Proceedings of Machine Learning Research vol 247:1–75, 2024 37th Annual Conference on Learning Theory

Pruning is Optimal for Learning Sparse Features in High-Dimensions

Nuri Mert Vural VURAL@CS.TORONTO.EDU
Department of Computer Science at University of Toronto, and Vector Institute

Murat A. Erdogdu ERDOGDU@CS.TORONTO.EDU
Department of Computer Science, and of Statistical Sciences at University of Toronto, and Vector Institute

Editors: Shipra Agrawal and Aaron Roth

Abstract

While it is commonly observed in practice that pruning networks to a certain level of sparsity can improve the quality of the features, a theoretical explanation of this phenomenon remains elusive. In this work, we investigate this by demonstrating that a broad class of statistical models can be optimally learned using pruned neural networks trained with gradient descent, in high-dimensions.

We consider learning both single-index and multi-index models of the form $y = \sigma^*(\boldsymbol{V}^\top \boldsymbol{x}) + \epsilon$, where σ^* is a degree-p polynomial, and $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ with $r \ll d$, is the matrix containing relevant model directions. We assume that $\boldsymbol{V}$ satisfies a certain ℓ_q-sparsity condition for matrices and show that pruning neural networks proportional to the sparsity level of $\boldsymbol{V}$ improves their sample complexity compared to unpruned networks. Furthermore, we establish Correlational Statistical Query (CSQ) lower bounds in this setting, which take the sparsity level of $\boldsymbol{V}$ into account. We show that if the sparsity level of $\boldsymbol{V}$ exceeds a certain threshold, training pruned networks with a gradient descent algorithm achieves the sample complexity suggested by the CSQ lower bound. In the same scenario, however, our results imply that basis-independent methods such as models trained via standard gradient descent initialized with rotationally invariant random weights can provably achieve only suboptimal sample complexity.

1. Introduction

Neural network pruning, a technique aimed at reducing the number of weights by selectively removing certain connections or neurons, has attracted significant attention in recent years as a means to improve efficiency and scalability in deep learning (LeCun et al., 1989; Hassibi and Stork, 1992; Han et al., 2015; Frankle and Carbin, 2019). Beyond the computational advantages offered by pruning, empirical observations demonstrate that this method can also substantially improve the generalization performance of neural networks (Bartoldson et al., 2020; Jin et al., 2022).

Deep learning has challenged the classical learning theory and demonstrated that overparameterization will oftentimes improve generalization. In stark contrast, however, pruning overparametrized networks is also known to improve generalization, as observed in many empirical studies (LeCun et al., 1989; Hassibi and Stork, 1992; Bartoldson et al., 2020; Jin et al., 2022). In this context, our understanding of the effect of pruning remains elusive. As such, we focus on the following question:

Does pruning improve the quality of trained features in neural networks?

We answer this question in the affirmative. Indeed, we show that when the statistical model satisfies a certain sparsity condition, pruned neural networks trained with gradient descent can achieve optimal sample complexity, and learn significantly more efficiently compared to unpruned networks.

Feature learning in neural networks has been the focus of many recent works. A key characteristic in these models is their ability to learn low-dimensional latent features (Yehudai and Shamir, 2019; Ghorbani et al., 2020; Mousavi-Hosseini et al., 2023a). An apt scenario for studying this capability is the task of learning multi-index models (Damian et al., 2022; Mousavi-Hosseini et al., 2023a), where the response $y \in \mathbb{R}$ depends on the input $\boldsymbol{x} \in \mathbb{R}^d$ via the relationship $y = \sigma^*(\boldsymbol{V}^\top \boldsymbol{x}) + \epsilon$. Here, $\sigma^* : \mathbb{R}^r \to \mathbb{R}$ is the non-linear link function, and the matrix $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ contains the relevant *model directions*. Our main focus is the regime where there are few relevant directions when compared to the ambient input dimension, i.e. $r \ll d$. In the special case $r = 1$, this model also covers the single-index setting, which has been studied extensively; see e.g. Ba et al. (2022); Mousavi-Hosseini et al. (2023a); Arous et al. (2021); Damian et al. (2023) and the references therein. In the simplified single-index case, the sample complexity of learning the model direction is determined by the *information exponent* $k^\star$ of the link function σ^*, which is defined as the smallest order nonzero Hermite coefficient of σ^*. Arous et al. (2021) proved that SGD learns the direction in $n \geq O(d^{1 \vee k^\star - 1})$ samples, which is also tight for this algorithm. This, however, does not meet the corresponding Correlational Statistical Query (CSQ) lower bound in this setting which, roughly states that $n \geq \Omega(d^{k^\star/2})$ samples are necessary. Recently, Damian et al. (2023) showed that smoothing the loss landscape can close this gap and attain the CSQ lower bound.

It is important to highlight that the aforementioned studies consider single- or multi-index settings in their full generality, without any structural assumptions on the model directions. In practice, however, high-dimensional data often exhibits low-dimensional structures; thus, sparsity is a natural property to consider. It is reasonable to expect that with this additional structure, the corresponding CSQ lower bound would become smaller. However, it remains unclear whether the previously considered training methods can still achieve this lower bound in the sparse setting.

In this paper, we introduce the concept of *soft sparsity* for the model directions $\boldsymbol{V}$ and derive a CSQ lower bound that depends on this sparsity level, which is always smaller than the lower bound in the general multi-index setting that only considers the worst-case sparsity scenario. Next, we demonstrate that pruned neural networks trained with a gradient-based method can achieve the optimal sample complexity suggested by this CSQ lower bound. Since the additional sparsity structure reduces the lower bound, basis-independent training methods such as gradient descent initialized with a symmetric distribution have provably suboptimal sample complexity; this implies a separation between pruning-based and existing training methods. We summarize our contributions below.

- We consider learning multi-index models of the form $y = \sigma^*(\boldsymbol{V}^\top \boldsymbol{x}) + \epsilon$ where the model directions $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ satisfy a certain soft sparsity. In Theorem 2, we prove a Correlational Statistical Query (CSQ) lower bound for this model, which also takes the inherent sparsity into account. The lower bound depends only on the sparsity level beyond a certain threshold. In this regime, our result shows that basis-independent training methods are always suboptimal.
- In the single-index case where $r = 1$, we prove that pruning the neural network with a sparsity level proportional to that of the model direction leads to a better sample complexity after training. Specifically, we consider polynomial link functions and show in Theorem 4 that the sample complexity achieved after pruning is optimal in the sense that, training after pruning can achieve the complexity suggested by the CSQ lower bound for any information exponent $k^\star \geq 1$.
- Finally, we consider the multi-index case with $r > 1$. Under an additional assumption implying that the information exponent is $k^\star = 2$, we prove in Theorem 5 that, pruned network trained with gradient descent can achieve the corresponding CSQ lower bound in this setting as well.

1.1. Related Work

Pruning and generalization. Pruning techniques have a rich history, spanning from classical methods that prune weights based on connectivity metrics like the Jacobian/Hessian (LeCun et al., 1989; Hassibi and Stork, 1992), to more recent approaches relying on weight magnitude (Han et al., 2015; Wen et al., 2016; Molchanov et al., 2017). Notably, iterative magnitude pruning, proposed by Han et al. (2015) demonstrated remarkable success in deep neural networks, sparking a surge in pruning research (Zhu and Gupta, 2018; Frankle et al., 2020; Gale et al., 2019; Liu et al., 2019).

Numerous studies demonstrate the beneficial effects of pruning on generalization (LeCun et al., 1989; Frankle and Carbin, 2019; Barsbey et al., 2021). Prior research treats pruning as an additional regularization technique, which requires weights to exhibit small norm (Giles and Omlin, 1994), achieve flat minima (Bartoldson et al., 2020), or enhance robustness to outliers (Jin et al., 2022). However, these studies are predominantly empirical and lack a theoretical foundation. Among the theoretical works, only Yang et al. (2023) examines random pruning within a specific statistical model. Our work extends their framework to encompass general polynomial link functions and data-dependent pruning algorithms, complementing generalization bounds with guarantees of optimality.

Lottery tickets and sparsity. Recent work has observed that overparameterized neural networks contain subsets, referred to as “winning tickets”, which can achieve comparable performance to the original network when trained independently (Frankle and Carbin, 2019). This phenomenon, known as the Lottery Ticket Hypothesis (LTH), has been extensively studied in the literature (Frankle et al., 2020; Gale et al., 2019; Chen et al., 2020; Zhou et al., 2019). Several recent works have focused on investigating the theoretical conditions for the existence of such subnetworks (Malach et al., 2020; Orseau et al., 2020) and the fundamental limitations of identifying them (Kumar et al., 2024). Our study takes a different approach by examining the training dynamics and generalization within the context of pruning. While previous works primarily focus on identifying subnetworks as predicted by the LTH, our research delves into the interplay between generalization and pruning methods.

Non-linear feature learning with neural networks. Recent theoretical studies have examined two scaling regimes in neural networks. In the “lazy” regime (Chizat et al., 2019), parameters remain largely unchanged from initialization, resembling kernel methods (Jacot et al., 2018; Du et al., 2019; Allen-Zhu et al.; Oymak and Soltanolkotabi, 2020). However, deep learning’s superiority over kernel models suggests they can go beyond this regime (Yehudai and Shamir, 2019; Ghorbani et al., 2020; Geiger et al., 2019). In contrast, the “mean-field” regime, where gradient descent converges to Wasserstein gradient flow, enables feature learning (Chizat et al., 2019; Mei et al., 2019; Chizat, 2022), but primarily applies to infinitely wide networks. Our paper explores a different setting, allowing for arbitrary-width neural networks without excessive overparameterization, while still employing mean-field scaling for weight initialization.

Feature learning with multiple-index teacher models. Learning an unknown low-dimensional function from data is fundamental in statistics (Li and Duan, 1989). Recent research in learning theory has considered this problem, aiming to demonstrate that neural networks can learn useful feature representations and outperform kernel methods (Ghorbani et al., 2020; Damian et al., 2022; Abbe et al., 2023). In particular, Abbe et al. (2022) investigates the necessary and sufficient conditions for learning with linear sample complexity in the mean-field limit, focusing on inputs confined to the hypercube. Closer to our setting are the recent works Damian et al. (2022); Mousavi-Hosseini et al. (2023a) which demonstrate a clear separation between NNs and kernel methods, leveraging the effect of representation learning. More recently, Dandi et al. (2024) shows that mini-batch SGD with finite number steps can learn a certain class of link functions with linear sample complexity.

Our work operates within a similar framework, incorporating an additional sparsity condition on relevant model directions. However, our analysis differs from previous work in two main aspects. First, our pruning results are constructive; we develop an explicit algorithm to establish the sample complexity of the pruned network trained via gradient descent. Second, pruning introduces a new dependency between weights and data, requiring an intricate analysis of gradient descent dynamics.

