla_w (-y^t a^t \sigma(w^{t\top} x^t + b^t))| \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$\leq \kappa_m^t + \frac{\eta}{\sqrt{M}} \sum_{m'=1}^{M} \sum_{i=p}^{q} \left[i \alpha_i \beta_{m',i} (\kappa_{m'}^t)^{i-1} (v_m^\top v_{m'} - \kappa_m^t \kappa_{m'}^t) \right] + \frac{|\kappa_m^t| + \eta C_1 d^{\frac{1}{2}}}{2} \eta^2 C_1^2 d$$

$$+ \eta v_m^\top (I - w^t w^{t\top}) Z^t.$$

Finally, we check that $|\kappa_m^t - \kappa_m^{t+1}| = \tilde{O}(\eta)$. From the above argument, we have

$$|\kappa_m^{t+1} - \kappa_m^t| \leq \eta \left| v_m^\top (I - w^t w^{t\top}) \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) \right| + \frac{|\kappa_m^t| \eta^2}{2} \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2$$

$$+ \frac{\eta^3}{2} |v_m^\top (I - w^t w^{t\top}) \nabla_w (-y^t a^t \sigma(w^{t\top} x^t + b^t))| \|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2.$$

The first term is bounded by $\tilde{O}(\eta)$ because $\left| v_m^\top (I - w^t w^{t\top}) \nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t) \right| = \tilde{O}(1)$. The second term is bounded by $\tilde{O}(\eta)$ when $\eta \leq c_4 d^{-1}$, because $\|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2 = \tilde{O}(d)$. The third term is bounded by $\tilde{O}(\eta)$ when $\eta \leq c_4 d^{-1}$, because $|v_m^\top (I - w^t w^{t\top}) \nabla_w (-y^t a^t \sigma(w^{t\top} x^t + b^t))| = \tilde{O}(1)$ and $\|\nabla_w y^t a^t \sigma(w^{t\top} x^t + b^t)\|^2 = \tilde{O}(d)$. Therefore, we obtained that $|\kappa_m^t - \kappa_m^{t+1}| = \tilde{O}(\eta)$. $\blacksquare$

B.4. Phase I: Weak Recovery for One Direction

Based on Lemma 26, we analyze the stochastic gradient update of the first-layer parameters. The goal of this subsection is to prove the formal version of Lemma 6.

Lemma 6 (formal) *Suppose that $|v_{m'}^\top v_m| \leq c_4 M^{-1}$ for all $m \neq m'$ and $M \leq c_4 d^{\frac{1}{2}}$, and $\eta^t = \eta \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$. Then, with high probability, there exists some time $t_1 \leq T_{1,1} = \Theta(\eta^{-1} M^{\frac{1}{2}} d^{\frac{p-2}{2}})$ such that the following holds:*

(i) $\kappa_1^{t_1} > c_2$, and

(ii) $|\kappa_m^{t_1}| \leq 5 c_3 M^{-1}$, for all $m = 2, \cdots, M$.

We start with the following lemma, which introduces (deterministic) auxiliary sequences that upper and lower bound the stochastic updates of κ_m^t.

Lemma 28 *Suppose that $|v_{m'}^\top v_m| \leq c_4 M^{-1}$ for all $m \neq m'$ and $M \leq c_4 d^{\frac{1}{2}}$. For all $s = 0, 1, \cdots, t$, we assume that*

(a) $\kappa_1^s \leq c_2$ (only required for (i): Lower bound),

(b) $|\kappa_m^s| \leq \kappa_1^s$ for all $m = 2, \cdots, M$, and

(c) $|\kappa_m^s| \leq C_2 c_3 M^{-1}$ for all $m = 2, \cdots, M$.

Then, taking $\eta^t = \eta \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$, we have the following bounds.

(i): Lower bound of κ_1^s: *For κ_1^t, we have*

$$\kappa_1^{s+1} \geq (1 - c_2)\kappa_1^0 + \eta(1 - c_2)\sum_{s'=0}^{s} \frac{p\alpha_p \beta_{1,p}}{\sqrt{M}}(\kappa_1^{s'})^{p-1}, \tag{B.4}$$

for $s = 0, 1, \cdots, t$. Consequently, by introducing an auxiliary sequence $(P^s)_{s=0}^{t+1}$ with $P^0 = (1 - c_2)\kappa_1^0$, and

$$P^{s+1} = P^s + \eta c_2 \frac{p\alpha_p \beta_{1,p}}{\sqrt{M}}(P^s)^{p-1} \quad (s = 0, 1, \cdots, t), \tag{B.5}$$

κ_1^s is lower-bounded by P^s for all $s = 0, 1, \cdots, t+1$, with high probability.

(ii): Upper bound of $\max_{m \neq 1}|\kappa_m^s|$: *For an auxiliary sequence $(Q^s)_{s=0}^{t+1}$ with $Q^0 = (1+c_2)\max\{\max_{m \neq 1}|\kappa_m^0|, \frac{1}{2}d^{-\frac{1}{2}}\}$, and*

$$Q^{s+1} = Q^s + (1 + c_2)\frac{\eta}{\sqrt{M}}p\max_m|\alpha_p \beta_{m,i}|(Q^s)^{p-1} + c_3\frac{\eta}{M^{\frac{3}{2}}}p\alpha_p \beta_{1,p}(\kappa_1^s)^{p-1},$$

$\max_{m \neq 1}|\kappa_m^s|$ is upper-bounded by Q^s for all $s = 0, 1, \cdots, t+1$, with high probability.

Proof.

38

(i) Lower bound of κ_1^s by P^s. If $\kappa_1^t \geq \frac{1}{2}d^{-\frac{1}{2}}$, we have $\kappa_1^t \geq \eta C_1 d^{\frac{1}{2}}$ by the choice of η. If $\kappa_1^s \geq \frac{1}{2}d^{-\frac{1}{2}}$, by Lemma 26, we have

$$\kappa_1^{s+1} \geq \kappa_1^s + \frac{\eta}{\sqrt{M}} \sum_{m=1}^{M} \sum_{i=p}^{q} \left[i\alpha_i \beta_{m,i}(\kappa_m^s)^{i-1}(v_1^\top v_m - \kappa_1^s \kappa_m^s) \right] - \kappa_1^s \eta^2 C_1^2 d + \eta v_1^\top (I - w^s w^{s\top}) Z^s$$

$$\geq \kappa_1^s + \frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(1 - (\kappa_1^s)^2)(\kappa_1^s)^{p-1} + \frac{\eta}{\sqrt{M}} \sum_{m=2}^{M} \sum_{i=p}^{q} \left[i\alpha_i \beta_{m,i}(\kappa_m^s)^{i-1}(v_1^\top v_m - \kappa_1^s \kappa_m^s) \right]$$

$$- \kappa_1^s \eta^2 C_1^2 d + \eta v_1^\top (I - w^s w^{s\top}) Z^s$$

$$\geq \kappa_1^s + \frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(1 - (\kappa_1^s)^2)(\kappa_1^s)^{p-1} - q^2 \eta \sqrt{M} \max_{m \neq 1} |\kappa_m^s|^{p-1} \max_{m \neq 1} |v_1^\top v_m| \max_{m,i} |\alpha_i \beta_{m,i}|$$

$$- q^2 \eta \sqrt{M} \max_{m \neq 1} |\kappa_m^s|^p \max_{m,i} |\alpha_i \beta_{m,i}| - \kappa_1^s \eta^2 C_1^2 d + \eta v_1^\top (I - w^s w^{s\top}) Z^s \qquad \text{(B.6)}$$

$$\geq \kappa_1^s + \frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(1 - (\kappa_1^s)^2)(\kappa_1^s)^{p-1} - q^2 \eta \sqrt{M}(\kappa_1^s)^{p-1} c_4 M^{-1}$$

$$- q^2 \eta \sqrt{M}(\kappa_1^s)^{p-1} C_2 c_3 M^{-1} - \eta \kappa_1^s c_4 M^{-\frac{1}{2}} d^{-\frac{p-2}{2}} C_1^2 + \eta v_1^\top (I - w^s w^{s\top}) Z^s.$$

Note that $(\kappa_1^s)^2 \leq c_2^2 \leq \frac{1}{5}c_2$, $q^2 \eta \sqrt{M}(\kappa_1^s)^{p-1} c_4 M^{-1} \leq \frac{1}{5}c_2 \frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(\kappa_1^s)^{p-1}$, $q^2 \eta \sqrt{M}(\kappa_1^s)^{p-1} C_2 c_3 M^{-1} \leq \frac{1}{5}c_2 \frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(\kappa_1^s)^{p-1}$, and $\eta \kappa_1^s c_4 M^{-\frac{1}{2}} d^{-\frac{p-2}{2}} C_1^2 \leq \frac{1}{5}c_2 \frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(\kappa_1^s)^{p-1}$, (where we used $\kappa_1^s \geq \frac{1}{2}d^{-\frac{1}{2}}$ for the last statement). Thus, we obtained that

$$\kappa_1^{s+1} \geq \kappa_1^s + (1 - \frac{4}{5}c_2)\frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(\kappa_1^s)^{p-1} + \eta v_1^\top (I - w^s w^{s\top}) Z^s. \qquad \text{(B.7)}$$

We prove the assertion by induction. Suppose that (B.4) holds for $s = 0, \cdots, \tau$ for some $\tau \leq t$. Note that this implies $\kappa_1^s \geq (1 - c_2)\kappa_1^0$ and $\kappa_1^s \geq \frac{1}{2}d^{-\frac{1}{2}}$. By applying (B.7), we have

$$\kappa_1^{\tau+1} \geq \kappa_1^\tau + (1 - \frac{4}{5}c_2)\frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(\kappa_1^\tau)^{p-1} + \eta v_1^\top (I - w^\tau w^{\tau\top}) Z^\tau$$

$$\geq \kappa_1^0 + \sum_{s=0}^{\tau}(1 - \frac{4}{5}c_2)\frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(\kappa_1^s)^{p-1} + \sum_{s=0}^{\tau} \eta v_1^\top (I - w^s w^{s\top}) Z^s. \qquad \text{(B.8)}$$

If $\tau \leq C_2 M(\kappa_1^0)^{2-2p}$, then

$$\sum_{s=0}^{\tau} \eta v_1^\top (I - w^s w^{s\top}) Z^s \leq \eta C_1 \sqrt{\tau} \leq c_4 C_1 \kappa_1^0 \leq c_2 \kappa_1^0, \qquad \text{(B.9)}$$

with high probability. If $\tau > C_2 M(\kappa_1^0)^{2-2p}$,

$$\sum_{s=0}^{\tau} \eta v_1^\top (I - w^s w^{s\top}) Z^s \leq \eta C_1 \sqrt{\tau} \leq \eta \tau C_1 C_2^{-\frac{1}{2}} M^{-\frac{1}{2}} (\kappa_1^0)^{p-1} \leq \frac{1}{5} c_2 \frac{\eta}{\sqrt{M}} \tau p\alpha_p \beta_{1,p}((1-c_2)\kappa_1^0)^{p-1}$$

$$\leq \frac{1}{5} \sum_{s=0}^{\tau} c_2 \frac{\eta}{\sqrt{M}} p\alpha_p \beta_{1,p}(\kappa_1^s)^{p-1}, \qquad \text{(B.10)}$$

with high probability. Applying the above evaluations to (B.8), we have

$$(\text{B.8}) \geq (1 - c_2)\kappa_1^0 + \sum_{s=0}^{\tau}(1 - c_2)\frac{\eta}{\sqrt{M}}p\alpha_p\beta_{1,p}(\kappa_1^s)^{p-1}.$$

Thus the (B.4) holds also for $s = \tau + 1$. The induction proves that (B.4) holds until $s = t$.

By repeatedly using (B.5), the update of $(P^t)_{t=0}^{\tau}$ is equivalent to

$$P^{t+1} = P^0 + \sum_{s=0}^{t}\frac{\eta}{\sqrt{M}}(1 - c_2)p\alpha_p\beta_{1,p}(P^s)^{p-1}$$

By comparing this and (B.4), we conclude that κ_1^s is lower bounded by P^s for $s = 1, 2, \cdots, t + 1$.

(ii) Upper bound of $\max_{m \neq 1}|\kappa_m^s|$ **by** Q^s**:** According to Lemma 26, $|\kappa_m^{s+1} - \kappa_m^s| \leq C_1\eta$ with high probability. Thus, the sign of κ_m^{s+1} is the same as that of κ_m^s, or $|\kappa_m^{s+1}| \leq C_1\eta$. Therefore,

$$|\kappa_m^{s+1}|$$

$$\leq \max\left\{C_1\eta, \left|\kappa_m^s + \frac{\eta}{\sqrt{M}}\sum_{m=1}^{M}\sum_{i=p}^{q}\left[i\alpha_i\beta_{m,i}(\kappa_m^s)^{i-1}(v_1^\top v_m - \kappa_m^t\kappa_m^s)\right]\right.\right.$$

$$\left.\left. + \frac{|\kappa_m^s| + \eta C_1 d^{\frac{1}{2}}}{2}\eta^2 C_1^2 d + \eta v_m^\top(I - w^s w^{s\top})Z^s\right|\right\}$$

$$\leq \max\left\{C_1\eta, \left|\kappa_m^s + \frac{\eta}{\sqrt{M}}\sum_{i=p}^{q}\left[i\alpha_i\beta_{1,i}(\kappa_1^s)^{i-1}(v_m^\top v_1 - \kappa_m^s\kappa_1^s)\right]\right.\right.$$

$$\left. + \frac{\eta}{\sqrt{M}}\sum_{m'\neq 1}^{M}\sum_{i=p}^{q}\left[i\alpha_i\beta_{m',i}(\kappa_{m'}^s)^{i-1}(v_m^\top v_{m'} - \kappa_m^s\kappa_{m'}^s)\right]\right.$$

$$\left.\left. + \frac{|\kappa_m^s| + \eta C_1 d^{\frac{1}{2}}}{2}\eta^2 C_1^2 d + \eta v_m^\top(I - w^s w^{s\top})Z^s\right|\right\}$$

$$\leq \max\left\{C_1\eta, \left|\kappa_m^t + \frac{\eta}{\sqrt{M}}\sum_{i=p}^{q}i\alpha_i\beta_{1,i}(\kappa_1^s)^{i-1}v_m^\top v_1\right.\right.$$

$$\left. + \frac{\eta}{\sqrt{M}}\sum_{m'\neq 1}^{M}\sum_{i=p}^{q}\left[i\alpha_i\beta_{m',i}(\kappa_{m'}^s)^{i-1}(v_m^\top v_{m'} - \kappa_m^s\kappa_{m'}^s)\right]\right.$$

$$\left.\left. + \frac{|\kappa_m^s| + \eta C_1 d^{\frac{1}{2}}}{2}\eta^2 C_1^2 d + \eta v_m^\top(I - w^s w^{s\top})Z^s\right|\right\}.$$

We have

$$\left|\sum_{i=p}^{q}i\alpha_i\beta_{1,i}(\kappa_1^s)^{i-1}v_m^\top v_1 + \sum_{m'\neq 1}^{M}\sum_{i=p}^{q}\left[i\alpha_i\beta_{m',i}(\kappa_{m'}^s)^{i-1}(v_m^\top v_{m'} - \kappa_m^s\kappa_{m'}^s)\right]\right|$$

$$\leq \left|\sum_{i=p}^{q}i\alpha_i\beta_{1,i}(\kappa_1^s)^{i-1}v_m^\top v_1\right| + \left|\sum_{i=p}^{q}\left[i\alpha_i\beta_{m,i}(\kappa_m^s)^{i-1}(1 - (\kappa_m^s)^2)\right]\right|$$

$$+ \left| \sum_{m' \neq 1}^{M} \sum_{i=p}^{q} \left[i\alpha_i \beta_{m',i} (\kappa_{m'}^s)^{i-1} (v_m^\top v_{m'} - \kappa_m^s \kappa_{m'}^s) \right] \right|$$

$$\leq q^2 \max_i |\alpha_i \beta_{1,i}| (\kappa_1^s)^{p-1} \max_{m'} |v_1^\top v_{m'}| + p \max_{m'} |\alpha_p \beta_{m',p}| |\kappa_m^s|^{p-1} + q^2 \max_{m',i} |\alpha_i \beta_{m',i}| |\kappa_m^s|^p$$

$$+ M q^2 \max_{m',i} |\alpha_i \beta_{m',i}| \max_{m \neq 1} |\kappa_m^s|^{p-1} (\max_{m',m} |v_m^\top v_{m'}| + \max_{m' \neq 1} |\kappa_m^s|)$$

$$\leq q^2 \max_i |\alpha_i \beta_{1,i}| (\kappa_1^s)^{p-1} c_4 M^{-1} + p \alpha_p \beta_{m,p} (\kappa_m^s)^{p-1} + q^2 \max_{m,i} |\alpha_i \beta_{m,i}| (\kappa_m^s)^{p-1} C_2 c_3 M^{-1}$$

$$+ M q^2 \max_{m,i} |\alpha_i \beta_{m,i}| \max_{m \neq 1} (\kappa_m^s)^{p-1} (c_4 M^{-1} + C_2 c_3 M^{-1})$$

$$\leq (1 + \frac{1}{3} c_2) p \max_m |\alpha_p \beta_{m,p}| \max_{m \neq 1} |\kappa_m^s|^{p-1} + c_3 \frac{p \alpha_p \beta_{1,p}}{M} (\kappa_1^s)^{p-1}.$$

Also, $\frac{|\kappa_m^s| + \eta C_1 d^{\frac{1}{2}}}{2} \eta^2 C_1^2 d \leq \frac{|\kappa_m^s| + c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}} C_1 d^{\frac{1}{2}}}{2} \eta c_4 M^{-\frac{1}{2}} d^{-\frac{p-2}{2}} C_1^2 \leq \frac{1}{3} c_2 \eta p \alpha_p \beta_{m,p} \max\{|\kappa_m^s|, \frac{1}{2} d^{-\frac{1}{2}}\}^{p-1}$. Therefore, $\kappa_m^{\tau+1}$ is bounded as

$$|\kappa_m^{\tau+1}| \leq \max \left\{ C_1 \eta, \left| \kappa_m^\tau + (1 + \frac{1}{3} c_2) \frac{\eta}{\sqrt{M}} p \max_m |\alpha_p \beta_{m,p}| \max_{m \neq 1} |\kappa_m^s|^{p-1} + c_3 \frac{\eta}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (\kappa_1^\tau)^{p-1} \right. \right.$$

$$+ \frac{1}{3} c_2 \eta p \alpha_p \beta_{1,p} \max\{|\kappa_m^s|, \frac{1}{2} d^{-\frac{1}{2}}\}^{p-1} + \left. \eta v_m^\top (I - w^s w^{s\top}) Z^\tau \right| \right\}$$

$$\leq \max \left\{ C_1 \eta, \left| \kappa_m^0 + (1 + \frac{2}{3} c_2) \sum_{s=0}^{\tau} \frac{\eta}{\sqrt{M}} p \max_m |\alpha_p \beta_{m,p}| \max\{\max_{m \neq 1} |\kappa_m^s|^{p-1}, (\frac{1}{2} d^{-\frac{1}{2}})^{p-1}\} \right. \right.$$

$$+ c_3 \sum_{s=0}^{\tau} \frac{\eta}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1} + \left. \sum_{s=0}^{\tau} \eta v_m^\top (I - w^s w^{s\top}) Z^s \right| \right\}$$

$$\leq \max \left\{ C_1 \eta, \max_{m \neq 1} |\kappa_m^0| + (1 + \frac{2}{3} c_2) \sum_{s=0}^{\tau} \frac{\eta}{\sqrt{M}} p \max_m |\alpha_p \beta_{m,p}| \max\{\max_{m \neq 1} |\kappa_m^s|^{p-1}, (\frac{1}{2} d^{-\frac{1}{2}})^{p-1}\} \right.$$

$$+ c_3 \sum_{s=0}^{\tau} \frac{\eta}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1} + \max_{m \neq 1} \left| \sum_{s=0}^{\tau} \eta v_m^\top (I - w^s w^{s\top}) Z^s \right| \right\}. \tag{B.11}$$