2. Preliminaries

Notations. Let $[n] \coloneqq \{1, \cdots, n\}$. We use $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_2$ to denote the Euclidean inner product and the norm, respectively. For matrices, $\|\cdot\|_2$ denotes the usual operator norm. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{A}_{i*}$ and $\boldsymbol{A}_{*j}$ denote the ith row and jth column of $\boldsymbol{A}$, respectively. S^{d-1} is the d-dimensional unit sphere. We use $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_d\}$ to denote the standard basis vectors in $\mathbb{R}^d$. We use $O(\cdot)$ and $\Omega(\cdot)$ to suppress constants in upper and lower bounds. We use $\tilde{O}(\cdot)$ to suppress poly-logarithmic terms in d in upper bounds. We use $o_d(\cdot)$ to denote vanishing terms as $d \to \infty$. We use $f \in \Theta(g)$ to denote $\Omega(g) \leq f \leq O(g)$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we use $\text{supp}(\boldsymbol{x}) \coloneqq \{i \in [d] \; : \; x_i \neq 0\}$. For a subset $\mathcal{J} \subseteq [d]$, we use $\boldsymbol{x}|_{\mathcal{J}} \in \mathbb{R}^d$ to denote the restriction of the vector $\boldsymbol{x}$ on $\mathcal{J}$, i.e., the coordinate indices that are not in $\mathcal{J}$ are set to be 0. For matrices, $\boldsymbol{A}|_{\mathcal{J}}$ denotes the matrix $\boldsymbol{A}$ with everything but the rows indexed by the elements in $\mathcal{J}$ set to 0. Finally, $\boldsymbol{x}|_{\text{top}(M)}$ denote the vector $\boldsymbol{x}$ with everything except M largest entries in magnitude set to 0.

Statistical model. For a link function $\sigma^* : \mathbb{R}^r \to \mathbb{R}$, we consider the multi-index model

$$y = \sigma^*(\boldsymbol{V}^\top \boldsymbol{x}) + \epsilon \quad \text{with} \quad \boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the input, ϵ is a zero-mean noise with $O(1)$ sub-Gaussian norm and $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ is an orthonormal matrix, i.e, $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_r$. We assume that σ^* is a polynomial of degree p, and it is normalized to satisfy $\mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_r)}[\sigma^*(\boldsymbol{z})] = 0$ and $\mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_r)}\left[\sigma^*(\boldsymbol{z})^2\right] = 1$. We consider the low-dimensional setting $r \ll d$ which, in the extreme case $r = 1$, covers single-index models. We are mainly interested in models where $\boldsymbol{V}$ exhibits *sparsity*; we use the following matrix norm:

$$\|\boldsymbol{V}\|_{2,q} \coloneqq \left\| \left(\|\boldsymbol{V}_{1*}\|_2, \cdots, \|\boldsymbol{V}_{d*}\|_2 \right) \right\|_q \quad \text{where} \quad q \in [0, 2),$$

where $\boldsymbol{V}_{i*}$ denotes the ith row of $\boldsymbol{V}$.[1] This is simply the usual ℓ_q norm of the vector with entries ℓ_2 norm of rows of $\boldsymbol{V}$. Since $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_r$, assuming that $\|\boldsymbol{V}\|_{2,q}$ is small constrains the model complexity significantly. Indeed, when $q = 0$, $\|\cdot\|_{2,q}$ counts the number of non-zero rows, serving as a measure of sparsity in high-dimensional settings. In the case $q \in (0, 2)$, small $\|\cdot\|_{2,q}$ norm allows all rows to potentially contain non-zero values, provided their ℓ_2 norms are all relatively small. When we have $\|\boldsymbol{V}\|_{2,q}^q \leq R_q$ for some R_q, we adopt a terminology from Raskutti et al. (2011) and refer to R_q as the *soft sparsity* level. Notably, the particular choice $\|\cdot\|_{2,q}$ is motivated by its coordinate-independent property, as $\|\boldsymbol{V}\boldsymbol{U}\|_{2,q} = \|\boldsymbol{V}\|_{2,q}$ for any orthonormal matrix $\boldsymbol{U} \in \mathbb{R}^{r \times r}$.

Two-layer Neural Networks. Denoting the ReLU activation with $\phi(t) = \max\{t, 0\}$, we consider learning with two-layer neural networks of the form

$$\hat{y}(\boldsymbol{x}; (\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b})) = \sum_{j=1}^{2m} a_j \phi(\langle \boldsymbol{W}_{j*}, \boldsymbol{x} \rangle + b_j) = \langle \boldsymbol{a}, \phi(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \rangle,$$

1. To be precise, $\|\cdot\|_{2,q}$ is not a norm when $q < 1$.

where $\boldsymbol{W} = \{\boldsymbol{W}_{j*}\}_{j=1}^{2m}$ is the $2m \times d$ matrix whose rows are denoted with $\boldsymbol{W}_{j*}$, $\boldsymbol{a} = \{a_j\}_{j=1}^{2m}$ is the second layer weights, $\boldsymbol{b} = \{b_j\}_{j=1}^{2m}$ is the biases. Note that $\phi(\cdot)$ is applied element-wise in the second equality. We define the population and the empirical risks respectively as

$$R((\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b}))= \frac{1}{2}\mathbb{E}\left[(\hat{y}(\boldsymbol{x}; (\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b})) - y)^2\right], \quad R_n((\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b}))= \frac{1}{2n}\sum_{i=1}^{n}(\hat{y}(\boldsymbol{x}_i; (\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b})) - y_i)^2$$

where the expectation above is over the data distribution.

Our training procedure consists of three-steps: (i) we first prune the network for dimension reduction, then (ii) we take a gradient descent iteration with a large step-size to train $\boldsymbol{W}$, and finally (iii) we train the second layer weights $\boldsymbol{a}$. We will provide the details of the algorithm, in particular the pruning step in Section 4. Similar to the previous works, e.g. Chizat et al. (2019); Damian et al. (2022); Dandi et al. (2023), we use symmetric initialization so that $\hat{y}(\boldsymbol{x}, (\boldsymbol{a}^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{b}^{(0)})) = 0$; we assume that the network has a width of $2m$ such that

$$a_j^{(0)} = -a_{2m-j}^{(0)}, \quad \boldsymbol{W}_{j*}^{(0)} = \boldsymbol{W}_{(2m-j)*}^{(0)} \in S^{d-1}, \quad b_j^{(0)} = b_{2m-j}^{(0)}, \quad \text{for } j \in [m]. \tag{2.1}$$

Particularly, we will use the following initialization for the second-layer weights and the biases,

$$a_j^{(0)} \sim \text{Unif}\{-1, 1\}, \text{ and } b_j^{(0)} \sim \mathcal{N}(0, 1), \ \ j \in [m]. \tag{2.2}$$

Initialization of $\boldsymbol{W}^{(0)}$ will depend on the pruning algorithm and be detailed later. Note that due to (2.1), the gradient of R_n with respect to $\boldsymbol{W}_{j*}$ at initialization can be written as follows:

$$\nabla_{\boldsymbol{W}_{j*}} R_n((\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b})) = \frac{-a_j}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i \phi'\left(\langle \boldsymbol{W}_{j*}, \boldsymbol{x}_i\rangle + b_j\right).$$

We simplify the notation to $\nabla_j R_n((\boldsymbol{a}, \boldsymbol{w}, \boldsymbol{b}))$ whenever $\boldsymbol{W}_{i*} = \boldsymbol{w}$ for all i.

Characteristics of the link function σ^* plays an important role in the complexity of learning. Indeed, recent works showed that the term in the Hermite expansion of σ^* with the smallest degree determines the sample complexity (Arous et al., 2021; Abbe et al., 2023). In line of these works, we also rely on Hermite expansions, for which we define the Hermite polynomials as follows.

Definition 1 (Hermite Polynomials) *The kth Hermite polynomial $H_{e_k} : \mathbb{R} \to \mathbb{R}$ is the degree k polynomial defined by*

$$H_{e_k}(t) = (-1)^k e^{t^2/2} \frac{d^k}{dt^k} e^{-t^2/2}.$$

3. Limitations of Basis Independent Methods: CSQ Lower Bounds

In this section, we explore the fundamental barriers under the *soft sparsity* structure we assume on the statistical model. Specifically, we establish a lower bound for Correlational Statistical Query (CSQ) methods within our framework. We note that the CSQ methods encompasses a wide class of algorithms under the squared error loss. We consider the function class

$$\mathcal{F}_{r,k} := \left\{ \boldsymbol{x} \to \frac{1}{\sqrt{rk!}} \sum_{j=1}^{r} H_{e_k}(\langle \boldsymbol{V}_{*j}, \boldsymbol{x}\rangle) \;\middle|\; \boldsymbol{V} \in \mathbb{R}^{d\times r},\ \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_r,\ \|\boldsymbol{V}\|_{2,q}^q \le r^{\frac{q}{2}} d^{\alpha\left(1-\frac{q}{2}\right)} \right\} \tag{3.1}$$

where $\alpha \in (0,1)$, H_{e_k} denotes the kth Hermite polynomial (see Definition 1), and for $q = 0$, we use the convention $\|\boldsymbol{V}\|_{2,0}^0 := \|\boldsymbol{V}\|_{2,0}$. We remark that the constraint $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_r$ directly implies $r \leq \|\boldsymbol{V}\|_{2,q}^q \leq r^{q/2} d^{1-q/2}$. Therefore, $\mathcal{F}_{r,k}$ covers all possible sparsity levels by varying the parameter α. We have the following result on the query complexity of CSQ methods.

Theorem 2 *Consider $\mathcal{F}_{r,k}$ with some $q \in [0,2)$ and $\alpha \in (0,1)$. For a sufficiently large d depending on (r,k,q,α), any CSQ algorithm for $\mathcal{F}_{r,k}$ that guarantees error $\varepsilon = \Omega(1)$ requires either queries of accuracy $\tau = \widetilde{O}\big(d^{-(\alpha \wedge \frac{1}{2})\frac{k}{2}}\big)$ or super-polynomially many queries in d.*

Using the heuristic $\tau \approx \frac{1}{\sqrt{n}}$ as in Damian et al. (2022), Theorem 2 implies that $n \geq \Omega\big(d^{(\alpha \wedge \frac{1}{2})k}\big)$ samples are necessary to learn a function in $\mathcal{F}_{r,k}$ unless the algorithm makes super-polynomial queries in d. This recovers the existing lower bound $\Omega\big(d^{k/2}\big)$ given in Damian et al. (2022); Abbe et al. (2023), when the constraint is sufficiently large, i.e., $\alpha > \frac{1}{2}$. Conversely, when the soft sparsity level is sufficiently small, i.e., $\alpha \leq \frac{1}{2}$, we observe that the complexity lower bound reads $\Omega\big(d^{\alpha k}\big)$. Remarkably, in Section 5, we prove that a pruned neural network trained with gradient descent can indeed attain this lower bound; thus, it achieves optimal sample complexity in this sense.

We note that $\|\boldsymbol{V}\|_{2,q}^q$ can be as small as r; thus, the CSQ lower bound in this regime can be significantly smaller than the unconstrained version $\Omega\big(d^{k/2}\big)$. On the other hand, methods that are independent of the underlying basis, such as gradient descent with symmetric initialization, cannot exploit the additional structure. As a result, these methods are constrained by the sample complexity lower bound of $\Omega\big(d^{k/2}\big)$ in the worst case. Finally, it is worth emphasizing that CSQ lower bounds do not directly apply to algorithms like SGD or one-step gradient descent due to non-adversarial noise. Nevertheless, under the square loss, queries of these algorithms fall under the correlational regime, thus the fundamental barrier CSQ lower bounds provide is frequently referred to when assessing the optimality of these methods; see e.g. Damian et al. (2022, 2023); Abbe et al. (2023).

4. Training Procedure: Pruning as Dimension Reduction

In this section, we outline the pruning procedure and how it effectively reduces the dimensionality of the learning problem, leading to the optimal sample complexity suggested by Theorem 2.