According to the update of Q^t, $Q^t \geq \frac{1}{2} d^{-\frac{1}{2}} \geq C_1 \eta$ for all t. Moreover, by $\eta \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$,

$$\max_{m \neq 1} \left| \sum_{s=0}^{\tau} \eta v_m^\top (I - w^s w^{s\top}) Z^s \right|$$

$$\leq \eta C_1 \sqrt{\tau}$$

$$\leq \begin{cases} \frac{1}{2} c_2 d^{-\frac{1}{2}} & (\tau \leq C_2 M d^{p-1}) \\ \frac{\tau c_2 \eta}{3\sqrt{M}} p \alpha_p \beta_{1,p} (\frac{1}{2} d^{-\frac{1}{2}})^{p-1} & (\tau > C_2 M d^{p-1}) \end{cases}$$

Therefore, (B.11) is further upper bounded as

$$\max_{m \neq 1} |\kappa_m^{\tau+1}| \leq (1 + c_2) \max\{\max_{m \neq 1} |\kappa_m^0|, \frac{1}{2} d^{-\frac{1}{2}}\}$$

$$+ (1 + c_2) \sum_{s=0}^{\tau} \frac{\eta}{\sqrt{M}} p \max_m |\alpha_p \beta_{m,p}| \max\{\max_{m \neq 1} |\kappa_m^s|, Q^s\}^{p-1} + c_3 \sum_{s=0}^{\tau} \frac{\eta}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1}$$

$$\text{(B.12)}$$

On the other hand, $Q^{\tau+1}$ is written as

$$Q^{\tau+1} \leq (1 + c_2) \max\{\max_{m \neq 1} |\kappa_m^0|, \frac{1}{2} d^{-\frac{1}{2}}\} + (1 + c_2) \sum_{s=0}^{\tau} \frac{\eta}{\sqrt{M}} p \max_m |\alpha_p \beta_{m,p}| (Q^s)^{p-1}$$

$$+ c_3 \sum_{s=0}^{\tau} \frac{\eta}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1} \tag{B.13}$$

Comparing (B.12) and (B.13) with the update of Q^t, we conclude that $\max_{m \neq 1} |\kappa_m^\tau| \leq Q^\tau$ ($\tau = 0, 1, \cdots, t+1$) holds by induction. ∎

The previous lemma assumed (a)-(c). Next we show that (b) and (c) hold along the trajectory via induction. (here we use a different notation for the coefficient in (c).)

Lemma 29 *Take $\eta^t = \eta \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$. Suppose that, for all $s = 0, 1, \cdots, t$,*

(a) $\kappa_1^s \leq c_2$,

(b) $|\kappa_m^s| \leq \kappa_1^s$ for all $m = 2, \cdots, M$, and

(c)' $|\kappa_m^s| \leq 4 c_3 M^{-1}$ for all $m = 2, \cdots, M$.

Then, if we have $M \leq c_4 d^{\frac{1}{2}}$ and (a) $\kappa_1^{t+1} \leq c_2$ for $t+1$, (b) and (c)' hold for $s = t+1$ with high probability.

Proof. First consider the case when

$$(1 + c_2) \frac{p \max_m |\alpha_p \beta_{m,p}|}{\sqrt{M}} (Q^s)^{p-1} > \frac{c_3}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (P^s)^{p-1} \tag{B.14}$$

holds for all $s = 0, 1, \cdots, t$. Then, for $s = 0, 1, \cdots, t$,

$$\frac{c_3}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (P^s)^{p-1} < c_2^{-1} c_3 (1 + c_2) \frac{p \max_m |\alpha_p \beta_{m,p}|}{\sqrt{M}} (Q^s)^{p-1} < c_2 \frac{p \max_m |\alpha_p \beta_{m,p}|}{\sqrt{M}} (Q^t)^{p-1},$$

and therefore

$$Q^{s+1} \leq Q^s + (1 + 2c_2) \frac{\eta}{\sqrt{M}} p \max_m |\alpha_p \beta_{m,i}| (Q^s)^{p-1}$$

for all $s = 0, 1, \cdots, t$. According to Lemma 19, we have

$$Q^s \leq \frac{Q^0}{\left(1 - \eta p (p-2)(1 + 3c_2)(\max_m |\alpha_p \beta_{m,p}|) M^{-\frac{1}{2}} (Q^0)^{(p-2)} s\right)^{\frac{1}{p-2}}} \tag{B.15}$$

for all $s = 0, 1, \cdots, t+1$ (here $(1+c)^{p-1}$ in the original bound is absorbed in $(1 + 3c_2)$).

On the other hand, P^s is lower bounded by

$$P^s \geq \frac{P^0}{\left(1 - \eta p(p-2)(1-c_2)(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}(P^0)^{(p-2)}s\right)^{\frac{1}{p-2}}} \quad \text{(B.16)}$$

for all $s = 0, 1, \cdots, t+1$. According to (B.1), $Q^0 = (1 + c_2)\max\{\max_{m\neq 1}|\kappa_m^0|, \frac{1}{2}d^{-\frac{1}{2}}\} \leq P^0 = (1 - c_2)\kappa_1^0$. Moreover, (B.1) implies that $\max_m |\alpha_p\beta_{m,p}|(1 + 3c_2)(\max_{m\neq 1}|\kappa_m^0|)^{p-2} \leq (1-c_2)\alpha_p\beta_{1,p}(\kappa_1^0)^{p-2}$, yielding that the denominator of (B.15) is larger than that of (B.16). Therefore, $Q^{t+1} \leq P^{t+1}$ holds, which implies that $\max_{m\neq 1}|\kappa_m^{t+1}| \leq \kappa_1^{t+1}$.

Next we check $Q^{t+1} < c_3M^{-1}$ (RHS is smaller than (c)' by a factor of 4). From $P^t \leq \kappa_1^t \leq c_2$ and (B.16),

$$t \leq \frac{\eta^{-1}M^{\frac{1}{2}}((P^0)^{-(p-2)} - (c_2)^{-(p-2)})}{p(p-2)(1-c_2)(\alpha_p\beta_{1,p})} \leq \frac{\eta^{-1}M^{\frac{1}{2}}(1+5c_2)\frac{\max_m|\alpha_p\beta_{m,p}|}{\alpha_p\beta_{1,p}}(P^0)^{-(p-2)}}{p(p-2)(1+3c_2)(\max_m|\alpha_p\beta_{m,p}|)}. \quad \text{(B.17)}$$

On the other hand, according to (B.15), $Q^{t+1} > c_3M^{-1}$ holds only if

$$t > \frac{\eta^{-1}M^{\frac{1}{2}}((Q^0)^{-(p-2)} - (c_3M^{-1})^{-(p-2)} - \eta M^{-\frac{1}{2}}p(p-2)(1+3c_2)(\max_m|\alpha_p\beta_{m,p}|))}{p(p-2)(1+3c_2)(\max_m|\alpha_p\beta_{m,p}|)}. \quad \text{(B.18)}$$

If $M \leq c_4d^{\frac{1}{2}}$, $(c_3M^{-1})^{-(p-2)} \leq (c_3c_4^{-1}d^{-\frac{1}{2}})^{-(p-2)} \leq c_2(\kappa_1^0)^{-(p-2)} \leq c_2(P^0)^{-(p-2)} \leq c_2(Q^0)^{-(p-2)}$. Moreover, $\eta M^{-\frac{1}{2}}p(p-2)(1 + 5c_2)(\max_m|\alpha_p\beta_{m,p}|) \leq c_2(Q^0)^{-(p-2)}$. Thus, (B.18) is further bounded by

$$\text{(B.18)} \geq \frac{\eta^{-1}M^{\frac{1}{2}}(1-2c_2)(Q^0)^{-(p-2)}}{p(p-2)(1+3c_2)(\max_m|\alpha_p\beta_{m,p}|)}. \quad \text{(B.19)}$$

According to (B.1), $(1 + 8c_2)\max_m|\alpha_p\beta_{m,p}|(\max_{m\neq 1}|\kappa_m^0|)^{p-2} < \alpha_p\beta_{1,p}(\kappa_1^0)^{p-2}$. By using this, we have (RHS of (B.19)) > (RHS of (B.17)). Thus, $Q^{t+1} > c_3M^{-1}$ does not hold, and we have obtained $Q^{t+1} \leq c_3M^{-1}$ (and (c)').

Now we consider the case when (B.14) holds for $s = 0, 1, \cdots, \tau_1 - 1$ but

$$(1 + c_2)\frac{p\max_m|\alpha_p\beta_{m,p}|}{\sqrt{M}}(Q^s)^{p-1} \leq \frac{c_3}{M^{\frac{3}{2}}}p\alpha_p\beta_{1,p}(P^s)^{p-1} \quad \text{(B.20)}$$

holds for $s = \tau_1 \leq t$. We show that (B.20) holds for all $s = \tau_1, \cdots, t+1$ in this case. Assuming the inequality holds for all $s = \tau_1, \cdots, \tau$ with $\tau \leq t$. For $s = \tau_1, \cdots, \tau$, the update of Q^s is evaluated as

$$Q^{s+1} \leq Q^s + 2c_3\frac{\eta}{M^{\frac{3}{2}}}p\alpha_p\beta_{1,p}(P^s)^{p-1}.$$

Thus, when $p \geq 3$,

$$Q^{\tau+1} \leq Q^{\tau_1} + \sum_{s=\tau_1}^{\tau} 2c_3\frac{\eta}{M^{\frac{3}{2}}}p\alpha_p\beta_{1,p}(P^s)^{p-1}$$

$$\leq Q^{\tau_1} + \frac{2c_3(1-c_2)^{-1}}{M}\left(P^{\tau+1} - P^{\tau_1}\right) \tag{B.21}$$

$$\leq \left(\frac{\alpha_p\beta_{1,p}}{\max_m|\alpha_p\beta_{m,p}|}(1+c_2)^{-1}M^{-1}\right)^{\frac{1}{p-1}}P^{\tau_1} + \frac{2c_3(1-c_2)^{-1}}{M}\left(P^{\tau+1} - P^{\tau_1}\right)$$

$$\leq \left(\frac{\alpha_p\beta_{1,p}}{\max_m|\alpha_p\beta_{m,p}|}(1+c_2)^{-1}M^{-1}\right)^{\frac{1}{p-1}}P^{\tau+1},$$

which yields (B.20) for $s = \tau + 1$. (B.20) with $s = t + 1$ implies that $Q^{t+1} \leq P^{t+1}$, and hence $\max_{m \neq 1}|\kappa_m^{t+1}| \leq \kappa_1^{t+1}$. Similar to (B.21), we have

$$Q^{t+1} \leq Q^{\tau_1} + \frac{2c_3(1-c_2)^{-1}}{M}\left(P^{t+1} - P^{\tau_1}\right) \leq Q^{\tau_1} + \frac{2c_3(1-c_2)^{-1}}{M}P^{t+1} \leq Q^{\tau_1} + 3c_3M^{-1}.$$

For Q^{τ_1}, as we proved $Q^{t+1} \leq c_3M^{-1}$ in the first case, we have $Q^{\tau_1} \leq c_3M^{-1}$. Thus, Q^{t+1} is bounded by $4c_3M^{-1}$, which yields (c)'. ∎

Proof of Lemma 6. Suppose that (a) holds for all $s = 0, 1, \cdots, T_{1,1}$, where

$$T_{1,1} = \left\lfloor \left(\eta p(p-2)(1-5c_2)(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}(P^0)^{(p-2)}\right)^{-1}\right\rfloor.$$

According to Lemma 29, if $M \leq c_4 d^{\frac{1}{2}}$, $\eta \leq c_4 M^{-\frac{1}{2}}d^{-\frac{p}{2}}$, and (a) holds for all $s = 0, 1, \cdots, T_{1,1}$, (b) and (c) of Lemma 29 holds with high probability for all $s = 0, 1, \cdots, T_{1,1}$ and the bounds of Lemma 28 holds for all $s = 0, 1, \cdots, T_{1,1}$.

From Lemma 28 and Lemma 19,

$$\kappa_1^t \geq P^t \geq \frac{P^0}{\left(1 - \eta p(p-2)(1-c_2)(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}(P^0)^{(p-2)}s\right)^{\frac{1}{p-2}}}. \tag{B.22}$$

However, at $t = T_{1,1}$,

$$\text{(RHS of (B.22))} \geq \frac{P^0}{\left(\eta p(p-2)(1-c_2)(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}(P^0)^{(p-2)}\right)^{\frac{1}{p-2}}}$$

$$= \frac{1}{\left(\eta p(p-2)(1-c_2)(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}\right)^{\frac{1}{p-2}}},$$

and thus RHS of (B.22) is clearly larger than 1. This yields the contradiction because $\kappa_1^{T_{1,1}} \leq 1$. Therefore, with high probability, there exists some $t_1 \leq T_{1,1}$ such that $\kappa_1^{t_1} > c_3$ and $\kappa_1^t \leq c_3$ ($t = 0, 1, \cdots, t_1 - 1$).

As for (ii), recall that $|\kappa_m^{t_1-1}| \leq 4c_3 M^{-\frac{1}{2}}$. Moreover, according to Lemma 26,

$$|\kappa_1^{t_1} - \kappa_1^{t_1-1}|$$
$$\leq C_1\eta \leq C_1 c_4 M^{-\frac{1}{2}}d^{-\frac{p}{2}} \leq c_3 M^{-\frac{1}{2}}.$$

Thus, $|\kappa_m^{t_1}| \leq |\kappa_m^{t_1-1}| + |\kappa_1^{t_1} - \kappa_1^{t_1-1}| \leq 4c_3 M^{-\frac{1}{2}} + c_3 M^{-\frac{1}{2}} \leq 5c_3 M^{-\frac{1}{2}}$. ∎

B.5. Phase II: Amplification of Alignment

In the previous section (Lemma 6), we proved that neurons in J_1 achieve a small constant (c_2) alignment. However, as the alignment κ_1^t becomes larger, the effect from other directions becomes non-negligible, while the signal from v_1 gets smaller because of the projection $(1 - w^t w^{t\top})$. Hence the previous weak recovery analysis is not sufficient to show that the neurons will continue to grow in the direction of v_1.

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

Lemma 30 *Suppose $\eta = \eta^t \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$, $|v_{m'}^\top v_m| \leq c_4 M^{-1}$ for all $m \neq m'$, and $M \leq c_4 d^{\frac{1}{2}}$, and consider a neuron that satisfies (i) and (ii) of Lemma 6. Then, with high probability, there exists some time $t_2 \leq T_{1,2} = \tilde{\Theta}(M^{\frac{1}{2}} \eta^{-1})$ such that $\kappa_1^{t_2} > 1 - c_2$.*

Similar to the Phase I analysis, we bound the update by deterministic auxiliary sequences. To simplify the notation, we let $t \leftarrow t - t_1$ throughout this subsection.

Lemma 31 *Suppose that $|v_{m'}^\top v_m| \leq c_4 M^{-1}$ for all $m \neq m'$ and $M \leq c_4 d^{\frac{1}{2}}$. For all $s = 0, 1, \cdots, t$, we assume that*

(a) $\kappa_1^s \leq 1 - c_2$ (only required for (i): Lower bound),

(b) $|\kappa_m^s| \leq \kappa_1^s$ for all $m = 2, \cdots, M$, and

(c) $|\kappa_m^s| \leq C_2 c_3 M^{-1}$ for all $m = 2, \cdots, M$.

Take $\eta^t = \eta \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$. Then, we have the following bounds.

(i): **Lower bound of κ_1^s:** *For an auxiliary sequence $(P^s)_{s=0}^{t+1}$ with $P^0 = (1 - c_2)\kappa_1^0$, and*

$$P^{s+1} = P^s + c_2 \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{m,p} (P^s)^{p-1},$$

κ_1^s is lower-bounded by P^s for all $s = 0, 1, \cdots, t+1$, with high probability.

(ii): **Upper bound of $\max_{m \neq 1} |\kappa_m^s|$:** *For an auxiliary sequence $(Q^s)_{s=0}^{t+1}$ with $Q^0 = 6 c_3 M^{-1}$, and*

$$Q^{s+1} = Q^s + (1 + c_2) \frac{\eta}{\sqrt{M}} p \max_m |\alpha_p \beta_{m,i}| (Q^s)^{p-1} + c_3 \frac{\eta}{M} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1},$$

$\max_{m \neq 1} |\kappa_m^s|$ is upper-bounded by Q^s for all $s = 0, 1, \cdots, t+1$, with high probability.

Proof.

(i) Lower bound of κ_1^s by P^s: If $\kappa_1^s \geq \frac{1}{2} d^{-\frac{1}{2}}$, by following the argument for (B.6) in Lemma 28, we have

$$
\begin{aligned}
\kappa_1^{s+1} &\geq \kappa_1^s + \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (1 - (\kappa_1^s)^2)(\kappa_1^s)^{p-1} - q^2 \eta \sqrt{M} \max_{m \neq 1} |\kappa_m^s|^{p-1} \max_{m \neq 1} |v_1^\top v_m| \max_{m,i} |\alpha_i \beta_{m,i}| \\
&\quad - q^2 \eta \sqrt{M} \max_{m \neq 1} |\kappa_m^s|^p \max_{m,i} |\alpha_i \beta_{m,i}| - \kappa_1^s \eta^2 C_1^2 d + \eta v_1^\top (I - w^s w^{s\top}) Z^s \\
&\geq \kappa_1^s + \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (1 - (1 - c_2)^2)(\kappa_1^s)^{p-1} - q^2 \eta \sqrt{M} |\kappa_1^s|^{p-1} c_4 M^{-1}
\end{aligned}
$$

$$- q^2 \eta \sqrt{M} (\kappa_1^s)^{p-1} C_2 c_3 M^{-1} - \eta \kappa_1^s c_4 M^{-\frac{1}{2}} d^{-\frac{p-2}{2}} C_1^2 + \eta v_1^\top (I - w^s w^{s\top}) Z^s.$$

Note that $1 - (1 - \kappa_1^s)^2 \geq \frac{9}{5} c_2$, $q^2 \eta \sqrt{M} (\kappa_1^s)^{p-1} c_4 M^{-1} \leq \frac{1}{5} c_2 \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1}$, $q^2 \eta \sqrt{M} (\kappa_1^s)^{p-1} C_2 c_3 M^{-1} \leq$
$\frac{1}{5} c_2 \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1}$ and $\eta \kappa_1^s c_4 M^{-\frac{1}{2}} d^{-\frac{p-2}{2}} C_1^2 \leq \frac{1}{5} c_2 \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1}$ (where we used $\kappa_1^s \geq$
$\frac{1}{2} d^{-\frac{1}{2}}$ for the last statement).

If $\kappa_1^s \geq \frac{1}{2} d^{-\frac{1}{2}}$ for all $s = 0, 1, \cdots, \tau$, following (B.9) and (B.10), we know that

$$\sum_{s=0}^{\tau} \eta v_1^\top (I - w^s w^{s\top}) Z^s \leq c_2 \kappa_1^0 + \frac{1}{5} \sum_{s=0}^{\tau} c_2 \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1},$$

with high probability.

Given $\kappa_1^s \geq \frac{1}{2} d^{-\frac{1}{2}}$ for all $s = 0, 1, \cdots, \tau$, we obtain

$$\kappa_1^{\tau+1} \geq (1 - c_2) \kappa_1^0 + c_2 \sum_{s=0}^{\tau} \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (\kappa_1^s)^{p-1}, \tag{B.23}$$

and $\kappa_1^{\tau+1} \geq \frac{1}{2} d^{-\frac{1}{2}}$. Therefore, $\kappa_1^\tau \geq \frac{1}{2} d^{-\frac{1}{2}}$ holds for all $\tau = 0, 1, \cdots, t$, and (B.23) holds for all $\tau = 0, 1, \cdots, t$ By comparing (B.23) with the update of P^τ, we obtain the desired bound.

(ii) Upper bound of $\max_{m \neq 1} |\kappa_m^s|$ **by** Q^s**:** We have already established the upper bound with $Q^0 = (1 + c_2) \max\{\max_{m \neq 1} |\kappa_m^0|, \frac{1}{2} d^{-\frac{1}{2}}\}$ in Lemma 28. Because the choice of $Q^0 = 6 c_3 M^{-1}$ is larger than that when $M \leq c_4 d^{\frac{1}{2}}$, the desired bound follows. ∎

Now we show that assumptions (b) and (c) in the previous lemma can be verified along the trajectory via induction.

Lemma 32 *Take $\eta^t = \eta \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$. Suppose that, for all $s = 0, 1, \cdots, t$,*

(a) $\kappa_1^s \leq 1 - c_2$,

(b) $|\kappa_m^s| \leq \kappa_1^s$ for all $m = 2, \cdots, M$, and

(c)' $|\kappa_m^s| \leq C_2 c_3 M^{-1}$ for all $m = 2, \cdots, M$.

Then, if $M \leq c_4 d^{\frac{1}{2}}$ and (a) $\kappa_1^{t+1} \leq 1 - c_2$ for $t + 1$, (b) and (c)' hold for $s = t + 1$ with high probability.

Proof. We only need to prove (c)': $\max_{m \neq 1} |\kappa_m^{t+1}| \leq 7 c_3 M^{-1}$. This is because if (a) and (c)' for $t + 1$ hold, then $\kappa_1^0 > c_2$ and Lemma 31 yields that $\kappa_1^{t+1} \geq P^{t+1} \geq P^0 = (1 - c_2) \kappa_1^0 \geq 2 c_3 M^{-1} \geq \max_{m \neq 1} |\kappa_m^{t+1}|$, which proves (b) for $t + 1$. If

$$(1 + c_2) \frac{p \max_m |\alpha_p \beta_{m,p}|}{\sqrt{M}} (Q^s)^{p-1} \leq \frac{c_3}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (P^s)^{p-1} \tag{B.24}$$

holds for all $s = 0, 1, \cdots, \tau$, according to Lemma 31, the update of Q^s is evaluated as

$$Q^{s+1} \leq Q^s + 2 c_3 \frac{\eta}{M^{\frac{3}{2}}} p \alpha_p \beta_{1,p} (P^s)^{p-1}$$

46

for all $s = 0, 1, \cdots, \tau$. Note that (B.24) holds for $s = 0$. Thus, if (B.24) holds for all $s = 0, 1, \cdots, \tau$, we have,

$$
\begin{aligned}
Q^{\tau+1} &\leq Q^{\tau_1} + \sum_{s=\tau_1}^{\tau} 2c_3 \frac{\eta}{M^{\frac{3}{2}}} p\alpha_p\beta_{1,p}(P^s)^{p-1} \\
&\leq Q^{\tau_1} + \frac{2c_2^{-1}c_3}{M}\left(P^{\tau+1} - P^{\tau_1}\right) \\
&\leq \left(\frac{\alpha_p\beta_{1,p}}{\max_m |\alpha_p\beta_{m,p}|}(1+c_2)^{-1}M^{-1}\right)^{\frac{1}{p-1}} P^{\tau_1} + \frac{2c_2^{-1}c_3}{M}\left(P^{\tau+1} - P^{\tau_1}\right) \\
&\leq \left(\frac{\alpha_p\beta_{1,p}}{\max_m |\alpha_p\beta_{m,p}|}(1+c_2)^{-1}M^{-1}\right)^{\frac{1}{p-1}} P^{\tau+1},
\end{aligned}
\tag{B.25}
$$

which implies (B.24) for $s = \tau + 1$. Thus, (B.24) holds for all $s = 0, 1, \cdots, t + 1$.

Moreover, similar to (B.25),

$$
Q^{t+1} \leq Q^0 + \frac{2c_2^{-1}c_3}{M}\left(P^{t+1} - P^0\right) \leq Q^0 + \frac{2c_2^{-1}c_3}{M}P^{t+1} \leq Q^0 + 2C_2c_3M^{-1}.
$$

Q^0 is bounded by $6c_3M^{-1}$ Thus, Q^{t+1} is bounded by $C_2c_3M^{-1}$, which yields (c). ∎

Proof of Lemma 30.   Suppose that (a) holds for all $s = 0, 1, \cdots, T_{1,2}$, where

$$
T_{1,2} = \left\lfloor \left(\eta p(p-2)c_2(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}(P^0)^{(p-2)}\right)^{-1} \right\rfloor.
$$

According to Lemma 32, if $M \leq c_4 d^{\frac{1}{2}}$, $\eta \leq c_4 M^{-\frac{1}{2}}d^{-\frac{p}{2}}$, and (a) for all $s = 0, 1, \cdots, T_{1,2}2$ hold, (b) and (c) of Lemma 32 holds with high probability for all $s = 0, 1, \cdots, T_{1,2}$ and the bounds of Lemma 31 holds for all $s = 0, 1, \cdots, T_{1,2}$.

According to Lemma 31 and Lemma 19,

$$
\kappa_1^t \geq P^t \geq \frac{P^0}{\left(1 - \eta p(p-2)c_2(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}(P^0)^{(p-2)}s\right)^{\frac{1}{p-2}}}.
\tag{B.26}
$$

However, at $t = T_{1,2}$,

$$
\text{(RHS of (B.26))} \geq \frac{P^0}{\left(\eta p(p-2)c_2(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}(P^0)^{(p-2)}\right)^{\frac{1}{p-2}}} = \frac{1}{\left(\eta(p-2)c_2(\alpha_p\beta_{1,p})M^{-\frac{1}{2}}\right)^{\frac{1}{p-2}}},
$$

and thus RHS of (B.26) is clearly larger than 1. This yields the contradiction because $\kappa_1^{T_{1,2}}$ should be smaller than 1. Therefore, with high probability, there exists some $t_2 \leq T_{1,2}$ such that $\kappa_1^{t_2} > 1 - c_2$. ∎

B.6. Phase III: Strong Recovery and Localization

In the previous section (Lemma 30), we proved that neurons can achieve alignment of $1 - c_2$, which sets up the local convergence argument. To simplify the notation, we let $t \leftarrow t - t_1 - t_2$ throughout this subsection. We write $\bar{v}_m = (I - v_1 v_1^\top) v_m$ and $\bar{\kappa}_m^t = \bar{v}_m^\top w_j^t$.

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

Lemma 33 *Take $\eta^t = \eta_1 \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$ for $0 \leq t \leq (T_{1,1} - t_1) + (T_{1,2} - t_2) - 1$, and $\eta^t = \eta_2 \leq \min\{c_4 \tilde{\varepsilon} M^{-\frac{1}{2}} d^{-1}, c_4 \tilde{\varepsilon}^2 M^{-\frac{1}{2}}\}$ for $(T_{1,1} - t_1) + (T_{1,2} - t_2) \leq t \leq T_{1,3} + (T_{1,1} - t_1) + (T_{1,2} - t_2) - 1$. Suppose $|v_{m'}^\top v_m| \leq c_4 M^{-1}$ for all $m' \neq m$ and $T_{1,3} = \tilde{\Theta}(\tilde{\varepsilon}^{-1} M^{\frac{1}{2}} \eta^{-1})$, and consider a neuron that satisfies $\kappa_1^0 \geq 1 - c_2$.*

Then, $\kappa_1^{(T_{1,1} - t_1) + (T_{1,2} - t_2) + T_{1,3}} > 1 - 3\tilde{\varepsilon}$ with high probability.

We bound the update by deterministic auxiliary sequences.

Lemma 34 *Let $0 < \bar{\varepsilon} < c_2$. If $\eta^t = \eta \leq \min\{c_4 \bar{\varepsilon} M^{-\frac{1}{2}} d^{-1}, c_4 \bar{\varepsilon}^2 M^{-\frac{1}{2}}\}$, $1 - 2c_2 \leq \kappa_1^0$, $\kappa_1^s \leq 1 - \bar{\varepsilon}$ for $s = 0, 1, \cdots, t$, and $|v_{m'}^\top v_m| \leq c_4 M^{-1}$ for all $m' \neq m$, we have the following bound:*

- **Lower bound of κ_1^s:**

$$\kappa_1^s \geq \kappa_1^0 - c_2 \bar{\varepsilon} + s \bar{\varepsilon} \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{m,p},$$

for all $s = 0, 1, \cdots, t+1$, with high probability.

Proof. If $\kappa_1^s \geq \frac{1}{2} d^{-\frac{1}{2}}$, by Lemma 28,

$$\kappa_1^{s+1} \geq \kappa_1^s + \frac{\eta}{\sqrt{M}} \sum_{m=1}^M \sum_{i=p}^q \left[i \alpha_i \beta_{m,i} (\kappa_m^s)^{i-1} (v_1^\top v_m - \kappa_1^s \kappa_m^s) \right] - \kappa_1^s \eta^2 C_1^2 d + \eta v_1^\top (I - w^s w^{s\top}) Z^s$$

$$\geq \kappa_1^s + \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (1 - (\kappa_1^s)^2)(\kappa_1^s)^{p-1} + \frac{\eta}{\sqrt{M}} \sum_{m=2}^M \sum_{i=p}^q \left[i \alpha_i \beta_{m,i} (\kappa_m^s)^{i-1} (v_1^\top v_m - \kappa_1^s \kappa_m^s) \right]$$

$$- \kappa_1^s \eta^2 C_1^2 d + \eta v_1^\top (I - w^s w^{s\top}) Z^s$$

$$\geq \kappa_1^s + \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (1 - (\kappa_1^s)^2)(\kappa_1^s)^{p-1} + \frac{\eta}{\sqrt{M}} \sum_{m=2}^M \sum_{i=p}^q \left[i \alpha_i \beta_{m,i} (\kappa_m^s)^{i-1} v_1^\top v_m (1 - (\kappa_1^s)^2) \right]$$

$$+ \frac{\eta}{\sqrt{M}} \sum_{m=2}^M \sum_{i=p}^q \left[i \alpha_i \beta_{m,i} (\kappa_m^s)^{i-1} \bar{\kappa}_m^s \kappa_1^s \right] - \kappa_1^s \eta^2 C_1^2 d + \eta v_1^\top (I - w^s w^{s\top}) Z^s. \tag{B.27}$$

We bound each term of (B.27) from now. If $\kappa_1^s \geq 1 - 3c_2$, the second term is bounded by

$$p \alpha_p \beta_{1,p} (1 - (\kappa_1^s)^2)(\kappa_1^s)^{p-1} \geq p \alpha_p \beta_{1,p} (1 - (\kappa_1^s)^2)(\kappa_1^s)^{p-1} \geq \frac{9}{5} p \alpha_p \beta_{1,p} (1 - \kappa_1^s).$$

Next, the third term is bounded by

$$\left| \sum_{m=2}^M \sum_{i=p}^q \left[i \alpha_i \beta_{m,i} (\kappa_m^s)^{i-1} v_1^\top v_m (1 - (\kappa_1^s)^2) \right] \right| \leq q^2 M \max_{m,i} |\alpha_i \beta_{m,i}| \max_{m \neq 1} |v_1^\top v_m| (1 - (\kappa_1^s)^2)$$

$$\leq 2q^2 M \max_{m,i} |\alpha_i \beta_{m,i}| c_4 M^{-1} (1 - (\kappa_1^s)^2)$$

$$\leq \frac{1}{5} p \alpha_p \beta_{1,p} (1 - \kappa_1^s).$$

Then we consider the fourth term,

$$\left| \sum_{m=2}^{M} \sum_{i=p}^{q} \left[i \alpha_i \beta_{m,i} (\kappa_m^s)^{i-1} \tilde{\kappa}_m^s \kappa_1^s \right] \right|$$

$$\leq q^2 \max_{i,m} |\alpha_i \beta_{m,i}| \sum_{m=2}^{M} |\kappa_m^s| |\bar{\kappa}_m^s|$$

$$\leq 2q^2 \max_{i,m} |\alpha_i \beta_{m,i}| \sum_{m=2}^{M} |\bar{\kappa}_m^s|^3 + 2q^2 \max_{i,m} |\alpha_i \beta_{m,i}| \sum_{m=2}^{M} |v_1^\top v_m|^2 |\bar{\kappa}_m^s|. \tag{B.28}$$

To upper bound the above, we consider the value of $\sum_{m=2}^{M} |\bar{\kappa}_m^s|^2$. This can be represented as $\sum_{m=2}^{M} |\bar{\kappa}_m^s|^2 = w^{s\top} (I - v_1 v_1^\top)^\top A (I - v_1 v_1^\top) w^s$, where $A = \sum_{m=2}^{M} v_m v_m^\top$.

Consider $\tilde{v}_m$ defined in Lemma 40 with coefficients $\{c_{m,m'}\}$. Let us define $B \in \mathbb{R}^{M \times M}$ as

$$B_{i,j} = \begin{cases} c_{i,j} & (j \leq i \leq M) \\ 0 & \text{(otherwise)}. \end{cases}$$

Then we have

$$\begin{pmatrix} \tilde{v_1} & \cdots & \tilde{v_M} \end{pmatrix} = \begin{pmatrix} v_1 & \cdots & v_M \end{pmatrix} B^\top.$$

Since the non-diagonal terms are bounded by $c_4 M^{-1} \leq \frac{1}{3} M^{-1}$ and the absolute value of diagonal terms is no smaller than $1 - 20c_2 \geq \frac{2}{3}$, we know that $B^\top$ is invertible and (the absolute value of) the maximum eigenvalue of $(B^\top)^{-1}$ is bounded by 3. Each v_m is computed as

$$v_m = \begin{pmatrix} \tilde{v_1} & \cdots & \tilde{v_M} \end{pmatrix} (B^\top)^{-1} \mathsf{e}_m^\top$$

and

$$A = \sum_{m=1}^{M} \begin{pmatrix} \tilde{v_1} & \cdots & \tilde{v_M} \end{pmatrix} (B^\top)^{-1} \mathsf{e}_m \mathsf{e}_m^\top ((B^\top)^{-1})^\top \begin{pmatrix} \tilde{v_1}^\top \\ \vdots \\ \tilde{v_M}^\top \end{pmatrix}.$$

Thus,

$$\lambda_{\max}(A) \leq \lambda_{\max}^2 \left(\begin{pmatrix} \tilde{v_1} & \cdots & \tilde{v_M} \end{pmatrix} \right) \lambda_{\max}^2 ((B^\top)^{-1}) \lambda_{\max} ((\sum_{m=1}^{M} \mathsf{e}_m \mathsf{e}_m^\top)) \leq 9.$$

Therefore, $\sum_{m=2}^{M} |\bar{\kappa}_m^s|^2 = w^{s\top} (I - v_1 v_1^\top)^\top A (I - v_1 v_1^\top) w^s \leq 9 \|(I - v_1 v_1^\top) w^s\|^2 = 9(1 - (\kappa_1^s)^2)$. Based on this, if $\kappa_1^s \geq 1 - 3c_2$, we have

$$(\text{B.28}) \leq 54q^2 \max_{i,m} |\alpha_i \beta_{m,i}| (1 - (\kappa_1^s)^2)^{\frac{3}{2}} + 6q^2 \max_{i,m} |\alpha_i \beta_{m,i}| \max_{m \neq 1} |v_1^\top v_m|^2 (1 - (\kappa_1^s)^2)^{\frac{1}{2}}$$

49

$$\leq \frac{1}{5} p \alpha_p \beta_{1,p} (1 - \kappa_1^s).$$

The fifth term $\kappa_1^s \eta^2 C_1^2 d$ of (B.27) is bounded by $\frac{\eta}{5\sqrt{M}} p \alpha_p \beta_{1,p} (1 - \kappa_1^s)$, when $\eta \leq c_4 \bar{\varepsilon} M^{-\frac{1}{2}} d^{-1}$.