Intuition. To gain intuition, we start with the population dynamics and consider a simplified single-index setting to demonstrate the resulting dimension reduction. Let

$$\sigma^*(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) = H_{e_2}(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) \;\text{ with }\; \boldsymbol{v} = \Big(d^{-\frac{1}{4}}, \cdots, d^{-\frac{1}{4}}, 0, 0, \cdots, 0\Big),$$

where the direction $\boldsymbol{v}$ is sparse, i.e. $\|\boldsymbol{v}\|_0 = \sqrt{d} \ll d$. Moreover, for clarity, let us fix the output layer weights to $a_j^{(0)} = 1$ and biases to $b_j^{(0)} = 0$ and consider the population gradient at initialization. To see why comparing gradients performs dimension reduction, we write

$$\nabla_j R((\boldsymbol{a}^{(0)}, \boldsymbol{e}_i, \boldsymbol{b}^{(0)})) = -\mathbb{E}\left[\sigma^*(\langle \boldsymbol{v}, \boldsymbol{x} \rangle)\phi'(\langle \boldsymbol{e}_i, \boldsymbol{x} \rangle)\boldsymbol{x}\right] = -\sqrt{\tfrac{2}{\pi}}\, \langle \boldsymbol{v}, \boldsymbol{e}_i \rangle\, \boldsymbol{v} + \tfrac{1}{\sqrt{2\pi}}\, \langle \boldsymbol{v}, \boldsymbol{e}_i \rangle^2\, \boldsymbol{e}_i \quad (4.1)$$

where $\boldsymbol{e}_i$ is the ith standard basis and constants are due to the Hermite coefficients of the ReLU activation $\phi(\cdot)$. Thus, we have

$$\|\nabla_j R((\boldsymbol{a}^{(0)}, \boldsymbol{e}_i, \boldsymbol{b}^{(0)}))\|_2^2 = \tfrac{2}{\pi}\boldsymbol{v}_i^2 + O(d^{-1}).$$

Algorithm 1 PruneNetwork

Inputs: (i) Data: $\mathcal{D} \coloneqq \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ (ii) Network width:[1] $m \in \mathbb{N}$ (iii) Sparsity level: $M \in [d]$ (iv) Shrinkage constant: $c \in (0, 1)$

1: Let $\tilde{\boldsymbol{e}}_i$ be as in (4.2), and initialize $\boldsymbol{a}^{(0)}$ and $\boldsymbol{b}^{(0)}$ as in (2.1)-(2.2)
2: Let $\widetilde{\nabla}_j R_n^{\pm}(\tilde{\boldsymbol{e}}_i) \coloneqq \nabla_j R_n^{\pm}\big((\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\big)|_{\text{top}(M)}$ and $\|\widetilde{\nabla} R_n^{\pm}(\tilde{\boldsymbol{e}}_i)\|_F^2 = \sum_{j=1}^{2m} \|\widetilde{\nabla}_j R_n^{\pm}(\tilde{\boldsymbol{e}}_i)\|_2^2$
3: $\mathcal{J} = \text{supp}(\widetilde{\nabla}_j R_n^{-}(\tilde{\boldsymbol{e}}_i))$ for some $j \in [m]$ with $b_j^{(0)} \geq 0$ if one exists, otherwise $\mathcal{J} = \emptyset$.
4: Sort $\|\widetilde{\nabla} R_n^{+}(\tilde{\boldsymbol{e}}_{j_1})\|_2 \geq \|\widetilde{\nabla} R_n^{+}(\tilde{\boldsymbol{e}}_{j_2})\|_2 \geq \cdots \geq \|\widetilde{\nabla} R_n^{+}(\tilde{\boldsymbol{e}}_{j_d})\|_2$ and $\mathcal{J} \leftarrow \mathcal{J} \cup \{j_1, \cdots, j_M\}$
5: Sort $\|\widetilde{\nabla} R_n^{-}(\tilde{\boldsymbol{e}}_{k_1})\|_2 \geq \|\widetilde{\nabla} R_n^{-}(\tilde{\boldsymbol{e}}_{k_2})\|_2 \geq \cdots \geq \|\widetilde{\nabla} R_n^{-}(\tilde{\boldsymbol{e}}_{k_d})\|_2$ and $\mathcal{J} \leftarrow \mathcal{J} \cup \{k_1, \cdots, k_M\}$
6: **Return:** $\mathcal{J}$

Since the entries of $\boldsymbol{V}$ scale with $d^{-1/4}$ in high dimensions, comparing the norm of gradients is equivalent to comparing the magnitude of each entry $\boldsymbol{v}_i$. Hence, non-zero coordinates of $\boldsymbol{V}$ can be picked up by pruning, which is effectively reducing the dimension of the problem from d to the sparsity level $\sqrt{d}$ in this example.

Algorithm 1 essentially extends the basic intuition above to general link functions σ^* and empirical gradients. However, such an extension requires us to handle two technical difficulties due to the bias in the Hermite expansion of the population gradient. In Section 6, we illustrate how each step in Algorithm 1 is designed to avoid those difficulties using the following arguments:

- (Data augmentation) We augment the feature vectors with an independent non-informative random variable, i.e., $\boldsymbol{x}' \leftarrow (\boldsymbol{x}, z)^T$ where $z \sim \mathcal{N}(0, 1)$ and independent of $\boldsymbol{x}$. For notational convenience, we assume that the augmented features $\boldsymbol{x}'$ (henceforth referred to as $\boldsymbol{x}$) is d-dimensional. Since the last entry of the feature vector is non-informative, we can assume $\boldsymbol{V}_{d*} = 0$, without loss of generality.

- (Shifted standard basis) We compare the magnitudes of the gradients initialized at

$$\tilde{\boldsymbol{e}}_j \coloneqq \begin{cases} c\boldsymbol{e}_j + \sqrt{1-c^2}\boldsymbol{e}_d, & j \in [d-1] \\ \boldsymbol{e}_d & j = d. \end{cases} \tag{4.2}$$

  Here, standard basis vectors are *shifted* by a factor of $c \in (0, 1)$ to make sure that the extra terms vanish (see Line 1 in Algorithm 1).

- (Even-odd decomposition) We consider the even and odd components of the activation separately, i.e., $\phi_{\pm}(t; b) = (\phi(t + b) \pm \phi(-t + b))/2$, and evaluate the gradient with these components (Line 2 in Algorithm 1)

$$\nabla_j R_n^{\pm}((\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})) \coloneqq \frac{1}{2}\left[\nabla_j R_n((\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) \pm \nabla_j R_n((\boldsymbol{a}^{(0)}, -\tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right].$$

Pruning Algorithm 1. The pruning algorithm is based on comparing gradient magnitudes at initialization to perform dimension reduction. The challenge lies in utilizing empirical gradients. To estimate the gradient magnitudes, we consider pruned empirical gradients , i.e., $\widetilde{\nabla}_j R_n^{\pm}(\tilde{\boldsymbol{e}}_i) \coloneqq \nabla_j R_n^{\pm}\big((\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\big)|_{\text{top}(M)}$ (Line 2). Improving on the sample mean estimator, which requires

Algorithm 2 Gradient-based Training

Inputs: (i) Data: $\mathcal{D} \coloneqq \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ (ii) Learning rate: $\eta_t > 0$ (iii) Weight Decay: $\lambda_t > 0$ (iv) Network width:[1] $m \in \mathbb{N}$ (v) Pruning Level: $M \in [d]$ (vi) Shrinkage constant: $c \in (0, 1)$

1: $\mathcal{J} \leftarrow \texttt{PruneNetwork}(\mathcal{D}, m, M, c)$

2: Re-initialize $\boldsymbol{a}^{(0)}$ and $\boldsymbol{b}^{(0)}$ as in as in (2.1)-(2.2), and

$$\boldsymbol{W}_{j*}^{(0)} \sim S_{\mathcal{J}}^{d-1}, \text{ and } \boldsymbol{W}_{j*}^{(0)} = \boldsymbol{W}_{(2m-j+1)*}^{(0)}, \; j \in [m].$$

3: Train the first layer weights: For $j \in [2m]$

$$\boldsymbol{W}_{j*}^{(1)} = \boldsymbol{W}_{j*}^{(0)} - \eta_1 \Big(\nabla_{\boldsymbol{W}_{j*}} R_n \left((\boldsymbol{a}^{(0)}, \boldsymbol{W}_{j*}^{(0)}, \boldsymbol{b}^{(0)}) \right) \big|_{\mathcal{J}} + \lambda_1 \boldsymbol{W}_{j*}^{(0)} \Big).$$

4: Re-initialize biases: For $j \in [m]$, let $b_j^{(1)} \sim \mathcal{N}(0, 1)$ and $b_j^{(1)} = b_{2m-j+1}^{(1)}$.

5: Train the second layer weights:

$$\boldsymbol{a}^{(t+1)} = \boldsymbol{a}^{(t)} - \eta_t \left(\nabla_{\boldsymbol{a}} R_n((\boldsymbol{a}^{(t)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) + \lambda_t \boldsymbol{a}^{(t)} \right), \; t \geq 2.$$

6: **Return:** $\hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) = \left\langle \boldsymbol{a}^{(T)}, \phi(\boldsymbol{W}^{(1)} \boldsymbol{x} + \boldsymbol{b}^{(1)}) \right\rangle$

$O(d)$ samples, pruned sample mean requires sample complexity of $\tilde{O}(d^\alpha)$ by leveraging the sparsity of population gradient, hence providing the desired sample complexity for the algorithm.

Having computed the empirical gradients, we proceed by evaluating and sorting the gradients (Lines 4 and 5). We keep the connections with larger gradient magnitude while pruning the remaining small entries.

Training Algorithm 2. After pruning the neural network, we perform a gradient-based training procedure. Let $S_{\mathcal{J}}^{d-1} \sim \text{Unif}\left\{ \boldsymbol{x} \in S^{d-1} \;\middle|\; \boldsymbol{x}_j = 0 \text{ for } j \in [d] \setminus \mathcal{J} \right\}$ denote the uniform distribution on the set of unit vectors supported on $\mathcal{J}$. The algorithm symmetrically re-initializes the neural network weights randomly restricted to $\mathcal{J}$, i.e.,

$$\boldsymbol{W}_{j*}^{(0)} \sim S_{\mathcal{J}}^{d-1} \text{ and } \boldsymbol{W}_{j*}^{(0)} = \boldsymbol{W}_{(2m-j+1)*}^{(0)}.$$

We consider a slightly modified version of the one-step gradient descent update used in recent works Damian et al. (2022); Ba et al. (2022, 2023), namely, we perform a gradient step restricted on set $\mathcal{J}$ (Line 3). Here, since both $\boldsymbol{W}^{(0)}$ and $\nabla_{\boldsymbol{W}} R_n\big((\boldsymbol{a}^{(0)}, \boldsymbol{W}^{(0)}, \boldsymbol{b}^{(0)})\big)|_{\mathcal{J}}$ are supported on $\mathcal{J}$, $\boldsymbol{W}^{(1)}$ is also supported on $\mathcal{J}$. Finally, after training the first layer weights $\boldsymbol{W}^{(0)}$, we again symmetrically re-initialize the biases and train the second-layer weights using gradient descent (Lines 4 and 5).

We note that Algorithm 2 as stated can be used to learn both single-index and multi-index models, and falls under the correlational query algorithms discussed in Section 3. However, in the multi-index setting, the algorithm needs a slight modification, which we detail in Section 5.2.