Putting things together, if $1 - 3c_2 \leq \kappa_1^s \leq 1 - \bar{\varepsilon}$ for all $s = 0, 1, \cdots, \tau$, we have

$$\kappa_1^{\tau+1} \geq \kappa_1^\tau + \frac{6}{5} \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (1 - \kappa_1^\tau) + \eta v_1^\top (I - w^\tau w^{\tau\top}) Z^\tau$$

$$\geq \kappa_1^0 + \sum_{s=0}^\tau \frac{6}{5} \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (1 - \kappa_1^s) + \eta \sum_{s=1}^\tau v_1^\top (I - w^s w^{s\top}) Z^s. \qquad \text{(B.29)}$$

Moreover, by $\eta \leq c_4 M^{-\frac{1}{2}} \bar{\varepsilon}^2$,

$$\left| \sum_{s=0}^\tau \eta v_1^\top (I - w^s w^{s\top}) Z^s \right|$$

$$\leq \eta C_1 \sqrt{\tau}$$

$$\leq \begin{cases} c_2 \bar{\varepsilon} & (\tau \leq C_2 M \bar{\varepsilon}^{-2}) \\ \dfrac{\tau \eta}{5\sqrt{M}} p \alpha_p \beta_{1,p} (1 - \kappa_1^s) & (\tau > C_2 M \bar{\varepsilon}^{-2}) \end{cases}$$

Therefore, if $1 - 3c_2 \leq \kappa_1^s \leq 1 - \bar{\varepsilon}$ for all $s = 0, 1, \cdots, \tau$, (B.29) is bounded by

$$\kappa_1^{\tau+1} \geq \kappa_1^0 - c_2 \bar{\varepsilon} + \sum_{s=0}^\tau \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} (1 - \kappa_1^s)$$

$$\geq \kappa_1^0 - c_2 \bar{\varepsilon} + \sum_{s=0}^\tau \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} \bar{\varepsilon}$$

$$\geq \kappa_1^0 - c_2 \bar{\varepsilon} + \tau \frac{\eta}{\sqrt{M}} p \alpha_p \beta_{1,p} \bar{\varepsilon}, \qquad \text{(B.30)}$$

and $1 - 3c_2 \leq \kappa_1^s$ holds for $s = \tau + 1$. By induction, $1 - 3c_2 \leq \kappa_1^s$ holds for all $s = 0, 1, \cdots, t+1$, and the bound (B.30) holds for all $\tau = 0, 1, \cdots, t$. ∎

Proof of Lemma 33. First, consider $0 \leq t \leq (T_{1,1} - t_1) + (T_{1,2} - t_2)$. By the choice of $\eta_1 \leq c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}$, $\eta \leq \min\{c_4 M^{-\frac{1}{2}} d^{-\frac{p}{2}}, c_4 \bar{\varepsilon}^2 M^{-\frac{1}{2}}\}$ is satisfied with $\bar{\varepsilon} = d^{-\frac{1}{2}}$ in Lemma 34. Thus, according to Lemma 34, until $\kappa_1^t > 1 - \bar{\varepsilon}$ holds (we let τ be the earliest time this condition holds), κ_1^t is lower bounded by $\kappa_1^0 - \bar{\varepsilon} \geq 1 - 2c_2$.

One can also see that $\kappa_1^t \geq 1 - 2c_2$ holds for all $\tau \leq t \leq (T_{1,1} - t_1) + (T_{1,2} - t_2)$. Suppose that there exists some $\tau' < (T_{1,1} - t_1) + (T_{1,2} - t_2)$ such that $\kappa_1^t < 1 - \bar{\varepsilon}$. Among such τ', we focus on the earliest time. According to Lemma 26, $|\kappa_1^{\tau'} - \kappa_1^{\tau'-1}| \leq C_1 \eta_1$, which implies that $\kappa_1^{\tau'} \geq 1 - 2\bar{\varepsilon}$. Then, According to Lemma 34, $\kappa_1^t \geq \kappa_1^{\tau'} - c_2 \geq 1 - 2c_2$ until κ_1^t gets larger than $1 - \bar{\varepsilon}$ again. By repeating this argument, we obtain that $\kappa_1^0 - \bar{\varepsilon} \geq 1 - 2c_2$ for all $0 \leq t \geq (T_{1,1} - t_1) + (T_{1,2} - t_2)$.

Then, we consider $(T_{1,1} - t_1) + (T_{1,2} - t_2) \leq t \leq (T_{1,1} - t_1) + (T_{1,2} - t_2) + T_{1,3}$. By the choice of $\eta_2 \leq \min\{c_4 \tilde{\varepsilon} M^{-\frac{1}{2}} d^{-1}, c_4 \tilde{\varepsilon}^2 M^{-\frac{1}{2}}\}$, $\eta \leq \min\{c_4 \bar{\varepsilon} M^{-\frac{1}{2}} d^{-1}, c_4 \bar{\varepsilon}^2 M^{-\frac{1}{2}}\}$ is satisfied with $\bar{\varepsilon} = \tilde{\varepsilon}$

in Lemma 34. Suppose that $\kappa_1^s \leq 1 - \tilde{\varepsilon}$ holds for all $s = (T_{1,1} - t_1) + (T_{1,2} - t_2), \cdots, (T_{1,1} - t_1) + (T_{1,2} - t_2) + T_{1,3}$, where

$$T_{1,3} = \left\lfloor 3c_2 \left(\frac{\eta}{\sqrt{M}} \tilde{\varepsilon} p \alpha_p \beta_{1,p} \right)^{-1} \right\rfloor + 1.$$

According to Lemma 34, the bound of Lemma 34 holds for all $s = (T_{1,1} - t_1) + (T_{1,2} - t_2), \cdots, (T_{1,1} - t_1) + (T_{1,2} - t_2) + T_{1,3}$:

$$\kappa_1^s \geq 1 - 2c_2 - c_2\tilde{\varepsilon} + s\tilde{\varepsilon}\frac{\eta}{\sqrt{M}} p\alpha_p\beta_{1,p} \geq 1 - 3c_2 + s\tilde{\varepsilon}\frac{\eta}{\sqrt{M}} p\alpha_p\beta_{1,p}. \qquad (\text{B.31})$$

However, at $t = (T_{1,1} - t_1) + (T_{1,2} - t_2) + T_{1,3}$,

$$(\text{RHS of (B.31)}) \geq 1,$$

and thus RHS of (B.31) is clearly larger than $1 - \tilde{\varepsilon}$. This yields the contradiction. Therefore, with high probability, there exists some $t_3 \leq (T_{1,1} - t_1) + (T_{1,2} - t_2) + T_{1,3}$ such that $\kappa_1^{t_3} > 1 - \tilde{\varepsilon}$.

Finally, suppose that there exists some $t > t_3$ such that $\kappa_1^t < 1 - \tilde{\varepsilon}$. Among such $t > t_3$, we focus on the smallest τ. According to Lemma 26, $|\kappa_1^\tau - \kappa_1^{\tau-1}| \leq C_1\eta_2$, which implies that $\kappa_1^\tau \geq 1 - 2\tilde{\varepsilon}$. Then, According to Lemma 34, $\kappa_1^t \geq 1 - 3\tilde{\varepsilon}$ until κ_1^t gets larger than $1 - \tilde{\varepsilon}$ again. By repeating this argument, we obtain that $\kappa_1^t > 1 - 3\tilde{\varepsilon}$ for all $t_3 \leq t \leq (T_{1,1} - t_1) + (T_{1,2} - t_2) + T_{1,3}$, which concludes the proof. ∎

B.7. Expressivity of the Trained Feature Map

In this section we discuss the expressivity of the feature map after first-layer training. First, we consider the approximation of single-index polynomials and show the existence of suitable second-layer parameters with small approximation error and low norm.

ReLU activation. For $\sigma = \text{ReLU}$, we have the following result.

Lemma 35 *Suppose that $b_j \sim \text{Unif}([-C_b, C_b])$ with $C_b = \tilde{O}(1)$, and consider the approximation of degree-q polynomial $h(s)$, where $q = O_d(1)$. Let $v \in \mathbb{S}^{d-1}(1)$ and $v_- = -v$. Then, there exists $a_1, \ldots, a_{2N}$ such that*

$$\sup_{t=T_1+1,\cdots,T_1+T_2} \left| \frac{1}{2N} \sum_{j=1}^N a_j\sigma(v^\top x^t + b_j) - \frac{1}{2N} \sum_{j=N+1}^{2N} a_j\sigma(v_-^\top x^t + b_j) - h(v^\top x^t) \right| = \tilde{O}(N^{-1}),$$

Moreover, we have $\sum_{j=1}^{2N} a_j^2 = \tilde{O}(N)$ and $\sum_{j=1}^{2N} |a_j| = \tilde{O}(N)$.

Proof. According to Lemma 9 of [DLS22], if $b \sim \text{Unif}([-1, 1])$ and $\delta \sim \text{Unif}(\{-1, 1\})$, for any $k \geq 0$, there exists $v_k(\delta, b)$ with $|v_k(\delta, b)| \lesssim 1$ such that for all s with $|s| \leq 1$,

$$\mathbb{E}[v_k(\delta, b)\sigma(\delta s + b)] = s^k.$$

Thus, for $C_b = \tilde{O}(1)$, if $b \sim \text{Unif}([-C_b, C_b])$ and $\delta \sim \text{Unif}(\{-1, 1\})$, there exists $\bar{v}(\delta, b; h)$ with $|\bar{v}(\delta, b; h)| = \tilde{O}(1)$ such that for all s with $|s| \leq C_b$,

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

We take $C_b = \tilde{\Theta}(1)$ sufficiently large so that for all x^t ($t = T_1 + 1, \cdots, T_1 + T_2$), $|v^\top x^t| \leq C_b$ holds, with high probability. For $A = \tilde{\Theta}(N)$, 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 $a_j = \frac{N}{2} \int_{C_b(-1+\frac{i}{A})}^{C_b(-1+\frac{i+1}{A})} \bar{v}(1, b; h) \mathrm{d}b$ for $1 \leq j \leq N$, and $a_j = \frac{N}{2} \int_{C_b(-1+\frac{i}{A})}^{C_b(-1+\frac{i+1}{A})} \bar{v}(-1, b; h) \mathrm{d}b$ otherwise. 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 component by letting $a_j = 0$. In doing so, since $\sigma(s + b)$ is 1-Lipschitz with respect to s, we have

$$\sup_{t=T_1+1,\cdots,T_1+T_2} \left| \frac{1}{2N} \sum_{j=1}^{N} a_j \sigma(v^\top x^t + b_j) - \frac{1}{2N} \sum_{j=N+1}^{2N} a_j \sigma(v_-^\top x^t + b_j) - h(v^\top x^t) \right| = \tilde{O}(N^{-1}),$$

with high probability. Thus we obtain the assertion. ∎

Polynomial Activation   If σ is a degree-q polynomial, we have the following result.

Lemma 36 *Suppose that $b_j \sim \text{Unif}([-C_b, C_b])$ with $C_b = \tilde{O}(1)$, and consider the approximation of degree-q polynomial $h(s)$, where $q = O_d(1)$. Then, there exists $a_1, \ldots, a_N$ such that*

$$\sup_{t=T_1+1,\cdots,T_1+T_2} \left| \frac{1}{N} \sum_{j=1}^{N} a_j \sigma(v^\top x^t + b_j) - h(v^\top x^t) \right| = \tilde{O}(N^{-1})$$

with high probability, where $v \in \mathbb{S}^{d-1}$. Moreover, we have $\sum_{j=1}^{N} a_j^2 = \tilde{O}(N)$ and $\sum_{j=1}^{N} |a_j| = \tilde{O}(N)$.

The lemma depends on the following result.

Lemma 37 *Suppose that $C_b \geq q$. For any polynomial $h(s)$ with 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

$$g_{q-i}(s) = \sum_{j=0}^{i} c_{i,j} \int_{b=-q}^{0} \sigma(s+b+j) \mathrm{d}b$$

$$= 2C_b \mathbb{E}_{b \sim \mathrm{Unif}([-C_b, C_b])}\left[\left(\sum_{j=0}^{i} c_{i,j} \mathbb{1}[j-q \le b \le j]\right) \sigma(s+b)\right],$$

when $C_b \ge q$. 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 36.   We now discretize Lemma 37. For $A = \tilde{\Theta}(N)$ (with a sufficiently 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. Due to 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 $|v^\top x^t| = \tilde{O}(1)$ with high probability, we have

$$\sup_{t=T_1+1,\cdots,T_1+T_2} \left| \frac{1}{N} \sum_{j=1}^{N} a_j \sigma(v^\top x^t + b_j) - h(v^\top x^t) \right| = \tilde{O}(N^{-1})$$

with high probability, which yields the assertion.   ∎

Next, by using the expressivity results above, we show that there exists some a^* that can approximate the additive target function f_*.

Lemma 38 *If $J \gtrsim J_{\min} M^{C_p} \log d$, and σ is either of the ReLU activation or any univariate polynomial with degree q, there exists some parameters $a^* = (a_j^*)_{j=1}^{J} \in \mathbb{R}^J$ such that*

$$\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \left(\frac{1}{J} \sum_{j=1}^{J} a_j^* \sigma(\hat{w}_j^\top x^t + b_j) - \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f_m(v_m^\top x^t) \right)^2 \le C_1 M(|J_{\min}|^{-2} + \tilde{\varepsilon}^2),$$

where $\|a^*\|_2^2 = \tilde{O}(J^2 |J_{\min}|^{-1})$ and $\|a^*\|_1 = \tilde{O}(J M^{\frac{1}{2}})$.

53

Proof. We only discuss the case of the ReLU activation; the result for degree-q polynomial link functions follows from the exact same analysis. Let $\tilde{J}_{m,+}$ be a set of neurons satisfying $\hat{w}_j \geq 1 - 3\tilde{\varepsilon}$ and $\tilde{J}_{m,-}$ be a set of neurons with $\hat{w}_j \leq -1 + 3\tilde{\varepsilon}$. Also, we let $\tilde{J}_m = \tilde{J}_{m,+} \cup \tilde{J}_{m,-}$. According to Lemma B, when $J \gtrsim J_{\min} M^{C_p} \log d$, we have $|\tilde{J}_{m,+}|, |\tilde{J}_{m,-}| \geq J_{\min}$ with high probability.

If $j \notin \bigcup_m \tilde{J}_m$, we set $a_j = 0$. Also, if $\tilde{J}_{m,+}$ or $\tilde{J}_{m,-}$ contains more than $J_{\min}$ neurons, we ignore the rest by simply setting $a_j = 0$. Then,

$$
\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \left(\frac{1}{J} \sum_{j=1}^{J} a_j \sigma(\hat{w}_j^\top x^t + b_j) - \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f_m(v_m^\top x^t) \right)^2
$$

$$
= \frac{1}{M} \sum_{m,m'=1}^{M} \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \left(\frac{\sqrt{M}}{J} \sum_{j \in \tilde{J}_m} a_j \sigma(\hat{w}_j^\top x^t + b_j) - f_m(v_m^\top x^t) \right)
$$

$$
\left(\frac{\sqrt{M}}{J} \sum_{j \in \tilde{J}_{m'}} a_j \sigma(\hat{w}_j^\top x^t + b_j) - f_{m'}(v_{m'}^\top x^t) \right).
$$

For each m, we evaluate $\left| \frac{\sqrt{M}}{J} \sum_{j \in \tilde{J}_m} a_j \sigma(\hat{w}_j^\top x + b_j^*) - f_m(v_m^\top x) \right|$. For the j-th neuron in $\tilde{J}_{m,+}$, by using a_j in Lemma 35 with $N = |\tilde{J}_{\min}|$, we define $a_j^* \leftarrow \frac{J}{2|J_{\min}|\sqrt{M}} a_j$; similarly for the j-th neuron in $\tilde{J}_{m,-}$, we define $a_j^* \leftarrow \frac{J}{2|\tilde{J}_{\min}|\sqrt{M}} a_{N+j}$. We obtain that

$$
\left| \frac{\sqrt{M}}{J} \sum_{j \in \tilde{J}_m} a_j^* \sigma(\hat{w}_j^\top x + b_j^*) - f_m(v_m^\top x) \right|
$$

$$
\leq \left| \frac{\sqrt{M}}{J} \sum_{j \in \tilde{J}_m} a_j^* \sigma(\delta_j v_m^\top x + b_j^*) - f_m(v_m^\top x) \right| + \left| \frac{\sqrt{M}}{J} \sum_{j \in \tilde{J}_m} (a_j^* \sigma(\hat{w}_j^\top x + b_j^*) - a_j^* \sigma(\delta_j v_m^\top x + b_j^*)) \right|
$$

$$
\leq C_1 (|J_{\min}|^{-1} + \tilde{\varepsilon}).
$$

The norm can be calculated from the construction. ∎

B.8. Fitting the Second Layer

This subsection proves the generalization error in Lemma 21, which concludes the proof of Theorem 4 together with the guarantee for the first-layer training (Lemma B).

Let $\hat{a}$ be the regularized empirical risk minimizer with L^1 or L^2 norm regularization:

$$
\hat{a} := \operatorname*{arg\,min}_{a \in \mathbb{R}^J} \hat{\mathcal{L}}(a) + \frac{\bar{\lambda}}{r} \|a\|_r^r,
$$

where $\hat{\mathcal{L}}(a) := \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2-1} (y^t - \frac{1}{J} \sum_{j=1}^{J} a_j \sigma(\hat{w}_j^\top x^t + b_j))^2$ and $r \in \{1, 2\}$. Let $f_a(x) := \frac{1}{J} \sum_{j=1}^{J} a_j \sigma(\hat{w}_j^\top x + b_j)$. Then, by the equivalence between convex regularizer and norm-constraint, there exists $\bar{\lambda} > 0$ (which can be data dependent) such that

$$
\hat{\mathcal{L}}(\hat{a}) \leq \hat{\mathcal{L}}(a^*), \quad \|a\|_r \leq \|a^*\|_r \quad (r = 1 \text{ or } 2). \tag{B.32}
$$

Indeed, let $g_B : \mathbb{R}^J \to \mathbb{R} \cup \{\infty\}$ be the indicator function of the L^p-norm ball with radius $B > 0$:
$$g_B(x) = \begin{cases} 0 & (\|x\|_r \le B) \\ \infty & (\|x\|_r > B) \end{cases}.$$ Then, the minimizer of the following norm constraint optimization problem

$$\hat{x} := \arg\min_{x \in \mathbb{R}^J} \hat{\mathcal{L}}(x) + g_B(x),$$

should satisfy $\nabla \hat{\mathcal{L}}(\hat{x}) \in -\partial g_B(\hat{x})$ where $\partial g_B(\hat{x})$ is the subgradient of g_B at $\hat{x}$. We notice that any element v in $\partial g_B(\hat{x})$ can be expressed by $v = cv'$ for some $c \ge 0$ and $v' \in \partial \|x\|_r|_{x=\hat{x}}$. This means that the minimizer $\hat{x}$ of the norm-constrained problem is also the minimizer of the regularized objective $\hat{\mathcal{L}}(x) + c\|x\|_r$. By resetting the value c as $c \leftarrow c/(p\|\hat{x}\|^{r-1})$, it is also the minimizer of the objective $\hat{\mathcal{L}}(x) + c\|x\|_r^r$.