1. Note that the actual width of the network is $2m$ due to symmetric initialization.

5. Main Results

In this section, we present learning guarantees on Algorithm 2 when the data is generated from either a single-index or a multi-index model. We focus on single-index models first.

5.1. Learning Sparse Single-index Models with Pruning

In what follows, we define a complexity measure for the link function to be learned.

Definition 3 (Information exponent) *For the link function σ^*, we let $\sigma^* := \sum_{k=0}^{p} \frac{\gamma_k}{k!} H_{e_k}$ be its Hermite expansion. The information exponent of σ^*, which we denote by $k^\star$, is the index of the first non-zero Hermite coefficient of σ^*, i.e., $k^\star := \inf\{k \geq 1 \mid \gamma_k \neq 0\}$.*

Intuitively, information exponent measures the magnitude of information contained in the gradient at initialization, and larger $k^\star$ implies increased gradient descent complexity (Arous et al., 2021). The main result in the single-index setting relies on the above definition, and is given below.

Theorem 4 *Let $\|\boldsymbol{V}\|_{2,q}^q = \Theta\big(d^{(1-\frac{q}{2})\alpha}\big)$, for some $q \in [0, 2)$ and $\alpha \in (0, 1)$. For any $\varepsilon > 0$, consider Algorithm 2 with $m = \Theta\left(d^\varepsilon\right)$, $c = \frac{1}{\log d}$,*

$$\eta_1 = \tilde{O}\left(M^{\frac{k^\star - 1}{2}}\right), \;\; \lambda_1 = \frac{1}{\eta_1}, \;\; \eta_t = \frac{1}{\tilde{O}(m) + \lambda_t}, \;\; \lambda_t = \tilde{O}(m), \; t \geq 2, \;\; \text{and} \;\; T = \tilde{O}(1).$$

For every $\ell \in \mathbb{N}$, there exists a constant $d_{\ell,\varepsilon}$, depending on ℓ and ε, such that for $d \geq d_{\ell,\varepsilon}$, if

$$n = \tilde{O}\left(d^{\alpha k^\star}\right) \;\; \text{and} \;\; M = \tilde{O}\left(d^{\alpha}\right),$$

then, Algorithm 2 guarantees that with probability at least $1 - d^{-\ell}$

$$\mathbb{E}\left[\left(\hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) - y\right)^2\right] - \mathbb{E}[\epsilon^2] \leq \tilde{O}\left(\frac{1}{m} + \sqrt{\frac{M}{n}}\right) + o_d(1).$$

We observe that for any constraint level, the sample complexity in Theorem 4 reduces to $\tilde{O}(d^{\alpha k^\star})$ for $\alpha \in (0, 1)$, which improves upon the existing $O\big(d^{k^\star}\big)$ guarantees for gradient-based algorithms (Bietti et al., 2022; Mousavi-Hosseini et al., 2023b). Moreover, in the case $\alpha \leq 1/2$, the upper bound matches with the CSQ lower bound in Theorem 2. Finally, we observe that for the generalization error to be small, the width m and particularly the ambient dimension d need to be both sufficiently large; thus, the right hand side of the bound vanishes only in high-dimensions.

5.2. Learning Sparse Multi-index Models with Pruning

In this section, we consider multi-index models, i.e., the case $r > 1$. We consider Algorithm 2 with two minor modifications, following a similar construction to Damian et al. (2022) adapted to our pruning framework. Right after the pruning step, between Lines 1 and 2, we subtract an estimate of the first Hermite component from the response variable. We add this term back at the output, in Line 6. These modifications are given as follows.

$$\texttt{1.5:}\;\; y_i \leftarrow y_i - \langle \hat{\boldsymbol{\mu}}|_{\mathcal{J}}, \boldsymbol{x}_i \rangle\,,\; i \in [n] \text{ where } \hat{\boldsymbol{\mu}} := \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i,$$

$$\texttt{6:}\;\; \text{Return: } \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) = \langle \hat{\boldsymbol{\mu}}|_{\mathcal{J}}, \boldsymbol{x} \rangle + \left\langle \boldsymbol{a}^{(T)}, \phi(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) \right\rangle.$$

We will refer to the modified algorithm as Algorithm 2^{+}.

The following condition on the link function, referred to as *non-degeneracy* in Damian et al. (2022), is helpful in the analysis.

Assumption 1 *The link function $\sigma^* : \mathbb{R}^r \to \mathbb{R}$ satisfies that $\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}\boldsymbol{z}^\top] \in \mathbb{R}^{r\times r}$ is full rank.*

Under this assumption, σ^* has information exponent[2] $k^\star = 2$. Therefore, this condition is significantly more restrictive than the assumptions in the single-index case. This is, however, expected since recovering the entire principal subspace spanned by the model directions, i.e., the column space of $\boldsymbol{V}$, is significantly more challenging than recovering a single direction. Under this condition, we state the main result of the multi-index setting.

Theorem 5 *Suppose that Assumption 1 holds. Let $\|\boldsymbol{V}\|_{2,q}^{q} = \Theta\big(d^{(1-\frac{q}{2})\alpha}\big)$, for some $q \in [0, 2)$ and $\alpha \in (0, 1)$. For any $\varepsilon > 0$, consider Algorithm 2^{+} with $m = \Theta(d^{\varepsilon})$, $c = \frac{1}{\log d}$,*

$$\eta_1 = \tilde{O}(M)\,, \quad \lambda_1 = \frac{1}{\eta_1}, \quad \eta_t = \frac{1}{\tilde{O}(m) + \lambda_t}, \quad \lambda_t = \tilde{O}(m),\; t \geq 2, \;\; \text{and} \;\; T = \tilde{O}(1).$$

For every $\ell \in \mathbb{N}$, there exists a constant $d_{\ell,\varepsilon}$, depending on ℓ and ε, such that for $d \geq d_{\ell,\varepsilon}$, if

$$n = \tilde{O}\Big(d^{2\alpha}\Big) \;\; \text{and} \;\; M = \tilde{O}\Big(d^{\alpha}\Big),$$

then, Algorithm 2^{+} guarantees that with probability at least $1 - d^{-\ell}$

$$\mathbb{E}\left[\Big(\hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) - y\Big)^2\right] - \mathbb{E}[\epsilon^2] \leq \tilde{O}\left(\frac{1}{m} + \sqrt{\frac{M}{n}}\right) + o_d(1).$$

The above result states that the improvement in sample-complexity due to pruning extends to the multi-index setting as well. As in the single-index case, for all sparsity levels, gradient descent followed by pruning requires $\tilde{O}(d^{2\alpha})$, for the soft sparsity level $\Theta(d^{(1-q/2)\alpha})$ and $\alpha \in (0, 1)$, which improves over the existing $\tilde{O}(d^2)$ bound shown in Damian et al. (2022). It is worth noting that the bound in Damian et al. (2022) does not meet the CSQ lower bound in their setting. This gap, however, was later closed in Damian et al. (2023) via smoothing the loss. With the additional soft sparsity condition in Theorem 5, even smoothing will achieve suboptimal sample complexity

2. In Definition 3 , the information exponent is defined for $r = 1$. Similar to an argument by Abbe et al. (2023), we can generalize our definition to encompass multi-index settings by considering the degree of the lowest order Hermite components in σ^*. With this, Assumption 1 leads to an information exponent $k^\star = 2$ in the worst-case scenario, encompassing situations where the first Hermite component does not exist.

guarantee since the corresponding CSQ lower bound in this regime becomes smaller. Nevertheless, observing that the function class in (3.1) satisfies Assumption 1 for $r > 1$ and $k = 2$, our lower bound in Theorem 2 implies that the above result is tight in this sense, for $\alpha \leq 1/2$.

For the generalization error to be small in Theorem 5, we require the width m to be large. More crucially, this bound is small only in high-dimensions where the ambient dimension is large. Therefore, pruned neural networks learn useful representations via gradient descent, and achieves optimal sample complexity in the above sense in high-dimensions, also in the multi-index setting.

6. Technicalities Around Pruning

First Technical Difficulty. A technical difficulty arises due to the bias introduced by the first-order Hermite components. To illustrate a pathological case for this problem, we consider two models, one with and one without the first-order Hermite component:

$$y = \underbrace{\tfrac{1}{\sqrt{2}} H_{e_2}(\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle) + \tfrac{1}{\sqrt{2}} H_{e_2}(\langle \boldsymbol{v}_2, \boldsymbol{x} \rangle)}_{\text{no first-order Hermite component}} \qquad \text{and} \qquad \check{y} = y + \underbrace{\langle \boldsymbol{v}, \boldsymbol{x} \rangle}_{\substack{\text{first-order} \\ \text{Hermite component}}} \tag{6.1}$$

where we choose $\boldsymbol{v}_1 = \boldsymbol{e}_1$, $\boldsymbol{v}_2 = \boldsymbol{e}_2$, $\boldsymbol{v} = \frac{-1}{\sqrt{\pi}}(\boldsymbol{e}_1 + \boldsymbol{e}_2)$. Here, the second model, $\check{y}$, includes an additional first-order Hermite term to illustrate its effect.

For the first model, we can derive the population gradient in (4.1) as follows:

$$\nabla_j R((\boldsymbol{a}^{(0)}, \boldsymbol{e}_i, \boldsymbol{b}^{(0)})) = -\mathbb{E}\left[y \phi'(\langle \boldsymbol{e}_i, \boldsymbol{x} \rangle) \boldsymbol{x} \right] = \frac{-1}{2\sqrt{\pi}} \begin{cases} \boldsymbol{e}_1 & i = 1 \\ \boldsymbol{e}_2 & i = 2 \\ 0 & i > 2, \end{cases} \tag{6.2}$$

For the second model, denoted by $\nabla_j \check{R}$, the population gradient is given by:

$$\begin{aligned} \nabla_j \check{R}((\boldsymbol{a}^{(0)}, \boldsymbol{e}_i, \boldsymbol{b}^{(0)})) &= -\mathbb{E}\left[\check{y} \phi'(\langle \tilde{\boldsymbol{e}}_i, \boldsymbol{x} \rangle) \boldsymbol{x} \right] \\ &= -\underbrace{\mathbb{E}\left[\langle \boldsymbol{v}, \boldsymbol{x} \rangle \, \phi'(\langle \boldsymbol{e}_i, \boldsymbol{x} \rangle) \boldsymbol{x} \right]}_{\substack{\text{due to the additional} \\ \text{first-order Hermite term}}} - \underbrace{\mathbb{E}\left[y \phi'(\langle \boldsymbol{e}_i, \boldsymbol{x} \rangle) \boldsymbol{x} \right]}_{=(6.2)} = \frac{1}{2\sqrt{\pi}} \begin{cases} \boldsymbol{e}_2 & i = 1 \\ \boldsymbol{e}_1 & i = 2 \\ \boldsymbol{e}_1 + \boldsymbol{e}_2 & i > 2. \end{cases} \end{aligned} \tag{6.3}$$

We notice that in the first model, comparing the gradient magnitudes would recover the support, whereas in the second model the gradients evaluated at the support of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ ($i = 1, 2$) have smaller norms than other cases (see Appendix A for the details).

The issue described above arises from the presence of the first-order Hermite term in (6.3). To address this, we consider the even and odd components of the activation separately, as detailed in Section 4. This decomposition allows us to separate the first-order Hermite term from the higher-order terms in the Hermite expansion through even-odd decomposition, and eliminate the problematic bias of the first-order term illustrated in (6.2)-(6.3).