Lemma 21 *Suppose that $J = \Theta(J_{\min} M^{C_p} \log(d))$, and σ be either of the ReLU activation or any univariate polynomial with degree q. There exists $\lambda > 0$ such that the ridge estimator $\hat{a}$ satisfies*

$$\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim M^{\frac{1}{2}}(|J_{\min}|^{-1} + \tilde{\varepsilon}) + \sqrt{\frac{M^{C_p}\log(d)}{T_2}}$$

with probability $1 - o_d(1)$. Therefore, by taking $T_2 = \tilde{\Theta}(M^{C_p}\varepsilon^{-2})$, $\tilde{\varepsilon} = \tilde{\Theta}(M^{-\frac{1}{2}}\varepsilon)$, $J_{\min} = \tilde{\Theta}(M^{\frac{1}{2}}\varepsilon^{-1})$, and $J = \tilde{\Theta}(M^{C_p+\frac{1}{2}}\varepsilon^{-1})$, we have $\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim \varepsilon$.
On the other hand, for LASSO ($r = 1$) we have

$$\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim M^{\frac{1}{2}}(|J_{\min}|^{-1} + \tilde{\varepsilon}) + \sqrt{\frac{J_{\min}^{2/s} M^{2C_p/s+1}\log(d)^{2/s}}{T_2}}$$

with probability $1 - o_d(1)$, for arbitrary $s < \infty$ (where the hidden constant may depend on s). Therefore, by taking $T_2 = \tilde{\Theta}(M^{1+\frac{2C_p+1}{s}}\varepsilon^{-2-\frac{2}{s}})$, $\tilde{\varepsilon} = \tilde{\Theta}(M^{-\frac{1}{2}}\varepsilon)$, $J_{\min} = \tilde{\Theta}(M^{\frac{1}{2}}\varepsilon^{-1})$, and $J = \tilde{\Theta}(M^{C_p+\frac{1}{2}}\varepsilon^{-1})$, we have $\mathbb{E}_x[|f_{\hat{a}}(x) - f_*(x)|] \lesssim \varepsilon$ (here we ignore polylogarithmic factors).

Proof. Let $\mathcal{F}_{a^*} := \{f_a \mid \|a\|_r \le \|a^*\|_r\}$ and P_{T_2} be the empirical distribution of the second stage: $P_{T_2} := \frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2} \delta_{x_t}$. If we choose λ as mentioned above, $\hat{a}$ satisfies the condition (B.32), which yields that $f_{\hat{a}} \in \mathcal{F}_{a^*}$. Therefore, we have that

$$\begin{aligned}
&\|f_{\hat{a}} - f_*\|_{L^1(P_x)} \\
&= \|f_{\hat{a}} - f_*\|_{L^1(P_x)} - \|f_{\hat{a}} - f_*\|_{L^1(P_{T_2})} + \|f_{\hat{a}} - f_*\|_{L^1(P_{T_2})} \\
&\le \sup_{a \in \mathbb{R}^J : \|a\| \le \|a^*\|} \left[\|f_a - f_*\|_{L^1(P_x)} - \|f_a - f_*\|_{L^1(P_{T_2})} \right] + \|f_{\hat{a}} - f_*\|_{L^2(P_{T_2})}.
\end{aligned} \quad (B.33)$$

(1) First, we bound the second term $\|f_{\hat{a}} - f_*\|_{L^2(P_{T_2})}$. Since $\hat{\mathcal{L}}(f_{\hat{a}}) \le \hat{\mathcal{L}}(f_{a^*})$, we have that

$$\|f_{\hat{a}} - f_*\|_{L^2(P_{T_2})}^2 \le \|f_{a^*} - f_*\|_{L^2(P_{T_2})}^2 + \frac{2}{T_2}\sum_{t=T_1+1}^{T_1+T_2}(f_{a^*}(x^t) - f_{\hat{a}}(x^t))\nu^t.$$

By the Hoeffding inequality, we have that

$$\frac{2}{T_2}\sum_{t=T_1+1}^{T_1+T_2}(f_{a^*}(x^t)-f_{\hat{a}}(x^t))\epsilon^t = \tilde{O}\left(\sqrt{\frac{\|f_{a^*}-f_{\hat{a}}\|_{L^2(P_{T_2})}^2}{T_2}}\right),$$

with high probability. The right hand side can be further bounded by

$$\tilde{O}\left(\sqrt{\frac{\|f_{a^*}-f_{\hat{a}}\|_{L^2(P_{T_2})}^2}{T_2}}\right) \le \tilde{O}\left(\sqrt{\frac{\|f_{a^*}-f_*\|_{L^2(P_{T_2})}^2+\|f_{\hat{a}}-f_*\|_{L^2(P_{T_2})}^2}{T_2}}\right)$$

$$\le \frac{1}{2}\|f_{\hat{a}}-f_*\|_{L^2(P_{T_2})}^2+\frac{1}{2}\|f_{a^*}-f_*\|_{L^2(P_{T_2})}^2+\tilde{O}\left(\frac{1}{T_2}\right),$$

by the Cauchy-Schwarz inequality. Then, by moving the term $\frac{1}{2}\|f_{\hat{a}}-f_*\|_{L^2(P_{T_2})}^2$ in the right hand side to the left hand side, we have that

$$\|f_{\hat{a}}-f_*\|_{L^2(P_{T_2})}^2 \le 3\|f_{a^*}-f_*\|_{L^2(P_{T_2})}^2+\tilde{O}\left(\frac{1}{T_2}\right),$$

with high probability. This also yields that

$$\|f_{\hat{a}}-f_*\|_{L^2(P_{T_2})} \le \sqrt{3}\|f_{a^*}-f_*\|_{L^2(P_{T_2})}+\tilde{O}\left(\frac{1}{\sqrt{T_2}}\right)$$

$$= \tilde{O}\left(M^{\frac{1}{2}}(|J_{\min}|^{-\frac{1}{2}}+\tilde{\varepsilon}^{\frac{1}{2}})+\frac{1}{\sqrt{T_2}}\right),$$

where we used $\|f_{a^*}-f_*\|_{L^2(P_{T_2})}=\tilde{O}(M^{\frac{1}{2}}(|J_{\min}|^{-\frac{1}{2}}+\tilde{\varepsilon}^{\frac{1}{2}}))$ by Lemma 38.

(2) The first term in (B.33) can be bounded by the standard Rademacher complexity argument (e.g., Chapter 4 of [Wai19]). Specifically, its expectation can be bounded as

$$\mathbb{E}_{(x^t)_{t=T_1+1}^{T_1+T_2}}\left[\sup_{a\in\mathbb{R}^J}\left(\|f_a-f_*\|_{L^1(P_x)}-\|f_a-f_*\|_{L^1(P_{T_2})}\right)\right] \tag{B.34}$$

$$\le 2\mathbb{E}_{(x^t,\sigma_t)_{t=T_1+1}^{T_1+T_2}}\left[\sup_{a\in\mathbb{R}^J}\left(\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}\sigma_t|f_a(x^t)-f_*(x^t)|\right)\right]$$

$$\le 4\underbrace{\mathbb{E}_{(x^t,\sigma_t)_{t=T_1+1}^{T_1+T_2}}\left[\sup_{a\in\mathbb{R}^J}\left(\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}\sigma_t f_a(x^t)\right)\right]}_{=:\mathrm{Rad}(\mathcal{F}_{a^*})} +4\mathbb{E}_{(x^t,\sigma_t)_{t=T_1+1}^{T_1+T_2}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2}\sigma_t f_*(x^t)\right], \tag{B.35}$$

where $(\sigma_t)_{t=T_1+1}^{T_1+T_2}$ is the i.i.d. Rademacher sequence which is independent of $(x^t)_{t=T_1+1}^{T_1+T_2}$ and we used the vector valued contraction inequality of the Rademacher complexity in the last inequality [Mau16]. Unfortunately, $f_*(X)$ is neither bounded nor sub-exponential, and thus we cannot naively apply the Bernstein type concentration inequality to evaluate the right hand side.

Hence, we utilize Markov's inequality instead to convert (B.34) to a high probability bound on $\sup_{a \in \mathbb{R}^J : \|a\| \leq \|a^*\|} \left[\|f_a - f_*\|_{L^1(P_x)} - \|f_a - f_*\|_{L^1(P_{T_2})} \right]$.

From Lemma 48 of [DLS22] and its proof, for either of ReLU and polynomial activation, we have

$$\text{Rad}(\mathcal{F}_{a^*}) \lesssim \sqrt{\frac{1}{T_2} \frac{\|a^*\|_r}{J}} \max_j \{ (J \mathbb{E}_x [\sigma_j (\hat{w}_j^\top x + b_j)^s])^{1/s} \},$$

for arbitrary $s \leq 1/(1 - 1/r)$. However, since $\max_j \{\hat{w}_j, b_j\} = \mathcal{O}(1)$, we have $\max_j \{\mathbb{E}_x [\sigma_j (\hat{w}_j^\top x + b_j)^s]\} = \mathcal{O}(1)$ whenever $s < \infty$ by noting that $\hat{w}_j^\top x$ is a Gaussian distribution with variance $\text{Var}(\hat{w}_j^\top x) = \mathcal{O}(1)$, which yields that the right hand side can be bounded as

$$\text{Rad}(\mathcal{F}_{a^*}) \lesssim \sqrt{\frac{1}{T_2} \frac{\|a^*\|_2}{J^{1/2}}} \quad (r = 2),$$

$$\text{Rad}(\mathcal{F}_{a^*}) \lesssim \sqrt{\frac{1}{T_2} \frac{\|a^*\|_1}{J^{1-1/s}}} \quad (r = 1),$$

where arbitrary $s < \infty$ (the hidden constant depends on s).

Applying Lemma 17 to the second term of (B.35) yields that

$$\mathbb{E} \left[\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \sigma_t f_*(x^t) \right] \leq \sqrt{\mathbb{E} \left[\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} f_*^2(x^t) \right]}$$
$$= \frac{1}{\sqrt{T_2}} \sqrt{\mathbb{E}[f_*^2(X)]} \lesssim \frac{2^{q+1}}{\sqrt{T_2}}.$$

(3) By combining evaluations of (1) and (2) together and ignoring polylogarithmic factors, we obtain that

$$\|f_{\hat{a}} - f_*\|_{L^1(P_x)} \lesssim M^{\frac{1}{2}} (|J_{\min}|^{-1} + \tilde{\varepsilon}) + \frac{1}{\sqrt{T_2}} + \sqrt{\frac{1}{T_2} \frac{\|a^*\|_r}{J^{1/2}}}. \tag{B.36}$$

We set $J = \Theta(J_{\min} M^{C_p} \log d)$. Thus, for $r = 2$, by using Theorem 38, we have

$$(\text{B.36}) \lesssim M^{\frac{1}{2}} (|J_{\min}|^{-1} + \tilde{\varepsilon}) + \frac{1}{\sqrt{T_2}} + \sqrt{\frac{1}{T_2}} J^{\frac{1}{2}} |J_{\min}|^{-\frac{1}{2}}$$
$$\lesssim M^{\frac{1}{2}} (|J_{\min}|^{-1} + \tilde{\varepsilon}) + \sqrt{\frac{M^{C_p} \log(d)}{T_2}}.$$

Thus, by setting $T_2 = \tilde{\Theta}(M^{C_p} \varepsilon^{-2})$, $\tilde{\varepsilon} = \tilde{\Theta}(M^{-\frac{1}{2}} \varepsilon)$, and $J_{\min} = \tilde{\Theta}(M^{\frac{1}{2}} \varepsilon^{-1})$, we obtain that (B.36) $\lesssim \varepsilon$.

Similarly, for $r = 1$, by Theorem 38, we have

$$(\text{B.36}) \lesssim M^{\frac{1}{2}} (|J_{\min}|^{-1} + \tilde{\varepsilon}) + \frac{1}{\sqrt{T_2}} + \sqrt{\frac{1}{T_2} \frac{J M^{\frac{1}{2}}}{J^{1-1/s}}}$$

57

$$\lesssim M^{\frac{1}{2}}(|J_{\min}|^{-1} + \tilde{\varepsilon}) + \sqrt{\frac{J_{\min}^{2/s} M^{2C_p/s+1} \log(d)^{1/s}}{T_2}}.$$

Thus, by setting $T_2 = \tilde{\Theta}(M^{1+\frac{2C_p+1}{s}} \varepsilon^{-2-\frac{2}{s}})$, $\tilde{\varepsilon} = \tilde{\Theta}(M^{-\frac{1}{2}} \varepsilon)$, and $J_{\min} = \tilde{\Theta}(M^{\frac{1}{2}} \varepsilon^{-1})$ with a sufficiently large s, we obtain that (B.36) $\lesssim \varepsilon$. ∎

Appendix C. Proof of CSQ Lower Bounds

We consider the CSQ lower bound for the following class $\mathcal{F}_{d,M,\varsigma}^{p,q}$:

$$x \sim \mathcal{N}(0, I_d), \quad y = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f_m(v_m^\top x) + \nu,$$

where $f_m = \sum_{i=p}^{q} a_{m,i} \mathsf{He}_i$ with $v_k \in \mathbb{S}^{d-1}$, $|a_{m,p}| \gtrsim 1$, $\mathbb{E}_{t \sim \mathcal{N}(0,1)}[|f_m(t)|^2] = 1$ $(m = 1, \cdots, M)$, and $\nu \sim \mathcal{N}(0, \varsigma^2)$. For the lower bound we may assume $\varsigma = 0$ since this is the easiest case for the learner.

The correlational statistical query (CSQ) returns an expectation of the correlation between y and a query function $q \colon \mathcal{X} \to \mathbb{R}$ up to an arbitrary (adversarial) error bounded by τ.

Definition 39 (Correlational statistical query) *For a function $g \colon \mathcal{X} \to \mathbb{R}$ and parameters ς, the correlational statistical query oracle $\mathrm{CSQ}(g, \varsigma, \tau)$ returns*

$$\mathbb{E}_{x,y}[yg(x)] + \nu,$$

where ν is an arbitrary noise that takes any value in $\nu \in [-\tau, \tau]$.

Without loss of generality, we assume $\|g\|_{L^2} = 1$. We prove the lower bounds on CSQ learner below.

C.1. Proof of Theorem 9(a)

We consider the following model with $\varsigma = 0$:

$$x \sim \mathcal{N}(0, I_d), \quad y = f_*(x) = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_m^\top x),$$

where $\{v_1, \cdots, v_M\}$ is a randomly sampled subset (without duplication) of the set S specified below. Also, the following lemma guarantees that when $M = \tilde{o}(d^{\frac{p}{4}})$ we have $|\mathbb{E}[y^2] - 1| = o(1)$.

Lemma 40 *For any A and d, there exists a set S of A unit vectors in $\mathbb{R}^d$ such that, for any $u, v \in S$, $u \neq v$, the inner product $|u^\top v|$ is bounded by $d^{-\frac{1}{2}} \sqrt{2 \log A}$.*

Proof. Let us sample A independent vectors $v_1, \cdots, v_A$ from the d-dimensional hypercube $\left[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}} \right]^d$. For each pair of v_i and v_j $(i \neq j)$, Hoeffding's inequality yields

$$\mathbb{P}[|v_i^\top v_j| \geq t] \leq 2e^{-t^2 d}.$$

By setting $t = d^{-1/2}\sqrt{2\log A}$, we have $|v_i^\top v_j| \leq d^{-1/2}\sqrt{2\log A}$ with probability no less than $1 - \frac{2}{A^2}$. Taking the union bound, $|v_i^\top v_j| \leq d^{-1/2}\sqrt{2\log A}$ holds for all (i,j) with probability no less than $1 - \frac{A(A-1)}{A^2} > 0$. This proves the existence of the desired S. ∎

As a result, we obtain a set of functions $\left\{ \frac{1}{\sqrt{p!}}\mathsf{He}_p(v^\top x) \mid v \in S \right\}$ with small pairwise correlation:

$$\int \frac{1}{\sqrt{p!}}\mathsf{He}_p(u^\top x) \cdot \frac{1}{\sqrt{p!}}\mathsf{He}_p(v^\top x) \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2}} \, \mathrm{d}x = (u^\top v)^p \leq d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}.$$

Based on this calculation, we know that each correlational query cannot obtain information of the true function, except for the case when the true function is in a polynomial-sized set, as shown in [Szö09].

Lemma 41 *Suppose that $\mathcal{F} = \{f_1, \cdots, f_K\}$ is a finite set of functions such that $|\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[f_i(x)f_j(x)]| \leq \varepsilon$ for all pairs of f_i and f_j with $f_i \neq f_j$. Then, for any query h satisfying $\|h\|_{L^2} \leq 1$, there are at most $\frac{2}{\tau^2 - \varepsilon}$ functions f_i that satisfy $\left|\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[f(x)h(x)]\right| \geq \tau$.*

Proof. Let

$$S_+ := \left\{ i \in [K] \mid \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[f_i(x)h(x)] > \tau \right\} \text{ and } S_- := \left\{ i \in [K] \mid \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[f_i(x)h(x)] < -\tau \right\}.$$

Then, because $\|h\| \leq 1$, Cauchy-Schwarz inequality yields

$$|S_+|^2 \tau^2 \leq \mathbb{E}\left[h(x) \sum_{i\in S_+} f_i(x) \right]^2 \leq \mathbb{E}\left[\left(\sum_{i\in S_+} f_i(x) \right)^2 \right] \lesssim |S_+| + \varepsilon(|S_+|^2 - |S_+|).$$

Therefore, we have that

$$|S_+| \leq \frac{1-\varepsilon}{\tau^2 - \varepsilon} \leq \frac{1}{\tau^2 - \varepsilon}.$$

The same argument applies to $|S_-|$. ∎

Proof of Theorem 9(a).

Consider the number of queries Q, the sequence of query $\{g_1, \cdots, g_Q\}$, and tolerance τ.

According to Lemmas 40 and 41, if

$$Q \cdot \frac{2d^C}{\tau^2 - d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}} \leq A, \tag{C.1}$$

for some $C > 0$, there exists at least $A(1 - d^{-C})$ vectors $v \in S$ such that

$$\left| \mathbb{E}\left[\frac{1}{\sqrt{p!}}\mathsf{He}_p(v^\top x)g_i(x) \right] \right| \leq \tau \quad (i = 1, \cdots, Q).$$

Now we consider the minimum value of τ to satisfy (C.1). If we take

$$A \gtrsim Qd^{\frac{p}{4}+C} \text{ and } \tau \gtrsim d^{-\frac{p}{4}}(2C\log Qd)^{\frac{p}{4}},$$

we have

$$\text{(LHS of (C.1))} \lesssim Q \cdot d^{\frac{p}{4}}(C \log Qd)^{-\frac{p}{4}} \lesssim A(C \log Qd)^{-\frac{p}{4}} \leq A,$$

which confirms (C.1).

Now consider the width-M additive model, for each query, we return the value of

$$\mathbb{E}\left[\frac{1}{\sqrt{M}} \sum_{m=1}^{M-1} \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_m^\top x) g_i(x)\right]$$

so that the learner cannot find the true v_M among $A(1 - d^{-C})$ possible directions, with probability at least $1 - d^{-C}$. Failing to do so incurs an L^2-error of $\Omega\left(\frac{1}{M}\right)$. This is because for two sets of vectors $\{v_m\}_{m=1}^M$ and $\{\tilde{v}_{\tilde{m}}\}_{\tilde{m}=1}^M$ in the set S, we have

$$\mathbb{E}\left(\frac{1}{\sqrt{M}} \sum_{m=1}^M \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_m^\top x) - \frac{1}{\sqrt{M}} \sum_{m=1}^M \frac{1}{\sqrt{p!}} \mathsf{He}_p(\tilde{v}_m^\top x)\right)^2 \geq 2 - 2\frac{1}{M} \sum_{m,\tilde{m}=1}^M (\tilde{v}_{\tilde{m}}^\top v_m)^p,$$

and $\tilde{v}_{\tilde{m}}^\top v_m = 1$ holds for at most $M - 1$ pairs and $|\tilde{v}_{\tilde{m}}^\top v_m| \leq d^{-\frac{p}{2}}(2 \log A)^{\frac{p}{2}}$ for the others if $\{v_m\}_{m=1}^M \neq \{\tilde{v}_{\tilde{m}}\}_{\tilde{m}=1}^M$. Therefore, we conclude that if $\tau \gtrsim M^{-\frac{1}{2}} d^{-\frac{p}{4}}(C \log Qd)^{\frac{p}{4}}$, the CSQ learner cannot achieve an L^2-error smaller than $O\left(\frac{1}{M}\right)$ with probability more than d^{-C}. ∎

C.2. Noisy CSQ and Proof of Theorem 9(b)

Now we prove the latter part of Theorem 9. We first explain why a lower bound with $\Omega(1)$ error cannot be achieved by naively extending the argument for Theorem 9(a).

A naive argument would go as follows. Suppose we construct some $\hat{f}(x) = \frac{1}{\sqrt{M}} \sum_{m=1}^M \frac{1}{\sqrt{p!}} \mathsf{He}_p(\hat{v}_m^\top x)$, where $\{\hat{v}_1, \cdots, \hat{v}_M\} \subset S$ using queries with tolerance τ_0. Then $\|f_*(x) - \hat{f}(x)\|_{L^2} \leq 1$ entails that a constant fraction of $\{v_1, \cdots, v_M\}$ should be identified. Therefore, we may use such a CSQ learner to learn a single-index function $\frac{1}{\sqrt{M}} \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_1^\top x)$ as follows. If we add $\frac{1}{\sqrt{M}} \sum_{m=2}^M \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_m^\top x)$ and apply the CSQ learner for $\frac{1}{\sqrt{M}} \sum_{m=1}^M \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_m^\top x)$, the learner would identify $\Omega(1)$-fraction of $\{v_1, \cdots, v_M\}$ with probability $\Omega(1)$. On the other hand, according to Theorem 9(a), learning $\frac{1}{\sqrt{M}} \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_1^\top x)$ requires $\tau \lesssim M^{-\frac{1}{2}} d^{-\frac{p}{4}}(C \log Qd)^{\frac{p}{4}} = \tilde{O}(M^{-\frac{1}{2}} d^{-\frac{p}{4}})$, with high probability. We may identify v_1 with high probability by repeating this process for $\tilde{O}(1)$ rounds; the CSQ lower bound for single-index model therefore implies that $\tau_0 \lesssim M^{-\frac{1}{2}} d^{-\frac{p}{4}}$.

The mistake in the above derivation is that, for the additive model $\frac{1}{\sqrt{M}} \sum_{m=1}^M \frac{1}{\sqrt{p!}} \mathsf{He}_p(v_m^\top x)$, since the target direction v_1 to be hidden is not known by the oracle beforehand, the (adversarial) oracle should prevent the identification of as many directions $v_1, v_2, ..., v_M$ as possible; whereas in the single-index setting $\frac{1}{\sqrt{p!}} \mathsf{He}_p(v_1^\top x)$, the oracle only need to "hide" one direction. Consequently, we cannot directly connect the identification of $\Omega(1)$-fraction of target directions in the additive model setting to the CSQ lower bound for learning single-index model.

To overcome this issue, we introduce the following sub-class of CSQ algorithms with i.i.d. noise. Because the noise is not adversarial but random, the oracle for the single index model does not use the information of the target direction, and hence the lower bound for single-index model now

implies the failure of learning $\Omega(1)$ fraction of directions. On the other hand, since the noise is no longer adversarial, our lower bound in Theorem 9(b) has weaker the query dependence compared to Theorem 9(a).

Definition 42 (Noisy CSQ) *For a function $g\colon \mathcal{X} \to \mathbb{R}$ and parameters (ς, τ), noisy correlational statistical query oracle* $\mathrm{NoisyCSQ}(g, \varsigma, \tau)$ *returns*

$$\mathbb{E}_{x,y}[yg(x)] + \nu,$$

where ν follows from the following clipped Gaussian distribution:

$$\nu = \max\{-\tau, \min\{\tilde{\nu}, \tau\}\}, \quad \tilde{\nu} \sim \mathcal{N}(0, \varsigma^2).$$

The clipping operation matches the noisy CSQ with (ς, τ) with the standard CSQ with a tolerance τ. The following theorem gives a lower bound for noisy CSQ algorithms to learn a single-index polynomial.

Theorem 43 *For any $p \geq 0$, $\varsigma > 0$, $1 > \tau > 0$, $Q > 0$, $C \gg 0$ and $M = \tilde{o}(d^{\frac{p}{4}})$, consider learning $f(x) = \frac{1}{\sqrt{p!}}\mathsf{He}_p(v^\top x)$, where v is sampled from some distribution over $\mathbb{S}^{d-1}$. Suppose that*

$$\varsigma \lesssim \frac{\tau}{\sqrt{\log Qd}}. \tag{C.2}$$

Then, for any learner using Q noisy correlational queries $\mathrm{NoisyCSQ}(g, \varsigma, \tau)$, the tolerance τ must satisfy

$$\tau \lesssim Q^{\frac{1}{2}}\sqrt{d^{-\frac{p}{2}}(\log dQ)^{\frac{p}{2}+2}}.$$

Otherwise, the learner cannot return $\hat{f}(x) = \frac{1}{\sqrt{p!}}\mathsf{He}_p(\hat{v}^\top x)$ such that $\|f(x) - \hat{f}(x)\|_{L^2} \leq 1$ with probability more than $O(d^{-C})$.

Proof. Due to the choice (C.2), we know that the clipping operation on the Gaussian noise does not make a difference with probability $1 - d^{-C}$. Thus in the following we simply consider that the pure Gaussian noise is added to the expectation. We assume that the distribution where v is sampled from is the uniform distribution over the set S consisting of A vectors, which is defined in Lemma 40. Here A is taken as

$$A \simeq Qd^{\frac{p}{2}+2},$$

so that it satisfies

$$Q \cdot \frac{2d^C}{2d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}} - d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}} \leq A.$$

As an intermediate claim, we show that if $\mathbb{E}[q_1(x)y], \cdots, \mathbb{E}[q_i(x)y]$ are bounded by $d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}$, then $\mathbb{E}[q_{i+1}(x)y]$ is also bounded by $2d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}$ with probability at least $1 - O(Q^{-1}d^{-C})$. Assume that $\mathbb{E}[q_1(x)y], \cdots, \mathbb{E}[q_i(x)y]$ are bounded by $d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}$. According to Lemma 41, for each query $q(x)$, there are at most $\frac{2}{d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}}$ vectors that has correlation larger than $d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}$.

61

Thus, there are at least $(1 - d^{-C})\frac{2Qd^C}{d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}}$ possible vectors that satisfy the assumption. Under this, use Lemma 44 with $a_{\max} = \sqrt{2d^{-\frac{p}{2}}(2\log M)^{\frac{p}{2}}}$ and $D = \Omega(\frac{Qd^C}{d^{-\frac{p}{2}}(\log A)^{\frac{p}{2}}})$. Then, when

$$\varsigma \gtrsim Q^{\frac{1}{2}}\sqrt{d^{-\frac{p}{2}}(\log dQ)^{\frac{p}{2}+1}},$$

we cannot find the desired vector with probability more than $O(D^{-1}) = O(\frac{Q^{-1}d^{-C}}{d^{\frac{p}{2}}(2\log A)^{-\frac{p}{2}}})$. This also implies that we cannot find any vector that satisfies $|\mathbb{E}[q_{i+1}(x)y]| > d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}$ with probability more than $O(Q^{-1}d^{-C})$; this is because otherwise we can select one of $\frac{2}{2d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}} - d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}}$ vectors that satisfy $|\mathbb{E}[q_{i+1}(x)y]| > d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}$ and output as the prediction of the true vector, which would succeed with probability more than $O(\frac{Q^{-1}d^{-C}}{d^{\frac{p}{2}}(2\log A)^{-\frac{p}{2}}})$.

Now, we obtain that, with probability at least $1 - O(d^{-C})$, $\mathbb{E}[q_1(x)y], \cdots, \mathbb{E}[q_Q(x)y]$ are all bounded by $d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}$. Again, there are at least $(1 - d^{-C})\frac{2Qd^C}{d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}}$ possible vectors that satisfy the conditions $\mathbb{E}[q_1(x)y], \cdots, \mathbb{E}[q_Q(x)y] \le d^{-\frac{p}{2}}(2\log A)^{\frac{p}{2}}$. Under this, we apply Lemma 44 with the same $a_{\max}$ and D^{-1} as previously, and hence we cannot identify the right vector with probability more than $O(D^{-1}) = O(\frac{Q^{-1}d^{-C}}{d^{\frac{p}{2}}(2\log A)^{-\frac{p}{2}}}) \lesssim O(d^{-C})$.

Therefore, we cannot return the correct vector in A with probability more than $O(d^{-C})$. ∎

Lemma 44 *Let $A_i = (a_{i,1}, \cdots, a_{i,Q}) \in \mathbb{R}^Q$ $(i = 1, \cdots, D)$ be sequences of Q real values satisfying $|a_{i,j}| \le a_{\max}$. Suppose that one of A_i is uniformly randomly chosen an observation $(b_1, \cdots, b_Q) \in \mathbb{R}^Q$ is generated as*

$$b_j \sim \mathcal{N}(a_{i,j}, \varsigma^2) \quad (j = 1, \cdots, Q).$$

Then, if

$$\varsigma \gtrsim Q^{\frac{1}{2}}a_{\max}\sqrt{\log D}, \tag{C.3}$$

any algorithm cannot identify which A_i is selected with probability more than $1 - O(D^{-1})$.

Proof. The optimal strategy is to calculate the likelihood function and select the one with which the index i takes the largest value. Let the likelihood function of the i-th sequence be $p_i(b)$ for $B \in \mathbb{R}^Q$. We aim to bound the success probability by $O(D^{-1})$.

$$\int \frac{1}{D}\max_i p_i(B)\mathrm{d}B \lesssim \frac{1}{D}. \tag{C.4}$$

To simplify the discussion, we assume $\varsigma = 1$ (because scaling does not affect whether the statement holds). Then, (C.3) implies $a_{\max}^2 \le 2Q^{-1}$. We have

$$\text{(C.4)} \times D \le \frac{1}{(2\pi)^{\frac{Q}{2}}}\int \exp\left(-\frac{\|B\|^2}{2} + \max_i B^\top A_i\right)\mathrm{d}B. \tag{C.5}$$

We bound the expectation of $\exp(\max_i B^\top A_i)$ conditioned on $\|B\|$. By Hoeffding's inequality, $\max_i B^\top A_i \leq a_{\max}\|B\|\sqrt{\log \delta^{-1}}$ with probability at least $1 - \delta$. Thus,

$$
\begin{aligned}
\mathbb{E}_{B\sim\mathbb{S}^{Q-1}(\|B\|)}[\exp(\max_i B^\top A_i)] &\leq \sum_{i=1}^{\infty} \frac{1}{2^i} \exp\left(a_{\max}\|B\|\sqrt{\log(D2^i)}\right) \\
&\lesssim \int_{t=1}^{\infty} \exp\left(a_{\max}\|B\|\sqrt{\log(D2^t)} - t\log 2\right)\mathrm{d}t \\
&\lesssim e^{a_{\max}\|B\|\sqrt{\log D}} \int_{t=1}^{\infty} \exp\left(a_{\max}\|B\|\sqrt{t\log 2} - t\log 2\right)\mathrm{d}t \\
&\lesssim a_{\max}\|B\| \exp\left(\frac{a_{\max}^2\|B\|^2}{4} + a_{\max}\|B\|\sqrt{\log D}\right) + 1.
\end{aligned}
$$

Applying this to (C.5) yields

$$
\text{(C.5)} \lesssim \frac{1}{(2\pi)^{\frac{Q}{2}}} \int a_{\max}\|B\| \exp\left(-\frac{(1-Q^{-1})\|B\|^2}{2} + a_{\max}\|B\|\sqrt{\log D}\right)\mathrm{d}B + 1.
$$

Here we used $a_{\max}^2 \leq 2Q^{-1}$. We bound the first term as follows:

$$
\begin{aligned}
&\frac{1}{(2\pi)^{\frac{Q}{2}}} \int a_{\max}\|B\| \exp\left(-\frac{(1-Q^{-1})\|B\|^2}{2} + a_{\max}\|B\|\sqrt{\log D}\right)\mathrm{d}B \\
&= \frac{(1-Q^{-1})^{-\frac{Q+1}{2}}}{(2\pi)^{\frac{Q}{2}}} \int a_{\max}\|B'\| \exp\left(-\frac{\|B'\|^2}{2} + \frac{a_{\max}\|B'\|\sqrt{\log D}}{\sqrt{1-Q^{-1}}}\right)\mathrm{d}B' \quad (B' = (1-Q^{-1})^{\frac{1}{2}}B) \\
&= \frac{2a_{\max}(1-Q^{-1})^{-\frac{Q+1}{2}}}{2^{\frac{Q}{2}}\Gamma(\frac{Q}{2})} \int_{s=0}^{\infty} e^{-\frac{s^2}{2} + \frac{a_{\max}s\sqrt{\log D}}{\sqrt{1-Q^{-1}}}} s^Q \mathrm{d}s \quad (s = \|B'\|) \\
&= \frac{a_{\max}2^{\frac{Q+1}{2}}(1-Q^{-1})^{-\frac{Q+1}{2}}}{2^{\frac{Q}{2}}\Gamma(\frac{Q}{2})} \left[\frac{\sqrt{2}a_{\max}\sqrt{\log D}}{\sqrt{1-Q^{-1}}}\Gamma\left(\frac{Q}{2}+1\right){}_1F_1\left(\frac{Q}{2}+1;\frac{3}{2};\frac{a_{\max}^2\log D}{2(1-Q^{-1})}\right)\right. \\
&\qquad\qquad\qquad\qquad \left.+ \Gamma\left(\frac{Q+1}{2}\right){}_1F_1\left(\frac{Q+1}{2};\frac{1}{2};\frac{a_{\max}^2\log D}{2(1-Q^{-1})}\right)\right], \quad \text{(C.6)}
\end{aligned}
$$

where ${}_1F_1(x_1; x_2; x_3)$ is the confluent hypergeometric function of the first kind defined as

$$
{}_1F_1(x_1; x_2; x_3) = \sum_{n=0}^{\infty} \frac{x_1(x_1+1)\cdots(x_1+n-1)}{x_2(x_2+1)\cdots(x_2+n-1)}\frac{x_3^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!}\prod_{i=0}^{n-1}\frac{(x_1+i)x_3}{x_2+i}.
$$

We can evaluate ${}_1F_1\left(\frac{Q}{2}+1;\frac{3}{2};\frac{a_{\max}^2\log D}{2(1-Q^{-1})}\right)$ as

$$
{}_1F_1\left(\frac{Q}{2}+1:\frac{3}{2};\frac{a_{\max}^2\log D}{2(1-Q^{-1})}\right) = \sum_{n=0}^{\infty}\frac{1}{n!}\prod_{i=0}^{n-1}\frac{(\frac{Q}{2}+1+i)\frac{a_{\max}^2\log D}{2(1-Q^{-1})}}{\frac{3}{2}+i} \lesssim 1,
$$

if $\left(\frac{Q}{2}+1\right)\frac{a_{\max}^2 \log D}{2(1-Q^{-1})} \leq \frac{3}{2} \Leftrightarrow 1 \gtrsim Qa_{\max}^2 \log D$ holds. In the same way, we have $_1F_1\left(\frac{Q+1}{2}; \frac{1}{2}; \frac{a_{\max}^2 \log D}{2(1-Q^{-1})}\right) \lesssim$ 1 if $1 \gtrsim Qa_{\max}^2 \log D$ holds. Also, $(1-Q^{-1})^{-\frac{Q+1}{2}} \lesssim 1$. Thus, we have

$$\text{(C.6)} \lesssim \frac{a_{\max}2^{\frac{Q+1}{2}}}{2^{\frac{Q}{2}}\Gamma(\frac{Q}{2})}\left[a_{\max}\sqrt{\log D}\Gamma\left(\frac{Q}{2}+1\right) + \Gamma\left(\frac{Q+1}{2}\right)\right]$$

$$\leq \frac{a_{\max}2^{\frac{1}{2}}}{\Gamma(\frac{Q}{2})}\left[a_{\max}\sqrt{\log D}\Gamma\left(\frac{Q}{2}+1\right) + \Gamma\left(\frac{Q+1}{2}\right)\right]$$

$$\leq a_{\max}^2\sqrt{2}\sqrt{\log D}\left(\frac{Q}{2}+1\right) + a_{\max}\sqrt{2}\left(\frac{Q}{2}\right)^{\frac{1}{2}}$$

$$\lesssim 1$$

where we used $\Gamma(x + \frac{1}{2}) = \int_{t=0}^\infty e^{-t}t^{x+\frac{1}{2}}dt \leq \left(\int_{t=0}^\infty e^{-t}t^x dt\right)^{\frac{1}{2}}\left(\int_{t=0}^\infty e^{-t}t^{x-1}dt\right)^{\frac{1}{2}} \leq (\Gamma(x + 1))^{\frac{1}{2}}(\Gamma(x))^{\frac{1}{2}} = x^{\frac{1}{2}}\Gamma(x)$ (by Hölder's inequality; this argument is borrowed from [Qi10]) and $a_{\max}Q^{\frac{1}{2}}\sqrt{\log D} \leq 1$. Therefore, we have successfully obtained (C.4) and the assertion follows. ∎

Proof. [Proof of Theorem 9(b)] Consider learning the following model

$$x \sim \mathcal{N}(0, I_d), \quad y = f_*(x) = \frac{1}{\sqrt{M}}\sum_{m=1}^M \frac{1}{\sqrt{p!}}\mathsf{He}_p(v_m^\top x),$$

where $\{v_1\cdots, v_M\}$ is a randomly sampled subset (without duplication) of the set S used in the proof of Lemma 43. Recall that Lemma 40 guarantees that when $M = \tilde{o}(d^{\frac{p}{4}})$ we have $|\mathbb{E}[y^2] - 1| = o(1)$.