Second Technical Difficulty. The second technical difficulty arises due to the presence of magnitude mismatch within the entries of $\boldsymbol{V}$. To illustrate, let us consider the following case: For a small

$0 < \varepsilon \ll d^{-1/2}$ and constants γ_2 and γ_4 specified later, let

$$\sigma^*(\langle \boldsymbol{v}, \boldsymbol{x}\rangle) = \tfrac{\gamma_2}{\sqrt{2}} H_{e_2}(\langle \boldsymbol{v}, \boldsymbol{x}\rangle) + \tfrac{\gamma_4}{\sqrt{4!}} H_{e_4}(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)$$
$$\text{with } \boldsymbol{v} = \Big(\sqrt{1-(\sqrt{d}-1)\varepsilon^2}, \underbrace{\varepsilon, \cdots, \varepsilon}_{\sqrt{d}-1 \text{ many}}, 0, 0, \cdots, 0\Big) \quad (6.4)$$

where $\boldsymbol{v}$ is sparse, i.e. $\|\boldsymbol{v}\|_0 = \sqrt{d} \ll d$, and the first entry of $\boldsymbol{v}$ is significantly larger than the rest. The population gradient in this case is given by

$$\begin{aligned}\nabla_j R((\boldsymbol{a}^{(0)}, \boldsymbol{e}_i, \boldsymbol{b}^{(0)})) &= -\mathbb{E}\left[\sigma^*(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)\phi'(\langle \boldsymbol{e}_i, \boldsymbol{x}\rangle)\boldsymbol{x}\right] \\ &= \underbrace{-\boldsymbol{v}\left(\sqrt{2}\gamma_2\tilde{\gamma}_2\boldsymbol{v}_i + \frac{2\gamma_4\tilde{\gamma}_4}{\sqrt{6}}\boldsymbol{v}_i^3\right)}_{\text{informative term}} \underbrace{-\boldsymbol{e}_i\left(\frac{\tilde{\gamma}_4\gamma_2}{\sqrt{2}}\boldsymbol{v}_i^2 + \frac{\tilde{\gamma}_6\gamma_4}{\sqrt{4!}}\boldsymbol{v}_i^4\right)}_{\text{extra term}}, \quad (6.5)\end{aligned}$$

where $\tilde{\gamma}_i$ denotes the i^{th} Hermite coefficients of the ReLU activation $\phi(\cdot)$. The informative term contains the information about the direction $\boldsymbol{v}$ while the extra term appears due to the properties of Hermite polynomials. Here, a very large $\boldsymbol{v}_i$ might cause extra terms to be comparable to the informative terms, leading to cancellation. As detailed in Appendix A, we can find $(\gamma_2, \gamma_4, \varepsilon)$ such that for $i = 1$ (corresponding to largest entry in $\boldsymbol{V}$), the informative and extra terms cancel each other in (6.6), i.e., informative term $\approx$ $-$extra term, making the algorithm require exponentially many samples to find the largest entry.

On the other hand, we observe that if $\boldsymbol{v}_i$'s vanish with d in (6.6), the informative term would dominate since it scales with $O(\boldsymbol{v}_i)$ whereas the extra term scales with $O(\boldsymbol{v}_i^2)$. To make sure that is the case in the presence of very large entries in $\boldsymbol{V}$, we use data augmentation and compare the magnitude of gradients evaluated at a shifted standard basis, as detailed in Section 4. Note that in this case,

$$\begin{aligned}\nabla_j R((\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})) &= -\mathbb{E}\left[\sigma^*(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)\phi'(\langle \tilde{\boldsymbol{e}}_i, \boldsymbol{x}\rangle)\boldsymbol{x}\right] \\ &= \underbrace{-c\boldsymbol{v}\left(\sqrt{2}\gamma_2\tilde{\gamma}_2\boldsymbol{v}_i + c^2\frac{2\gamma_4\tilde{\gamma}_4}{\sqrt{6}}\boldsymbol{v}_i^3\right)}_{\text{informative term}} \underbrace{-c^2\boldsymbol{e}_i\left(\frac{\tilde{\gamma}_4\gamma_2}{\sqrt{2!}}\boldsymbol{v}_i^2 + c^2\frac{\tilde{\gamma}_6\gamma_4}{\sqrt{4!}}\boldsymbol{v}_i^4\right)}_{\text{extra term}}, \quad (6.6)\end{aligned}$$

where a sufficiently small $c > 0$ ensures that the informative term dominates the right-hand side.

7. Discussion

We studied how pruning impacts the sample complexity of learning single and multi-index models. Our results show that pruning the network to a sparsity level proportional to the soft sparsity of relevant model directions significantly improves sample complexity. Moreover, we supported our results with a sparsity-aware CSQ lower bound which revealed that if the sparsity level exceeds a certain threshold, the sample complexity of training a pruned network cannot be improved in general. Conversely, the gap between our lower bound and the CSQ lower bound for the general dense case suggests that basis-independent methods, such as gradient descent initialized with a rotationally independent distribution, cannot achieve the sample complexity of the pruned network.

We outline a few limitations of our current work and discuss directions for future research.

– In our work, we considered training network weights with a single gradient step. However, recent research suggests that using multiple gradient descent steps in the multi-index setting yields improved sample complexity compared to single-step algorithms Abbe et al. (2023); Dandi et al. (2024). Therefore, considering pruning with a multi-step gradient descent algorithm can provide a more complete picture. Particularly, investigating pruning in the context of incremental (or curriculum) learning presents an interesting direction for future research.

– In the gradient-based algorithm, we considered a somewhat unconventional initialization, leveraging the symmetry it introduces. It would be interesting to examine cases where we train a network with multiple neurons starting from a more standard initialization. This analysis is challenging due to the interactions between the neurons.

– The results presented in this paper are based on the assumption that the input distribution follows an isotropic Gaussian distribution. Recent works Mousavi-Hosseini et al. (2023b); Ba et al. (2023) showed that there is an intricate interplay between the model and the important covariance directions, and the overall performance of neural networks is governed by their interplay. Studying the effect of pruning in this regime and also extending our results to other distributions (Roy et al., 2021), for example via zero-biased transformations (Goldstein and Reinert, 1997; Goldstein and Wei, 2019), is a topic for future research.

Acknowledgements

We would like to thank Berivan Isik and Alireza Mousavi-Hosseini for their helpful discussions and draft feedback. MAE was partially supported by NSERC Grant [2019-06167], CIFAR AI Chairs program, and CIFAR AI Catalyst grant.

References

Emmanuel Abbe, Enric Boix Adserà, 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, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 4782–4887. PMLR, 2022.

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2552–2623. PMLR, 12–15 Jul 2023.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR.

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.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 37932–37946. Curran Associates, Inc., 2022.

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 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Melih Barsbey, Milad Sefidgaran, Murat A. Erdogdu, Gaël Richard, and Umut Simsekli. Heavy tails in SGD and compressibility of overparametrized neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29364–29378, 2021.

Brian R. Bartoldson, Ari S. Morcos, Adrian Barbu, and Gordon Erlebacher. The generalization-stability tradeoff in neural network pruning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Sébastien Bubeck. Convex optimization: Algorithms and complexity, 2015.

Anthony Carbery and James Wright. Distributional and l-q norm inequalities for polynomials over convex bodies in r-n. *Mathematical Research Letters*, 8:233–248, 2001.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained BERT networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Lénaïc Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Trans. Mach. Learn. Res.*, 2022.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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 *Advances in Neural Information Processing Systems*, volume 36, pages 752–784. Curran Associates, Inc., 2023.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452. PMLR, 02–05 Jul 2022.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time, 2023.

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. *CoRR*, abs/2402.03220, 2024. doi: 10.48550/ARXIV.2402.03220.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Murat Erdogdu, Mohsen Bayati, and Lee H Dicker. Scalable approximations for generalized linear problems. *Journal of Machine Learning Research*, 20(7):1–45, 2019.

Murat A Erdogdu. Newton-stein method: a second order method for glms via stein's lemma. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1216–1224, 2015.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3259–3269. PMLR, 13–18 Jul 2020.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *ArXiv*, abs/1902.09574, 2019.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy learning in deep neural networks: an empirical study. *CoRR*, abs/1906.08034, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

C. Lee Giles and Christian W. Omlin. Pruning recurrent neural networks for improved generalization performance. *IEEE Trans. Neural Networks*, 5(5):848–851, 1994.

Larry Goldstein and Gesine Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability*, 7(4), November 1997. ISSN 1050-5164. doi: 10.1214/aoap/1043862419. URL http://dx.doi.org/10.1214/aoap/1043862419.

Larry Goldstein and Xiaohan Wei. Non-gaussian observations in nonlinear compressed sensing via stein discrepancies. *Information and Inference: A Journal of the IMA*, 8(1):125–159, 2019.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann, 1992.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8580–8589, 2018.

Tian Jin, Michael Carbin, Daniel M. Roy, Jonathan Frankle, and Gintare Karolina Dziugaite. Pruning's effect on generalization through the lens of training and regularization. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Tanishq Kumar, Kevin Luo, and Mark Sellke. No free prune: Information-theoretic barriers to pruning at initialization. *CoRR*, abs/2402.01089, 2024.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *Annals of Statistics*, 17: 1009–1052, 1989.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6682–6691. PMLR, 2020.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR, 2019.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a.

Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. In *Advances in Neural Information Processing Systems*, volume 36, pages 71449–71485. Curran Associates, Inc., 2023b.

Laurent Orseau, Marcus Hutter, and Omar Rivasplata. Logarithmic pruning is all you need. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE J. Sel. Areas Inf. Theory*, 1(1): 84–105, 2020.

Iosif Pinelis. Optimum Bounds for the Distributions of Martingales in Banach Spaces. *The Annals of Probability*, 22(4):1679 – 1706, 1994. doi: 10.1214/aop/1176988477.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over q-balls. *IEEE transactions on information theory*, 57(10): 6976–6994, 2011.

Abhishek Roy, Krishnakumar Balasubramanian, and Murat A Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. *Advances in Neural Information Processing Systems*, 34:8913–8926, 2021.

Terence Tao. Topics in random matrix theory. 2012.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 2010.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Hongru Yang, Yingbin Liang, Xiaojie Guo, Lingfei Wu, and Zhangyang Wang. Theoretical characterization of how neural network pruning affects its generalization, 2023. URL https://openreview.net/forum?id=dn6_PK73hAY.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 6594–6604, 2019.

Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3592–3602, 2019.

Michael Zhu and Suyog Gupta. To prune, or not to prune: Exploring the efficacy of pruning for model compression. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018.

Contents

Appendix A. Further Discussion for Section 4

In this section, we detail the examples discussed in Section 4. Recall that ϕ is the ReLU activation with the Hermite expansion $\phi = \sum_{k\geq 0} \frac{\tilde{\gamma}_k}{k!} H_{e_k}$. Notably, the coefficients are $\tilde{\gamma}_1 = \frac{1}{2}$, $\tilde{\gamma}_2 = \frac{1}{\sqrt{2\pi}}$, , $\tilde{\gamma}_3 = 0$, $\tilde{\gamma}_4 = \frac{-1}{\sqrt{2\pi}}$, and $\tilde{\gamma}_6 = \frac{3}{\sqrt{2\pi}}$ (see (C.5) with $b = 0$).

First, we consider the setting in (6.1). In this case, for $\boldsymbol{w} \in S^{d-1}$, we have

$$\begin{aligned}\mathbb{E}\left[y\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}\right] &= \sqrt{2}\tilde{\gamma}_2 \langle \boldsymbol{v}_1, \boldsymbol{w}\rangle \boldsymbol{v}_1 + \sqrt{2}\tilde{\gamma}_2 \langle \boldsymbol{v}_2, \boldsymbol{w}\rangle \boldsymbol{v}_2 + \tfrac{\tilde{\gamma}_4}{\sqrt{2}}(\langle \boldsymbol{w}, \boldsymbol{v}_1\rangle^2 + \langle \boldsymbol{w}, \boldsymbol{v}_2\rangle^2)\boldsymbol{w} \\ &= \tfrac{1}{\sqrt{\pi}} \langle \boldsymbol{e}_1, \boldsymbol{w}\rangle \boldsymbol{e}_1 + \tfrac{1}{\sqrt{\pi}} \langle \boldsymbol{e}_2, \boldsymbol{w}\rangle \boldsymbol{e}_2 - \tfrac{1}{2\sqrt{\pi}}\big(\langle \boldsymbol{w}, \boldsymbol{e}_1\rangle^2 + \langle \boldsymbol{w}, \boldsymbol{e}_2\rangle^2\big)\boldsymbol{w}, \end{aligned} \tag{A.1}$$

using an argument by Erdogdu et al. (2019) and

$$\mathbb{E}\left[\check{y}\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}\right] = \tilde{\gamma}_1 \boldsymbol{v} + \mathbb{E}\left[y\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}\right] = \tfrac{-1}{2\sqrt{\pi}}(\boldsymbol{e}_1 + \boldsymbol{e}_2) + \mathbb{E}\left[y\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}\right], \tag{A.2}$$

where we used the defined values in (6.1). From (A.1)-(A.2), we deduce

$$\mathbb{E}\left[y\phi'(\langle \boldsymbol{e}_i, \boldsymbol{x}\rangle)\boldsymbol{x}\right] = \tfrac{1}{2\sqrt{\pi}}\begin{cases}\boldsymbol{e}_1 & i = 1\\ \boldsymbol{e}_2 & i = 2\\ 0 & i > 2,\end{cases} \quad \text{and} \quad \mathbb{E}\left[\check{y}\phi'(\langle \boldsymbol{e}_i, \boldsymbol{x}\rangle)\boldsymbol{x}\right] = \tfrac{-1}{2\sqrt{\pi}}\begin{cases}\boldsymbol{e}_2 & i = 1\\ \boldsymbol{e}_1 & i = 2\\ \boldsymbol{e}_1 + \boldsymbol{e}_2 & i > 2,\end{cases}$$

confirming (6.2) and (6.3).

For (6.4), let us consider $\gamma_2 = 1$, $\gamma_4 = 2\sqrt{3}$, and $\varepsilon = e^{-d}$. Using (6.6), we can show that the population gradient in this case satisfies:

$$\|\mathbb{E}\left[y\phi'(\langle \boldsymbol{e}_i, \boldsymbol{x}\rangle)\boldsymbol{x}\right]\|_2 = \begin{cases} O\big(d^{\frac{1}{4}}e^{-d}\big), & i = 1\\ O(e^{-d}), & i = 2, \cdots, \sqrt{d}\\ 0, & i > \sqrt{d}.\end{cases}$$

We note that in this case, an exponentially large sample size in d is required to differentiate between $i = 1$ then $i = d$ using empirical gradients.

Appendix B. Preliminaries for Proofs

Additional Notation: Unless otherwise stated, Z follows the standard Gaussian distribution with a dimension depending on the context. We let $C_{\sigma^*} := \mathbb{E}[\|\nabla\sigma^*(Z)\|_2^2]^{1/2}$. We use S_M^{d-1} to denote the M-sparse d-dimensional unit vectors, i.e., $S_M^{d-1} := \{\boldsymbol{x} \in S^{d-1} \mid \|\boldsymbol{x}\|_0 \leq M\}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, $\sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots \geq \sigma_{d_1 \wedge d_2}(\boldsymbol{A})$ denotes the singular values of $\boldsymbol{A}$. For $\mathcal{J}_1 \subseteq [d_1]$ and $\mathcal{J}_2 \subseteq [d_2]$, we let $\boldsymbol{A}|_{\mathcal{J}_1},\ \boldsymbol{A}|_{\mathcal{J}_1 \times \mathcal{J}_2} \in \mathbb{R}^{d_1 \times d_2}$ such that

$$(\boldsymbol{A}|_{\mathcal{J}_1})_{ij} = \begin{cases} \boldsymbol{A}_{ij} & i \in \mathcal{J}_1 \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad (\boldsymbol{A}|_{\mathcal{J}_1 \times \mathcal{J}_2})_{ij} = \begin{cases} \boldsymbol{A}_{ij} & i \in \mathcal{J}_1 \text{ and } j \in \mathcal{J}_2 \\ 0 & \text{otherwise.} \end{cases}$$

In the following, $C, K > 0$ are constants that might take different values in different statements. For reader's convenience, we track on which variable they depend. For a set E,

$$\mathbb{1}_E(\boldsymbol{x}) := \begin{cases} 1 & \boldsymbol{x} \in E \\ 0 & \text{otherwise} \end{cases}$$

We use $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ to denote the dataset.

Additional Definitions: For notational simplicity, we assume that

$$|\sigma^*(\boldsymbol{z})| \leq C_1(1 + \|\boldsymbol{z}\|_2^2)^{C_2} \ \text{ for some }\ C_1 > 0,\ C_2 \geq \frac{1}{2}.$$

We note that since σ^* is a polynomial this assumption will always hold. Furthermore, in the proof, we particularly consider the model

$$y := \sigma^*(\boldsymbol{V}^\top \boldsymbol{x}) + \sqrt{\Delta}\epsilon, \tag{B.1}$$

where $\Delta > 0$ and ϵ has sub-Gaussian tails, i.e., $\mathbb{P}\left[|\epsilon| > t\right] \leq 2e^{-t^2}$.

We recall that $\phi(t) = \max\{0, t\}$ denotes the ReLU activation. To be precise, we define the initialization considered in Algorithm 2 mathematically as follows:

$$\boldsymbol{W}_{j*}^{(0)} = \left(\sum_{i \in \mathcal{J}} \boldsymbol{W}_{ji}^2\right)^{-1} (\boldsymbol{W}_{j1}\mathbb{1}_{1 \in \mathcal{J}}, \cdots, \boldsymbol{W}_{jd}\mathbb{1}_{d \in \mathcal{J}}) \tag{INIT}$$

where $\mathcal{J}$ is the output of `PruneNetwork` (see Algorithm 1), $\boldsymbol{W} \in \mathbb{R}^{m \times d}$, $\boldsymbol{W}_{ij} \sim_{iid} \mathcal{N}(0, 1)$, and $\boldsymbol{W}$ is independent of $\mathcal{D}$. As for definition (B.1), in the multi-index setting, we use

$$\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}\boldsymbol{z}^\top] := \boldsymbol{D} \in \mathbb{R}^{r \times r} \ \ \textit{and}\ \ \mathbb{E}[\sigma^*(\boldsymbol{V}^\top \boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top := \boldsymbol{H}, \tag{DEF-H}$$

which follows from the second order Stein's lemma (Erdogdu, 2015). Without loss of generality, we assume $\boldsymbol{D}$ is diagonal.

Appendix C. Hermite Expansion in the Multi-Index Setting

C.1. Background on Tensors

In the following, we will use the tensor representation of multivariate Hermite polynomials. Therefore, we introduce some new notation to work with tensors: We denote tensors with boldface uppercase letters, (e.g. $\boldsymbol{T}$). Unless specified, we assume that tensors take a value from an abstract inner product space, denoted with $\mathcal{H}$, with an inner product, of $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. For a k-tensor $\boldsymbol{T_k} : (\mathbb{R}^d)^{\otimes k} \to \mathcal{H}$ and an index tuple $(i_1, \cdots, i_k) \in [d]^k$, we use $\boldsymbol{T_k}|_{i_1\cdots i_k} \coloneqq \boldsymbol{T_k}[\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \cdots, \boldsymbol{e}_{i_k}]$, where $\{\boldsymbol{e}_i\}_{i\in[d]}$ is the standard basis for $\mathbb{R}^d$. We define the inner product and Frobenius norm for k-tensors $\boldsymbol{T_k}, \tilde{\boldsymbol{T}}_{\boldsymbol{k}} : (\mathbb{R}^d)^{\otimes k} \to \mathcal{H}$ as

$$\left\langle \boldsymbol{T_k}, \tilde{\boldsymbol{T}}_{\boldsymbol{k}} \right\rangle \coloneqq \sum_{(i_1,\cdots,i_k)\in[d]^k} \left\langle \boldsymbol{T_k}|_{i_1\cdots i_k}, \tilde{\boldsymbol{T}}_{\boldsymbol{k}}|_{i_1\cdots i_k} \right\rangle_{\mathcal{H}} \quad \text{and} \quad \|\boldsymbol{T_k}\|_F \coloneqq \sqrt{\langle \boldsymbol{T_k}, \boldsymbol{T_k}\rangle}. \tag{C.1}$$

We use $sym(\cdot)$ to denote symmetrization operator, i.e.,

$$sym(\boldsymbol{T_k})[\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \cdots, \boldsymbol{e}_{i_k}] = \frac{1}{k!} \sum_{\tau\in S_k} \boldsymbol{T_k}[\boldsymbol{e}_{\tau(i_1)}, \boldsymbol{e}_{\tau(i_2)}, \cdots, \boldsymbol{e}_{\tau(i_k)}] \tag{C.2}$$

where S_k is the set of permutations for $[k]$. We say a tensor is symmetric if $\boldsymbol{T_k} = sym(\boldsymbol{T_k})$. For a vector $\boldsymbol{u} \in \mathbb{R}^d$, $\boldsymbol{u}^{\otimes k} : (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}$ is a symmetric k-tensor defined as $\boldsymbol{u}^{\otimes k}[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k] = \prod_{i=1}^k \langle \boldsymbol{u}, \boldsymbol{v}_i\rangle$.

C.1.1. AUXILIARY TENSOR RESULTS

In this part, we present some useful tensor related result that we will use in the following.