According to Theorem 43, for any learner using Q noisy correlational queries with parameters (ς, τ), to learn a univariate polynomial $\frac{1}{\sqrt{p!M}}\mathsf{He}_p(v^\top x)$, the tolerance must satisfy

$$\tau \lesssim \frac{Q^{\frac{1}{2}}(\log dQ)^{\frac{p}{4}+1}}{M^{\frac{1}{2}}d^{\frac{p}{4}}},$$

otherwise, the learning will fail with probability more than $1 - O(d^{-C})$.

If an algorithm learns $\mathcal{F} \subset \mathcal{F}_{d,M,\varsigma}^{p,q}$ with Q noisy correlation queries and returns a function with L^2-error smaller than 1, we know that the algorithm need to identify at least $\Omega(1)$-fraction of directions $\{v_1, \cdots, v_M\}$. If so, we can use such a learner to solve the single-index polynomials, by adding $M - 1$ random functions $\frac{1}{\sqrt{p!}}\mathsf{He}_p(v_m^\top x)$ to the given single-index polynomial and then apply the algorithm. The lower bound therefore follows from the single-index CSQ lower bound stated in Theorem Theorem 43. ∎

Appendix D. Proof of SQ Lower Bound

This section considers the SQ lower bound for $\mathcal{F}_{d,M,\varsigma}^{p,q}$. The statistical query oracle is formally defined as follows, which covers the previous CSQ as a special case.

Definition 45 *For a function $g\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a tolerance $\tau > 0$, statistical query oracle* $\mathrm{SQ}(g, \tau)$ *returns any value in*

$$[\mathbb{E}_{x,y}[g(x,y)] - \tau, \mathbb{E}_{x,y}[g(x,y)] + \tau].$$

In the following, we assume bounded queries $\mathrm{SQ}\colon \mathbb{R}^d \times \mathbb{R} \to [-1, 1]$.

As mentioned in the main text, one motivation of our consideration of SQ learner is the existence of efficient SQ algorithms that can solve multi-index regression beyond the CSQ complexity. Specifically, the algorithm proposed in [CM20] learns single-index polynomials (i.e., the case of $K = 1$) with sample complexity $\tilde{O}(d)$ for any constants $p, q = O_d(1)$; this result can also include the multi-index model up to $K = O(1)$, under a certain non-degeneracy condition. In the first stage of their algorithm, the labels y are transformed so that the information exponent is reduced to 2, which enables a warm start; after that, projected gradient descent exponentially converges to the relevant directions. Although [CM20] only considered the noiseless case (i.e., $\varsigma = 0$), it is easy to extend their strategy to the noisy setting. Specifically, the warm-start algorithm can handle label noise with standard concentration arguments, and for the second stage, $\tilde{O}(d)$ sample complexity is also obtained despite the loss of exponential convergence.

However, Theorem 11 suggests that such linear-in-d complexity is no longer feasible for SQ learners to learn our additive model class. Specifically, our lower bound implies that for $M \asymp d^\gamma$ with $\gamma > 0$, the exponent in the dimension dependence can be made arbitrarily large by varying $p, q = O_d(1)$. This highlights the fundamental computational hardness of the larger M setting.

D.1. Superorthogonal Polynomials

Theorem 11 relies on the existence of *superorthogonal* polynomials defined in Lemma 13. Recall that superorthogonality means that a polynomial and its 2-, $\cdots$, K-th exponentiations are orthogonal, in a sense of inner product with respect to the standard Gaussian, to any polynomials up to degree L whose expectation is 0. For $K = 1$, $f(x) = \mathrm{He}_{L+1}(x)$ satisfies the condition. However, for $K \geq 2$, it is far from trivial that such a function exists. We defer the proof of Proposition 13 to Appendices D.2 and D.3, and proceed to explain how Proposition 13 is used in the SQ lower bound.

We utilize the following fact that the expectation $\mathbb{E}[q(x, y)]$ can be Taylor expanded with respect to a small perturbation of y, due to the Gaussian noise added to y.

Lemma 46 *Suppose that $|g(x, y)| \leq 1$ for any $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. Then, for $\delta \ll 1$, we have*

$$\mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\varsigma^2)}[g(x, z + \delta + \varepsilon)] = \sum_{k=0}^{K} a_k(x, z)\delta^k + O(\delta^{K+1}),$$

where $a_k(x, z) = \frac{1}{k!} \int g(x, w)\left(\frac{\mathrm{d}^k}{\mathrm{d}z^k} e^{-(w-z)^2/2\varsigma^2}\right)\mathrm{d}w = O(1)$.

Proof. The proof follows from the change-of-variables in integration. Specifically, by letting $w = z + \delta + \varepsilon$,

$$\mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\varsigma^2)}[g(x, z + \delta + \varepsilon)]$$
$$= \int g(x, z + \delta + \varepsilon)e^{-\varepsilon^2/2\varsigma^2}\mathrm{d}\varepsilon$$

$$= \int g(x,w)e^{-(w-z-\delta)^2/2\varsigma^2}\mathrm{d}w$$

$$= \sum_{k=0}^{K} \underbrace{\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}z^k}\int g(x,w)e^{-(w-z)^2/2\varsigma^2}\mathrm{d}w}_{=:a_k(x,z)} \cdot \delta^i$$

$$+ \frac{1}{(K+1)!}\frac{\mathrm{d}^{K+1}}{\mathrm{d}z^{K+1}}\int g(x,w)e^{-(w-z)^2/2\varsigma^2}\mathrm{d}w\bigg|_{z=z'} \cdot \delta^{K+1} \text{ (for some } z')$$

$$= \sum_{k=0}^{K}\int g(x,w)a_k(x,z)\delta^k + O(\delta^{K+1}).$$

Note that each $a_i(x,z)$ is $O(1)$ because $|g(x,w)| \le 1$. ∎

Proof of Theorem 11. Suppose for sake of contradiction $\tau \gtrsim d^\rho$. Let us take $A = e^{\sqrt{d}}$ in Lemma 40. Then, we have a set of unit vectors $S \subseteq \mathbb{S}^{d-1}$ such that two distinct vectors have an inner product at most $O(d^{-\frac{1}{4}})$. To construct the target function, we randomly draw $\{v_m\}_{m=1}^M$ from S and let $f_m = f$ for $m = 1, \cdots, M$, where f is a superorthogonal polynomial from Proposition 13 with $K \ge \frac{2\rho}{\gamma} + 2$ and $L \ge 4(\gamma + \rho) - 1$. In addition, we construct a different target function $\frac{1}{\sqrt{M}}\sum_{i=1}^M f({v'_m}^\top x)$ with $\{v'_m\}_{m=1}^M$ in the same fashion.

For each m, we prove that the following holds for at least $1 - \tilde{O}(e^{-\sqrt{d}}(M^2 d^{2\rho}))$ fraction of random choices of v_m, v'_m:

$$\mathbb{E}\left[g\left(x, \frac{1}{\sqrt{M}}\sum_{m'=1}^{m-1}f({v'_{m'}}^\top x) + \frac{1}{\sqrt{M}}\sum_{m'=m}^{M}f({v'_{m'}}^\top x) + \varepsilon\right)\right]$$

$$= \mathbb{E}\left[g\left(x, \frac{1}{\sqrt{M}}\sum_{m'=1}^{m}f({v'_{m'}}^\top x) + \frac{1}{\sqrt{M}}\sum_{m'=m+1}^{M}f({v'_{m'}}^\top x) + \varepsilon\right)\right] + o(M^{-1}d^{-\rho}). \qquad \text{(D.1)}$$

This is to say, when M is large, swapping one single-index task results in small change in the query value. To see this, from Lemma 46, we have

$$\mathbb{E}_\varepsilon\left[q\left(x, \frac{1}{\sqrt{M}}\sum_{m'=1}^{m-1}f({v'_{m'}}^\top x) + \frac{1}{\sqrt{M}}\sum_{m'=m}^{M}f(v_{m'}^\top x) + \varepsilon\right)\right]$$

$$= \mathbb{E}_\varepsilon\left[q\left(x, \frac{1}{\sqrt{M}}\sum_{m'=1}^{m-1}f({v'_{m'}}^\top x) + \frac{1}{\sqrt{M}}\sum_{m'=m+1}^{M}f(v_{m'}^\top x) + \varepsilon\right)\right]$$

$$+ \sum_{k=1}^{K}a_k\left(x, \frac{1}{\sqrt{M}}\sum_{m'=1}^{m-1}f({v'_{m'}}^\top x) + \frac{1}{\sqrt{M}}\sum_{m'=m+1}^{M}f(v_{m'}^\top x)\right)\frac{f^k(v_m^\top x)}{M^{\frac{i}{2}}} + o(M^{-\frac{K}{2}}). \qquad \text{(D.2)}$$

Note that $|\int f(v_i^\top x)f(v_j^\top x)e^{-\|x\|^2/2}\mathrm{d}x| \lesssim d^{-(L+1)/4} \lesssim M^{-1}d^{-\rho}$ if $v_i \ne v_j$. Now from Lemma 41, for each k, the number of v_m that satisfy

$$\left|\mathbb{E}\left[a_k\left(x, \frac{1}{\sqrt{M}}\sum_{m'=1}^{m-1}f({v'_{m'}}^\top x) + \frac{1}{\sqrt{M}}\sum_{m'=m+1}^{M}f(v_{m'}^\top x)\right)\frac{f^k(v_m^\top x)}{M^{\frac{i}{2}}}\right]\right| \ge 2M^{-1}d^{-\rho},$$

is at most $O(M^2 d^{2\rho})$. Thus, except for $O(KM^2 d^{2\rho})$ choices of v_m, taking expectation of (D.2) yields

$$
(\text{D.2}) = \mathbb{E}_\varepsilon \left[g\left(x, \frac{1}{\sqrt{M}} \sum_{m'=1}^{m-1} f(v'_{m'}{}^\top x) + \frac{1}{\sqrt{M}} \sum_{m'=m+1}^{M} f(v_{m'}^\top x) + \varepsilon \right) \right]
$$
$$
+ \sum_{k=1}^{K} \mathbb{E}\left[a_k\left(x, \frac{1}{\sqrt{M}} \sum_{m'=1}^{m-1} f(v'_{m'}{}^\top x) + \frac{1}{\sqrt{M}} \sum_{m'=m+1}^{M} f(v_{m'}^\top x) \right) \right] \mathbb{E}\frac{\left[f^k(v_m^\top x) \right]}{M^{\frac{i}{2}}} + o(M^{-\frac{K}{2}}).
$$
$$(\text{D.3})$$

Thus in similar fashion,

$$
(\text{D.3}) = \mathbb{E}\left[g\left(x, \frac{1}{\sqrt{M}} \sum_{m'=1}^{m} f(v_{m'}^\top x) + \frac{1}{\sqrt{M}} \sum_{m'=m+1}^{M} f(v'_{m'}{}^\top x) + \varepsilon \right) \right] + o(M^{-1} d^{-\rho}),
$$

which yields (D.1). By recursively applying (D.1), we have

$$
\mathbb{E}\left[q\left(x, \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f(v_m^\top x) + \varepsilon \right) \right] = \mathbb{E}\left[q\left(x, \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f(v'_m{}^\top x) + \varepsilon \right) \right] + o(d^{-\rho})
$$

for $(e^{\sqrt{d}} - \tilde{O}(M^2 d^{2\rho}))^M \geq e^{\sqrt{d}M}(1 - e^{-\Omega(\sqrt{d})})$ choices of $v_1, \cdots, v_M$.

Therefore, we may return the value of $\mathbb{E}\left[g(x, \frac{1}{\sqrt{M}} \sum_{m=1}^{M} f(\bar{v}_m^\top x) + \varepsilon) \right]$ for a specific choice of $\{\bar{v}_m\}_{m=1}^{M}$ fixed *a priori*, so that each query only removes at most $e^{-\Omega(\sqrt{d})}$ fraction of the possible choices of $v_1, \cdots, v_M$, but gives no information about the remaining directions. Thus, with probability at least $1 - Q e^{-\Omega(\sqrt{d})}$, $e^{\sqrt{d}M}(1 - Q e^{-\Omega(\sqrt{d})})$ possible choices of $v_1, \cdots, v_M$ are equally likely. On the other hand, for each choice of $\frac{1}{\sqrt{M}} \sum_{m=1}^{M} f(v_m^\top x)$, there are at most $e^{\sqrt{d}M/2}$ possible choices of $f(v'_1{}^\top), \cdots, f(v'_M{}^\top)$ if the L^2-error between $\frac{1}{\sqrt{M}} \sum_{m=1}^{M} f(v_m^\top x)$ and $\frac{1}{\sqrt{M}} \sum_{m=1}^{M} f(v'_m{}^\top x)$ is less than 1; so in order to output a function with small $O(1)$ error, we need to isolate one of the $O(e^{\sqrt{d}M})$ possible hypotheses. This completes the proof of Theorem 11. ∎

D.2. Reparameterization of Polynomials

The proof of Proposition 13 requires several new techniques. We begin by introducing an auxiliary class of polynomials $\{h_a(x)\}$ parameterized by $a = (a_{k,l}) \in [-1, 1]^{L \times K}$ that satisfies the three properties below. This is to avoid adjusting coefficients of $f(x)$ directly, because solving $\int (f(x))^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x = 0$ $(1 \leq k \leq K, 1 \leq l \leq L)$ as simultaneous high-order equations of the coefficients would be difficult.

(P1) $h_a(x) \not\equiv 0$.

(P2) For every $1 \leq k \leq K, 1 \leq l \leq L$, $\int (h_a(x))^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x$ is continuous with respect to a.

(P3) $\int (h_a(x))^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x > 0$ holds if $a_{k,l} = 1$, and $\int (h_a(x))^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x < 0$ holds if $a_{k,l} = -1$.
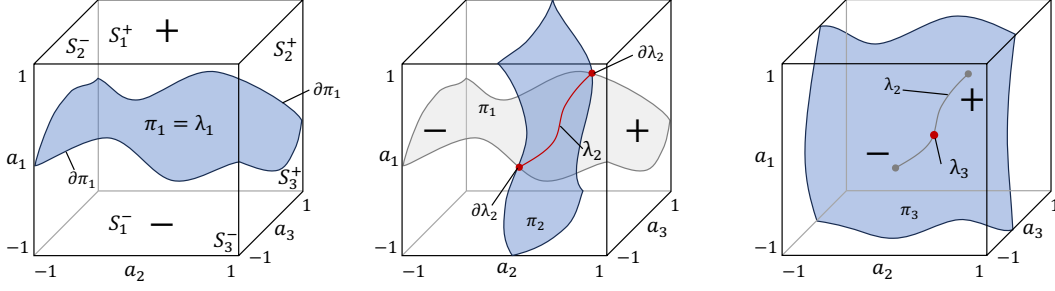
Figure 2: Illustration of the proof for $I = 3$. For each $i = 1, 2, 3$, a 2-dimensional curved surface π_i on which $A_i(a) = 0$ divides the hypercube. First, we take $\lambda_1 = \pi_1$. Then, we take the intersection between λ_1 and π_2, which is a curved line and connects one of its boundary on S_3^+ and the other in S_3^-. Finally, we consider the intersection of λ_2 and π_3. Because λ_2 connects the points in S_3^+ and S_3^- while π_3 divides the hypercube into the part containing S_3^+ and the one containing S_3^-, $\lambda_3 = \lambda_2 \cap \pi_3$ is not an empty set and $A_1(a) = A_2(a) = A_3(a) = 0$ holds on λ_3.

The following lemma shows that these three properties entails the existence of a desired superorthogonal polynomial.

Lemma 47 *If $\{h_a(x)\}$ satisfies (P1)-(P3), there exists some coefficient a such that $\int (h_a(x))^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x = 0$ holds for every $1 \le k \le K$ and $1 \le l \le L$ but $\int (h_a(x))^2 e^{-x^2/2} \mathrm{d}x > 0$.*

Proof. As handling two subscripts k, l can be notation-heavy, we prove the following restated claim. For $I \in \mathbb{N}$, we consider a vector $a \in [-1, 1]^I$ and functions $A_i \colon \mathbb{R}^I \to \mathbb{R}$, satisfying

(P2)' For every i, $A_i(a)$ is continuous with respect to a, and

(P3)' $A_i(a) > 0$ holds if $a_i = 1$, and $A_i(a) < 0$ holds if $a_i = -1$,

and prove that there exists some $a \in [-1, 1]^I$ such that $A_i(a) = 0$ for all i. Regarding as (k, l) and A_i as $\int (h_a(x))^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x$, this is equivalent to the assertion in the above lemma.

We name each surface of the hypercube $[-1, 1]^I$ as follows: the surface with $a_i = 1$ is denoted as S_i^+ and the surface with $a_i = -1$ as S_i^-. Because of (P2)' and (P3)', the hypercube $[-1, 1]^I$ is divided into two parts by the curved surface $\pi_i \subseteq [-1, 1]^I$, on which $A_i(a) = 0$ holds, and the surface is homeomorphic to $[-1, 1]^{I-1}$. Clearly, one part contains S_i^+ and the other contains S_i^-.

Now, we inductively see that there exists $\lambda_i \subset [-1, 1]^I$ homeomorphic to $[-1, 1]^{I-i}$ on which $A_j(a) = 0$ holds. This is true for $i = 1$, by taking $\lambda_1 = \pi_1$. When this is true for i, and the boundary of λ_i is on all surfaces of the hypercube except for $S_1^+, \cdots, S_i^+$ and $S_1^-, \cdots, S_i^-$, we can take $\lambda_{i+1} \subseteq \lambda_i \cap \pi_{i+1}$, on which $A_j(a) = 0$ holds for $j = 1, \cdots, i + 1$, which is homeomorphic to $[-1, 1]^{I-(i+1)}$, and the boundary of which is on all surfaces of the hypercube except for $S_1^+, \cdots, S_{i+1}^+$ and $S_1^-, \cdots, S_{i+1}^-$. Therefore by induction, we obtain $\lambda_I \subseteq \lambda_{I-1} \cap \pi_I$, which contains at least one point and on which $A_j(a) = 0$ holds for $j = 1, \cdots, I$. See Figure 2 for illustration of the case where $I = 3$. ∎

Our next goal is to construct $\{h_a(x)\}$ that satisfy the above properties. (P1) and (P2) are easily checked in the following construction. To meet (P3), we introduce an auxiliary class of functions $\{h_a^*(x)\}$ and will later approximate it using polynomials in Section D.3.

We first provide a high-level sketch on the construction of the auxiliary class $\{h_a^*(x)\}$ using Figure 3. In particular, we need to adjust the following value to satisfy (P3)

$$\int (f(x))^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x. \tag{D.4}$$

First, we fix the exponent k and consider to make the value (D.4) all 0 for $l = 1, 2, \cdots, L$. Directly adjusting coefficients of $f(x)$ by regarding (D.4) as simultaneous high-order polynomials of the coefficients of $f(x)$ would be difficult. Instead, we re-parameterize the problem as an almost linear simultaneous equation with respect to the new parameters.

Specifically, we consider a piecewise linear function defined as follows. First, we focus on (D.4) for a specific k by considering a $\{0, 1\}$-valued function (See Figure 3(a)). We divide (an interval of) the real line into equal intervals, which are indexed by i_1 and i_2, with width ε, and for each interval, we assign a value of 1 to the left half and assign the remaining portion a value of 0. Then, we consider the value of (D.4), which is the expectation of a multiplication of this function and a Hermite polynomial $\mathsf{He}_l(x)$. If the interval gets small, this is approximately equal to taking expectations of a multiplication between a constant function and the Hermite polynomial. In other words, the value of (D.4) gets closer to 0 as the interval gets small smaller.

To modify the integral value for a specific l, we move the right end of each interval beginning from x_i, proportionally to the value of $\mathsf{He}_l(x_i)$. If we move each right ends $O(a_l \varepsilon \mathsf{He}_l(x_i))$ (a_l is a scalar), then (D.4) for the specific l changes almost *linearly* with respect to a_l, while (D.4) for the other l remains (approximately) the same. In this way, we fine-tune the value of (D.4) for each l separately by changing the parameter a_l.

Next, we consider how to simultaneously address different exponentiations $k = 1, \ldots, K$. We divide each interval of width ε into K different sub-intervals indexed by $1, \ldots, K$ (see Figure 3(b)). For the j-th sub-interval, we let the height of the indicator function to be $\frac{j}{K}$. If we change the length of the j-th sub-interval proportionally to $\varepsilon \alpha_{i_1,i_2,j}$ $(j = 1, 2, \cdots, K)$, this is approximately equivalent to setting the value of each $(f(x))^k$ around x_{i_1,i_2} proportionally to

$$\left[V \cdot \begin{pmatrix} \alpha_{i_1,i_2,1} \\ \alpha_{i_1,i_2,2} \\ \vdots \\ \alpha_{i_1,i_2,K} \end{pmatrix} \right]_k, \tag{D.5}$$

where $V \in \mathbb{R}^{K \times K}$ is a variant of the Vandermonde matrix defined as follows:

$$V := \begin{bmatrix} \frac{1}{K} & \frac{2}{K} & \cdots & 1 \\ (\frac{1}{K})^2 & (\frac{2}{K})^2 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ (\frac{1}{K})^K & (\frac{2}{K})^K & \cdots & 1 \end{bmatrix}.$$

The following lemma implies that we can change each coordinate of $V \cdot (\alpha_{i_1,i_2,1}, \alpha_{i_1,i_2,2}, \vdots, \alpha_{i_1,i_2,K})^\top$ in (D.5) arbitrarily by adjusting the values of $\alpha_{i_1,i_2,1}, \cdots, \alpha_{i_1,i_2,K}$. In this way, we can control the contribution of $(f(x))^k$ to the integral values separately for different k at each interval.

(a) Approximating an integral value with indicator functions
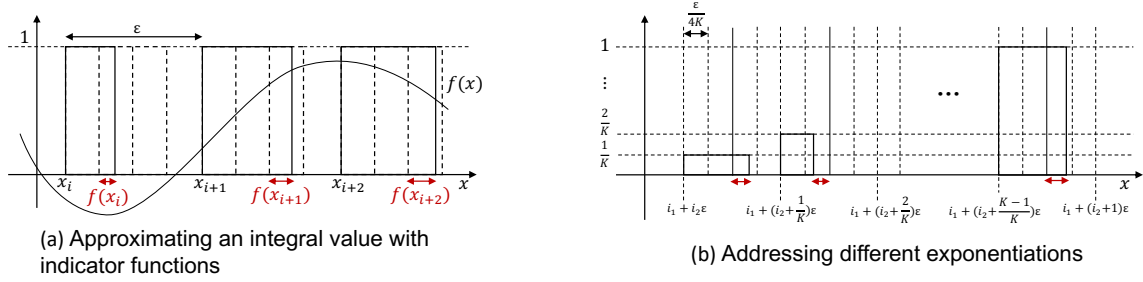
(b) Addressing different exponentiations

Figure 3: Approximation via piecewise constant function. Figure 3(a): shifting the right end of each indicator function proportionally to $f(x_i)$ is approximately equivalent to subtracting $O(f(x))$ from $(h_a^*(x))^k$ in the sense of integral value. Figure 3(b): By considering the staircase function, we can simultaneously modify the contribution of the different exponents of $h_a^*(x)$ to the integral.

Putting it all together, we can adjust the integral values of (D.4) separately for each k and l to satisfy (P3). After obtaining a function class $\{h_a^*(x)\}$ sketched above, approximating indicator functions by polynomials yields the desired class $\{h_a(x)\}$ that satisfies (P1)-(P3) in Section D.3.

Lemma 48 *The matrix V is invertible.*

Proof. Let

$$
\tilde{V}_x := \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{K+1} \\ (x_1)^2 & (x_2)^2 & \cdots & (x_{K+1})^2 \\ \vdots & \vdots & & \vdots \\ (x_1)^K & (x_2)^K & \cdots & (x_{K+1})^K \end{bmatrix}.
\tag{D.6}
$$

Because of the formula for the determinant of the block matrix, we have

$$
\det(\tilde{V}_x) = (-1)^n \cdot \det\left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_K \\ (x_1)^2 & (x_2)^2 & \cdots & (x_K)^2 \\ \vdots & \vdots & & \vdots \\ (x_1)^K & (x_2)^K & \cdots & (x_K)^K \end{bmatrix} - 1 \cdot \begin{pmatrix} x_{K+1} \\ \vdots \\ x_{K+1}^K \end{pmatrix} (1, \cdots, 1) \right).
$$

It is a well-known fact on the Vandermonde matrix that $\det(\tilde{V}_x) \neq 0$ when $x_i \neq x_j$ for all $i \neq j$. Let $x_1 = \frac{1}{K}, x_2 = \frac{2}{K}, \cdots, x_K = 1, x_{K+1} = 0$. Then, the LHS of (D.6) is $\det(\tilde{V}_x) \neq 0$ and the RHS of (D.6) is equal to $(-1)^n \det(V)$. Therefore, we have obtained that $\det(V) \neq 0$. ∎

Now we formalize the above proof sketch. We let

$$
M_1 = \max_{a \in [-1,1]^K} \left\| V^{-1} a \right\|_\infty
$$

and define the function class $\{h_a^*(x)\}$.

70

Definition 49 (An auxiliary class $\{h_a^*(x)\}$) *Let $\mathbb{1}_{s,t}(x)$ $(s < t)$ be an indicator function satisfying $\mathbb{1}_{s,t}(x) = 1$ for $s \leq x \leq t$ and $= 0$ otherwise. Fix $K, L, A_1,$ and A_2, and let $\varepsilon := \frac{1}{A_2}$. We define a class of functions $\{h_\alpha^\dagger(x)\}$, parameterized by $\alpha = (\alpha_{i_1,i_2,j}) \in [-1, 1]^{2A_1 \times A_2 \times K}$,*

$$h_\alpha^\dagger(x) := \sum_{i_1=-A_1}^{A_1-1} \sum_{i_2=0}^{A_2-1} \sum_{j=1}^{K} \frac{j}{K} \cdot \mathbb{1}_{i_1+i_2\varepsilon+\frac{4(j-1)}{4K}\varepsilon, i_1+i_2\varepsilon+\frac{4(j-1)+2+\alpha_{i_1,i_2,j}}{4K}\varepsilon}(x). \tag{D.7}$$

Then, we construct a map from $a \in \mathbb{R}^{K \times L}$ to α as follows. For each i_1, i_2, we define $(\alpha(a)_{i_1,i_2,j})_{j=1}^{K}$ as

$$\begin{pmatrix} \alpha(a)_{i_1,i_2,1} \\ \alpha(a)_{i_1,i_2,2} \\ \vdots \\ \alpha(a)_{i_1,i_2,K} \end{pmatrix} := \frac{1}{M_1 M_2} V^{-1} \begin{pmatrix} \sum_{l=1}^{L} a_{1,l} \mathsf{He}_l(i_1 + i_2\varepsilon) \\ \sum_{l=1}^{L} a_{2,l} \mathsf{He}_l(i_1 + i_2\varepsilon) \\ \vdots \\ \sum_{l=1}^{L} a_{K,l} \mathsf{He}_l(i_1 + i_2\varepsilon) \end{pmatrix},$$

where $M_2 := \max_{1 \leq l \leq L} \max_{-A_1 \leq x \leq A_1} |\mathsf{He}_l(x)|$. Based on this, we define $h_a^*(x)$ by

$$h_a^*(x) := h_{\alpha(a)}^\dagger(x)$$

From the definitions of M_1, M_2, we have $\|\alpha(a)\|_\infty \leq 1$. Thus in (D.7), each interval of the indicator function does not overlap with the others and the right end is contained in $[i_1 + i_2\varepsilon + \frac{4(j-1)}{4K}\varepsilon, i_1 + i_2\varepsilon + \frac{4(j-1)+2}{4K}\varepsilon]$.

Lemma 50 *There exist constants A_1 and A_2 such that $h^*(a)$ defined in Theorem 49 satisfy the property (P3). Specifically, when $a_{i,j} = 1$, we have*

$$\int (h_a^*(x))^i \mathsf{He}_j(x) e^{-x^2/2} \mathrm{d}x > 0,$$

and when $a_{i,j} = -1$, we have

$$\int (h_a^*(x))^i \mathsf{He}_j(x) e^{-x^2/2} \mathrm{d}x < 0,$$

for all $1 \leq i \leq L$ and $1 \leq j \leq K$.

Proof. First, we decompose the integral as follows:

$$\int (h_a^*(x))^k \mathsf{He}_l e^{-x^2/2} \mathrm{d}x = \sum_{i_1=-A_1}^{A_1-1} \sum_{i_2=0}^{A_2-1} \sum_{j=1}^{K} \left(\frac{j}{K}\right)^k \int_{i_1+i_2\varepsilon+\frac{4(j-1)\varepsilon}{4K}}^{i_1+i_2\varepsilon+\frac{(4(j-1)+2+\alpha_{i_1,i_2,j})\varepsilon}{4K}} \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x$$

$$= \sum_{i_1=-A_1}^{A_1-1} \sum_{i_2=0}^{A_2-1} \sum_{j=1}^{K} \left(\frac{j}{K}\right)^k \int_{i_1+i_2\varepsilon+\frac{4(j-1)\varepsilon}{4K}}^{i_1+i_2\varepsilon+\frac{(4(j-1)+2)\varepsilon}{4K}} \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x \tag{D.8}$$

$$+ \sum_{i_1=-A_1}^{A_1-1} \sum_{i_2=0}^{A_2-1} \sum_{j=1}^{K} \left(\frac{j}{K}\right)^k \int_{i_1+i_2\varepsilon+\frac{4(j-1)+2\varepsilon}{4K}}^{i_1+i_2\varepsilon+\frac{(4(j-1)+2+\alpha_{i_1,i_2,j})\varepsilon}{4K}} \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x$$

$$\tag{D.9}$$

71

For the first term, we have

$$
\begin{aligned}
\text{(D.8)} &= \sum_{j=1}^{K} \frac{2}{4K}\left(\frac{j}{K}\right)^{k} \sum_{i_1=-A_1}^{A_1-1} \sum_{i_2=0}^{A_2-1} \int_{i_1+i_2\varepsilon}^{i_1+(i_2+1)\varepsilon} \mathsf{He}_l(x)e^{-x^2/2}\mathrm{d}x + O(A_2^{-1}) \\
&= \sum_{j=1}^{K} \frac{2}{4K}\left(\frac{j}{K}\right)^{k} \int_{-A_1}^{A_1} \mathsf{He}_l(x)e^{-x^2/2}\mathrm{d}x + O(A_2^{-1}) \\
&= O(A_2^{-1}; A_1)O(e^{-A_1^2/2}) + O(A_2^{-1}), \tag{D.10}
\end{aligned}
$$

where $O(A_2^{-1}; A_1)$ is the big-O notation that treats A_1 as a constant. Moreover, by the definition of $\alpha(a)$,

$$
\begin{aligned}
\text{(D.9)} &= \sum_{i_1=-A_1}^{A_1-1} \sum_{i_2=0}^{A_2-1} \int_{i_1+i_2\varepsilon}^{i_1+(i_2+1)\varepsilon} \sum_{j=1}^{K}\left(\frac{j}{K}\right)^{k} \frac{\alpha_{i_1,i_2,j}}{4K} \mathsf{He}_l(x)e^{-x^2/2}\mathrm{d}x + O(A_2^{-1}; A_1) \\
&= \sum_{i_1=-A_1}^{A_1-1} \sum_{i_2=0}^{A_2-1} \frac{1}{M_1 M_2} \int_{i_1+i_2\varepsilon}^{i_1+(i_2+1)\varepsilon} \sum_{j=1}^{L} a_{k,j}\mathsf{He}_j(i_1+i_2\varepsilon)\mathsf{He}_l(x)e^{-x^2/2}\mathrm{d}x + O(A_2^{-1}; A_1) \\
&= \sum_{i_1=-A_1}^{A_1-1} \sum_{i_2=0}^{A_2-1} \frac{1}{M_1 M_2} \int_{i_1+i_2\varepsilon}^{i_1+(i_2+1)\varepsilon} \sum_{j=1}^{L} a_{k,j}\mathsf{He}_j(x)\mathsf{He}_l(x)e^{-x^2/2}\mathrm{d}x + O(A_2^{-1}; A_1) \\
&= \frac{a_{k,l}}{M_1 M_2} \int_{-A_1}^{A_1} \mathsf{He}_l^2(x)e^{-x^2/2}\mathrm{d}x + O(A_2^{-1}; A_1) \\
&= \frac{a_{k,l}}{M_1 M_2} \int \mathsf{He}_l^2(x)e^{-x^2/2}\mathrm{d}x + O(A_2^{-1}; A_1) + O(e^{-A_1^2/2}) \tag{D.11}
\end{aligned}
$$

Now, (D.10) and (D.11) yield

$$
\int (h_a^*(x))^k \mathsf{He}_l e^{-x^2/2}\mathrm{d}x = \frac{a_{k,l}}{M_1 M_2} \int \mathsf{He}_l^2(x)e^{-x^2/2}\mathrm{d}x + O(A_2^{-1}; A_1) + O(e^{-A_1^2/2}).
$$

Since $\frac{1}{M_1 M_2}\int \mathsf{He}_l^2(x)e^{-x^2/2}\mathrm{d}x > 0$, by taking A_1 sufficiently large and then taking A_2 sufficiently large, we obtain the assertion. $\blacksquare$

D.3. Polynomial Approximation of h_a^*

Finally, we consider the polynomial approximation of h_a^*, which can be reduced to polynomial approximation of each of the indicator functions. Approximation of the step/sign/indicator functions has been studied since the nineteenth century [Zol77; Akh90; EY07]. Among them, [EY07] considered the polynomial approximation of $\mathrm{sgn}(x)$ and proved the following result: the approximation error $L_m(\delta)$ with degree-m polynomials, in the interval of $[-1, -\delta] \cup [\delta, 1]$, satisfies

$$
\lim_{m\to\infty} \sqrt{m}\left(\frac{1+\delta}{1-\delta}\right)^{m} L_m(\delta) = \frac{1-\delta}{\sqrt{\pi\delta}}.
$$

This entails that, until m becomes larger than δ^{-1}, the error drops proportionally to $O(m^{-1/2}\delta^{-1/2})$. After that, the error exponentially decreases, proportionally to $e^{-\delta m}$. However, this result is not

directly applicable to our Gaussian setting, since the error bound is for a fixed interval. Also, the coefficients of the polynomial are not characterized (hence higher-order polynomials could become larger outside of the interval). Consequently, increasing the order of polynomials may not give smaller approximation error in expectation.

Therefore, we instead make use of the following fact of Hermite expansion.

Proposition 51 (Theorems 3 and 6 of [Muc70]) *Let* $U(x) = e^{-x^2/2}(1+|x|)^b$ *and* $V(x) = e^{-x^2/2}(1+|x|)^B$, *where* $b < 0$, $B \geq -2/3$, *and* $b \leq B - 1/3$. $s_n(x; f)$ *denotes the* n-*th partial sum of a Hermite series for the target function* f. *If*

$$\int_{-\infty}^{\infty} |f(x)| V(x)(1 + \log^+ |x| + \log^+ |f(x)|) < \infty,$$

then we have that

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} |s_n(x; f) - f(x)|^p U(x)^p \mathrm{d}x = 0.$$

We can therefore approximate $\mathbb{1}_{0,\infty}$ with an arbitrary accuracy with respect to the integral values.

Lemma 52 *There exists a sequence of polynomials* $\{p_n\}$ *such that*

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} |p_n(x) - \mathbb{1}_{0,\infty}(x)|^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x = 0$$

holds for all $1 \leq k \leq K$ *and* $1 \leq l \leq L$.

Proof. Let us take $b = -1$ and $B = 0$ in Proposition 51. Because $\mathbb{1}_{0,\infty}(x)$ is bounded, it is easy to see that $\int_{-\infty}^{\infty} \mathbb{1}_{0,\infty}(x) V(x)(1 + \log^+ |x| + \log^+ |\mathbb{1}_{0,\infty}(x)|) < \infty$ holds. We have that

$$0 = \lim_{n \to \infty} \int_{-\infty}^{\infty} |s_n(x; \mathbb{1}_{0,\infty}) - \mathbb{1}_{0,\infty}(x)|^K U(x)^K \mathrm{d}x$$

$$\geq \lim_{n \to \infty} \int_{-\infty}^{\infty} |s_n(x/(K+1); \mathbb{1}_{0,\infty}) - \mathbb{1}_{0,\infty}(x)|^K \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x.$$

Therefore, we arrive at a sequence of polynomials $g_n(x) := s_n(x/(K+1); \mathbb{1}_{0,\infty}(\cdot \times (K+1))$ such that

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} |s_n(x/(K+1); \mathbb{1}_{0,\infty}) - \mathbb{1}_{0,\infty}(x)|^k \mathsf{He}_l(x) e^{-x^2/2} \mathrm{d}x = 0$$

holds for all $1 \leq k \leq K$ and $1 \leq l \leq L$. ∎

Now we have obtained some p_n with which we can approximate each of the indicator functions that consists of h_a^* up to arbitrary accuracy. Each function h_a^* is not identically zero, and the integral value of h_a is continuous with respect to x. Therefore, there exists a class of polynomials $\{h_a(x)\}$ that satisfies (P1)-(P3).