Proposition 6 *Let $\boldsymbol{T_k} : (\mathbb{R}^d)^{\otimes k} \to \mathcal{H}$ be a symmetric k-tensor. For any k-tensor $\tilde{\boldsymbol{T}}_{\boldsymbol{k}}$, we have $\left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}, \boldsymbol{T_k} \right\rangle = \left\langle sym(\tilde{\boldsymbol{T}}_{\boldsymbol{k}}), \boldsymbol{T_k} \right\rangle$.*

Proof We have

$$\begin{aligned}
\left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}, \boldsymbol{T_k} \right\rangle &\overset{(a)}{=} \sum_{(i_1,\cdots,i_k)\in[d]^k} \frac{1}{k!} \sum_{\tau\in S_k} \left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}|_{i_1\cdots i_k}, \boldsymbol{T_k}[\boldsymbol{e}_{\tau(i_1)}, \cdots, \boldsymbol{e}_{\tau(i_k)}] \right\rangle \\
&\overset{(b)}{=} \sum_{(i_1,\cdots,i_k)\in[d]^k} \frac{1}{k!} \sum_{\tau\in S_k} \left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}|_{\tau(i_1)\cdots\tau(i_k)}, \boldsymbol{T_k}[\boldsymbol{e}_{i_1}, \cdots, \boldsymbol{e}_{i_k}] \right\rangle = \left\langle sym(\tilde{\boldsymbol{T}}_{\boldsymbol{k}}), \boldsymbol{T_k} \right\rangle,
\end{aligned}$$

where (a) follows since $\boldsymbol{T_k}$ is symmetric, and (b) follows by changing the indexing. ■

Lemma 7 *Let $\boldsymbol{T_{j+k}} : (\mathbb{R}^d)^{\otimes(j+k)} \to \mathbb{R}$ be a symmetric tensor. We define $\boldsymbol{\nabla^j T_{j+k}} : (\mathbb{R}^d)^{\otimes k} \to (\mathbb{R}^d)^{\otimes j}$ as*

$$\boldsymbol{\nabla^j T_{j+k}}[\boldsymbol{e}_{i_1}, \cdots, \boldsymbol{e}_{i_k}]|_{i_{k+1}\cdots i_{k+j}} \coloneqq \boldsymbol{T_{j+k}}[\boldsymbol{e}_{i_1}, \cdots, \boldsymbol{e}_{i_k}, \boldsymbol{e}_{i_{k+1}}, \cdots, \boldsymbol{e}_{i_{k+j}}]. \tag{C.3}$$

We have $\boldsymbol{\nabla^j T_{j+k}}$ is symmetric and $\|\boldsymbol{\nabla^j T_{j+k}}\|_F = \|\boldsymbol{T_{j+k}}\|_F$.

Proof Both statements follow from definitions in (C.1) and (C.2). ■

Lemma 8 *For $\boldsymbol{A} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{T_k} : (\mathbb{R}^r)^{\otimes k} \to \mathbb{R}$, let $\hat{\boldsymbol{T}}_{\boldsymbol{k}} : (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}$ such that $\hat{\boldsymbol{T}}_{\boldsymbol{k}}[\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k] = \boldsymbol{T_k}[\boldsymbol{A}^\top \boldsymbol{u}_1, \cdots, \boldsymbol{A}^\top \boldsymbol{u}_k]$. Then, $\|\hat{\boldsymbol{T}}_{\boldsymbol{k}}\|_F \geq \sigma_r^k(\boldsymbol{A})\|\boldsymbol{T_k}\|_F$.*

Proof Let singular value decomposition of $\boldsymbol{A}$ be $\boldsymbol{A} := \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{L}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{L} \in \mathbb{R}^{r\times r}$ are orthonormal vectors and $\boldsymbol{\Sigma}_{ii} = \sigma_i(\boldsymbol{A})$ for $i \in [r]$. First, we observe that for any $\boldsymbol{v} \in \mathbb{R}^d$ such that $\boldsymbol{v} \perp col(\boldsymbol{U})$, $\boldsymbol{A}^\top \boldsymbol{v} = 0$. Since Frobenius norm of a tensor is independent of the choice of basis, we can write that

$$\|\hat{\boldsymbol{T}}_{\boldsymbol{k}}\|_F^2 = \sum_{i_1,\cdots,i_k\in[r]^k} \tilde{\boldsymbol{T}}_{\boldsymbol{k}}\left[\boldsymbol{U}_{*i_1},\cdots,\boldsymbol{U}_{*i_k}\right]^2.$$

Hence, by definition

$$\|\hat{\boldsymbol{T}}_{\boldsymbol{k}}\|_F^2 = \sum_{i_1,\cdots,i_k\in[r]^k} \boldsymbol{T}_{\boldsymbol{k}}\left[\sigma_{i_1}(\boldsymbol{A})\boldsymbol{L}_{*i_1},\cdots,\sigma_{i_k}(\boldsymbol{A})\boldsymbol{L}_{*i_k}\right]^2 \overset{(a)}{\geq} \sigma_r^{2k}(\boldsymbol{A}) \sum_{i_1,\cdots,i_k\in[r]^k} \boldsymbol{T}_{\boldsymbol{k}}\left[\boldsymbol{L}_{*i_1},\cdots,\boldsymbol{L}_{*i_k}\right]^2$$
$$= \sigma_r^{2k}(\boldsymbol{A})\|\boldsymbol{T}_{\boldsymbol{k}}\|_F^2,$$

where we use the multi-linear property of tensors in (a). ∎

Lemmas for Hermite Tensors

Definition 9 (Hermite Tensors) *We define the Hermite tensor with a degree of k as $\boldsymbol{H}_{\boldsymbol{e_k}} : \mathbb{R}^d \to (\mathbb{R}^d)^{\otimes k}$ as*

$$\boldsymbol{H}_{\boldsymbol{e_k}}(\boldsymbol{x})|_{i_1,\cdots,i_k} := e^{\frac{\|\boldsymbol{x}\|_2^2}{2}}(-1)^k \frac{\partial^k}{\partial \boldsymbol{x}_{i_1}\cdots\partial \boldsymbol{x}_{i_k}}\left(e^{\frac{-\|\boldsymbol{x}\|_2^2}{2}}\right).$$

We use the following facts about Hermite tensors in our proofs.

Lemma 10 *For any orthonormal basis $\{\boldsymbol{b}_1,\cdots,\boldsymbol{b}_d\}$ and $\boldsymbol{x} \in \mathbb{R}^d$, we have*

$$\langle \boldsymbol{H}_{\boldsymbol{e_k}}(\boldsymbol{x}), \boldsymbol{b}_{i_1}\otimes\cdots\otimes\boldsymbol{b}_{i_d}\rangle = H_{e_{j_1}}(\langle \boldsymbol{b}_1,\boldsymbol{x}\rangle)\cdots H_{e_{j_d}}(\langle \boldsymbol{b}_d,\boldsymbol{x}\rangle),$$

where j_l is the number of occurrences of $l \in [d]$ in $(i_1,\cdots,i_k)$, i.e., $j_l = \mathbb{1}_{i_1=l} + \cdots + \mathbb{1}_{i_k=l}$.

Proof If $\{\boldsymbol{b}_1,\cdots,\boldsymbol{b}_d\}$ is the standard basis, the statement follows from Definition 9. To extend it for any orthonormal basis, let $\boldsymbol{B}$ denote the matrix with columns $\{\boldsymbol{b}_1,\cdots,\boldsymbol{b}_d\}$, let $h(\boldsymbol{x}) := \exp\left(-\|\boldsymbol{x}\|_2^2/2\right)$ and let $\nabla^k h(\boldsymbol{x}) : (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}$ represent the k^{th} derivative of h. We want to prove that for any $(i_1,\cdots,i_k) \in [d]^k$, $\nabla^k h(\boldsymbol{x})[\boldsymbol{B}\boldsymbol{e}_{i_1},\cdots,\boldsymbol{B}\boldsymbol{e}_{i_k}] \overset{(*)}{=} \nabla^k h(\boldsymbol{B}^\top\boldsymbol{x})[\boldsymbol{e}_{i_1},\cdots,\boldsymbol{e}_{i_k}]$, which will prove the statement. We will use proof by induction. We observe that $(*)$ holds for $k=1$. For $k>1$, by assuming $(*)$ holds for $k-1$, we have

$$\nabla^k h(\boldsymbol{x})[\boldsymbol{B}\boldsymbol{e}_{i_1},\cdots,\boldsymbol{B}\boldsymbol{e}_{i_k}] = \lim_{t\to 0}\frac{\left(\nabla^{k-1}h(\boldsymbol{x}+t\boldsymbol{B}\boldsymbol{e}_{i_k}) - \nabla^{k-1}h(\boldsymbol{x})\right)\left[\boldsymbol{B}\boldsymbol{e}_{i_1},\cdots,\boldsymbol{B}\boldsymbol{e}_{i_{k-1}}\right]}{t}$$
$$= \lim_{t\to 0}\frac{\left(\nabla^{k-1}h(\boldsymbol{B}^\top\boldsymbol{x}+t\boldsymbol{e}_{i_k}) - \nabla^{k-1}h(\boldsymbol{B}^\top\boldsymbol{x})\right)\left[\boldsymbol{e}_{i_1},\cdots,\boldsymbol{e}_{i_{k-1}}\right]}{t}$$
$$= \nabla^k h(\boldsymbol{B}^\top\boldsymbol{x})[\boldsymbol{e}_{i_1},\cdots,\boldsymbol{e}_{i_k}].$$

∎

Corollary 11 *Let $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ be an orthonormal matrix and $\boldsymbol{T_k} : (\mathbb{R}^r)^{\otimes k} \to \mathbb{R}$ be a symmetric k-tensor, and $\boldsymbol{H^{(r)}_{e_k}}$ and $\boldsymbol{H^{(d)}_{e_k}}$ denote k-degree Hermite tensor defined on $\mathbb{R}^r$ and $\mathbb{R}^d$ respectively. For $\tilde{\boldsymbol{T}}_{\boldsymbol{k}}[\boldsymbol{e}_{i_1}, \cdots, \boldsymbol{e}_{i_k}] := \boldsymbol{T_k}[\boldsymbol{V}^\top \boldsymbol{e}_{i_1}, \cdots, \boldsymbol{V}^\top \boldsymbol{e}_{i_k}]$, we have $\left\langle \boldsymbol{T_k}, \boldsymbol{H^{(r)}_{e_k}}(\boldsymbol{V}^\top \boldsymbol{x})\right\rangle = \left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}, \boldsymbol{H^{(d)}_{e_k}}(\boldsymbol{x})\right\rangle$.*

Proof It immediately follows from Lemma 10. ■

Lemma 12 *We have $\boldsymbol{H_{e_k}}(0) = (-i)^k \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\boldsymbol{w}^{\otimes k}\right]$, where $i = \sqrt{-1}$. Consequently, we have $\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\boldsymbol{w}^{\otimes 2k}\right] = (2k-1)!!sym(\boldsymbol{I}_d^{\otimes k})$.*

Proof See (Tao, 2012, Eqs. 2.159 and 2.160) and (Damian et al., 2022, Lemma 22). ■

C.2. Hermite Expansion of the Population Gradient

For a symmetric $(k+1)$-tensor $\boldsymbol{T_{k+1}} : (\mathbb{R}^r)^{\otimes k+1} \to \mathbb{R}$, we define a k-tensor $\boldsymbol{\nabla T_{k+1}} : (\mathbb{R}^r)^{\otimes k} \to \mathbb{R}^r$ as in (C.3) with $j = 1$. For the following, we use the following notation: For $b \in \mathbb{R}$,

$$\phi(\cdot + b) := \sum_{k\geq 0} \frac{\tilde{\gamma}_k(b)}{k!} H_{e_k} \text{ and } \sigma^* := \sum_{k\geq 0} \frac{1}{k!} \langle \boldsymbol{T_k}, \boldsymbol{H_{e_k}} \rangle ,$$

where $\tilde{\gamma}_k(b) \in \mathbb{R}$ and $\boldsymbol{T_k}$ is a symmetric k-tensor for $k \in \mathbb{N}$. The main statement of this part is given below.

Proposition 13 *For an orthonormal matrix $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{w} \in S^{d-1}$, we have*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\sigma^*(\boldsymbol{V}^\top \boldsymbol{x})\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}] = \boldsymbol{V} \sum_{k\geq 0} \tfrac{\tilde{\gamma}_{k+1}(b)}{k!} \boldsymbol{\nabla T_{k+1}}\Big[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\Big] \\ + \boldsymbol{w} \sum_{k\geq 0} \tfrac{\tilde{\gamma}_{k+2}(b)}{k!} \boldsymbol{T_k}\Big[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\Big] \end{aligned} \tag{C.4}$$

and

$$\tilde{\gamma}_k(b) = \begin{cases} 1 - \Phi(-b), & k = 1 \\ \frac{e^{\frac{-b^2}{2}}}{\sqrt{2\pi}} H_{e_{k-2}}(-b), & k \geq 2 \end{cases} \tag{C.5}$$

where $\Phi(b)$ is the CDF of the standard Gaussian distribution.

To prove Proposition 13, we will need two lemmas.

Lemma 14 *For $\boldsymbol{w} \in \mathbb{R}^d$ and $k \in \mathbb{N}$, let $\boldsymbol{T_k} := k\, sym(\boldsymbol{e}_l \otimes \boldsymbol{w}^{\otimes k-1})$. For $i_1, \cdots, i_k \in [d]$, we have $\boldsymbol{T_k}|_{i_1\cdots i_k} = j_l \boldsymbol{w}_1^{j_1} \times \cdots \times \boldsymbol{w}_l^{j_l - 1} \times \cdots \times \boldsymbol{w}_d^{j_d}$, where $j_l = \mathbb{1}_{i_1=l} + \cdots + \mathbb{1}_{i_k=l}$.*

Proof We have $\boldsymbol{T_k} \overset{(*)}{=} \boldsymbol{e}_l \otimes \boldsymbol{w}^{\otimes k-1} + \boldsymbol{w} \otimes \boldsymbol{e}_l \otimes \boldsymbol{w}^{\otimes k-2} + \boldsymbol{w}^{\otimes 2} \otimes \boldsymbol{e}_l \otimes \boldsymbol{w}^{\otimes k-3} + \cdots + \boldsymbol{w}^{\otimes k-1} \otimes \boldsymbol{e}_l$. Without loss of generality, we can assume $j_l > 0$ and $i_1, \cdots, i_{j_l} = l$ (since for $j_l = 0$, the statement is true). The statement follows from $(*)$ since in the right-hand side only j_l terms will be nonzero and the other terms will be equal to $\boldsymbol{w}^{\otimes k-1}|_{i_2,\cdots,i_k} = \boldsymbol{w}_1^{j_1} \times \cdots \times \boldsymbol{w}_l^{j_l - 1} \times \cdots \times \boldsymbol{w}_d^{j_d}$. ■

Lemma 15 *For* $\boldsymbol{w} \in S^{d-1}$, $l \in [d]$ *and* $k \in \mathbb{N}$, *we have* $\mathbb{E}\left[\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}_l \boldsymbol{H}_{\boldsymbol{e_k}}(\boldsymbol{x})\right] = \tilde{\gamma}_{k+2}(b)\boldsymbol{w}_l \boldsymbol{w}^{\otimes k} + \tilde{\gamma}_k(b)k\ sym(\boldsymbol{e}_l \otimes \boldsymbol{w}^{\otimes k-1})$.

Proof We recall that $\boldsymbol{H}_{\boldsymbol{e_k}}(\boldsymbol{x})|_{i_1\cdots i_k} = H_{e_{j_1}}(\boldsymbol{x}_1)\cdots H_{e_{j_d}}(\boldsymbol{x}_d)$, where $j_l = \mathbb{1}_{i_1=l} + \cdots + \mathbb{1}_{i_k=l}$. The for any fixed $(i_1, \cdots, i_k) \in [d]^k$,

$$
\begin{aligned}
\mathbb{E}\left[\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}_l \boldsymbol{H}_{\boldsymbol{e_k}}(\boldsymbol{x})|_{i_1\cdots i_k}\right] &= \mathbb{E}\left[\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)H_{e_{j_1}}(\boldsymbol{x}_1)\cdots H_{e_{j_l+1}}(\boldsymbol{x}_l)\cdots H_{e_{j_d}}(\boldsymbol{x}_d)\right] \\
&\quad + j_l\mathbb{E}\left[\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)H_{e_{j_1}}(\boldsymbol{x}_1)\cdots H_{e_{j_l-1}}(\boldsymbol{x}_l)\cdots H_{e_{j_d}}(\boldsymbol{x}_d)\right] \\
&= \tilde{\gamma}_{k+2}(b)\boldsymbol{w}_1^{j_1}\cdots \boldsymbol{w}_i^{j_l+1}\cdots \boldsymbol{w}_d^{j_d} + \tilde{\gamma}_k(b) j_l \boldsymbol{w}_1^{j_1}\cdots \boldsymbol{w}_i^{j_l-1}\cdots \boldsymbol{w}_d^{j_d} \\
&= \tilde{\gamma}_{k+2}(b)\boldsymbol{w}_l \boldsymbol{w}^{\otimes k}|_{i_1\cdots i_k} + \tilde{\gamma}_k(b)k\ sym(\boldsymbol{e}_l \otimes \boldsymbol{w}^{\otimes k-1})|_{i_1\cdots i_k},
\end{aligned}
$$

where we use Lemma 14 in the last line. ∎

Proof [Proof of Proposition 13] We fix $l \in [d]$. Since $\mathbb{E}[\phi(Z)^4] < \infty$, we have

$$
\begin{aligned}
\mathbb{E}[\sigma^*(\boldsymbol{V}^\top \boldsymbol{x})\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}_l] &= \sum_{k=0}^{\infty} \frac{1}{k!}\mathbb{E}\left[\left\langle \boldsymbol{T_k}, \boldsymbol{H}_{\boldsymbol{e_k}}(\boldsymbol{V}^\top \boldsymbol{x})\right\rangle \phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}_l\right] \\
&= \sum_{k=0}^{\infty} \frac{1}{k!}\left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}, \mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{e_k}}(\boldsymbol{x})\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}_l\right]\right\rangle, \qquad \text{(C.6)}
\end{aligned}
$$

where $\tilde{\boldsymbol{T}}_{\boldsymbol{k}}$ is defined in Corollary 11. For a fixed $k \in \mathbb{N}$, we have

$$
\begin{aligned}
\left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}, \mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{e_k}}(\boldsymbol{x})\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}_l\right]\right\rangle &\overset{(a)}{=} \tilde{\gamma}_{k+2}(b)\boldsymbol{w}_l\left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}, \boldsymbol{w}^{\otimes k}\right\rangle + \tilde{\gamma}_k(b)k\left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}, \boldsymbol{e}_l \otimes \boldsymbol{w}^{\otimes k-1}\right\rangle \\
&= \tilde{\gamma}_{k+2}(b)\boldsymbol{w}_l \boldsymbol{T_k}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right] + \tilde{\gamma}_k(b)k\ \boldsymbol{V}_{l*}^\top \boldsymbol{\nabla} \boldsymbol{T_k}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k-1}\right], \qquad \text{(C.7)}
\end{aligned}
$$

where (a) follows by Proposition 6 since $\tilde{\boldsymbol{T}}_{\boldsymbol{k}}$ symmetric. (C.4) follows from (C.6) and (C.7). For (C.5), see (Ba et al., 2023, Lemma 15). ∎

Corollary 16 *Let* $\phi_\pm(t, ; b) := \frac{\phi(t+b) \pm \phi(-t+b)}{2}$. *We have*

$$
\mathbb{E}[\sigma^*(\boldsymbol{V}^\top \boldsymbol{x})\phi'_+(\langle \boldsymbol{w}, \boldsymbol{x}\rangle; b)\boldsymbol{x}] = \boldsymbol{V}\sum_{\substack{k\geq 1 \\ k\ odd}} \frac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla} \boldsymbol{T_{k+1}}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right] + \boldsymbol{w}\sum_{\substack{k\geq 0 \\ k\ even}} \frac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right]
$$

$$
\mathbb{E}[\sigma^*(\boldsymbol{V}^\top \boldsymbol{x})\phi'_-(\langle \boldsymbol{w}, \boldsymbol{x}\rangle; b)\boldsymbol{x}] = \boldsymbol{V}\sum_{\substack{k\geq 0 \\ k\ even}} \frac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla} \boldsymbol{T_{k+1}}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right] + \boldsymbol{w}\sum_{\substack{k\geq 1 \\ k\ odd}} \frac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right]
$$

Proof We observe that $\phi_+(\cdot + b) = \sum_{\substack{k\geq 0 \\ k\ even}} \frac{\tilde{\gamma}_k(b)}{k!}H_{e_k}$ and $\phi_-(\cdot + b) = \sum_{\substack{k\geq 0 \\ k\ odd}} \frac{\tilde{\gamma}_k(b)}{k!}H_{e_k}$. By the argument in (C.6) and (C.7), the statement follows. ∎

C.3. Bounding the Higher Order Terms in the Hermite Expansion

Proposition 17 *For $N \in \mathbb{N} \cup \{-1, 0\}$, $\boldsymbol{w} \in S^{d-1}$ and $b \in \mathbb{R}$, let*

$$\zeta_N := \mathbb{E}\Big[\sigma^*(\boldsymbol{V}^\top \boldsymbol{x})\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)x\Big] - \boldsymbol{V}\sum_{k=0}^{N} \tfrac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right] - \boldsymbol{w}\sum_{k=0}^{N} \tfrac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right].$$

We have

$$\|\zeta_N\|_2 \leq (1+\sqrt{N+2})C_{\sigma^*}\begin{cases}\frac{\|\boldsymbol{V}^\top \boldsymbol{w}\|_2^{N+1}}{1-\|\boldsymbol{V}^\top \boldsymbol{w}\|^2} & \|\boldsymbol{V}^\top \boldsymbol{w}\|_2 > 0 \text{ or } N \geq 0 \\ 1 & \text{otherwise}.\end{cases}$$

Proof [Proof of Proposition 17] By Proposition 13, we know that

$$\zeta_N = \boldsymbol{V}\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right] + \boldsymbol{w}\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right].$$

Therefore,

$$\|\zeta_N\|_2 \overset{(a)}{=} \left\|\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right]\right\|_2 + \left|\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top \boldsymbol{w})^{\otimes k}\right]\right| \tag{C.8}$$

$$\begin{aligned}&\overset{(b)}{\leq} \left(\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+1}^2(b)\left\|\boldsymbol{V}^\top \boldsymbol{w}\right\|_2^{2k}}{k!}\right)^{\frac{1}{2}}\left(\sum_{k\geq N+1}\frac{1}{k!}\left\|\boldsymbol{\nabla T_{k+1}}\left[\left(\frac{\boldsymbol{V}^\top \boldsymbol{w}}{\|\boldsymbol{V}^\top \boldsymbol{w}\|_2}\right)^{\otimes k}\right]\right\|_2^2\right)^{\frac{1}{2}} \\ &\quad + \left(\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+2}^2(b)\left\|\boldsymbol{V}^\top \boldsymbol{w}\right\|_2^{2k}}{k!}\right)^{\frac{1}{2}}\left(\sum_{k\geq N+1}\frac{1}{k!}\boldsymbol{T_k}\left[\left(\frac{\boldsymbol{V}^\top \boldsymbol{w}}{\|\boldsymbol{V}^\top \boldsymbol{w}\|_2}\right)^{\otimes k}\right]^2\right)^{\frac{1}{2}} \\ &\overset{(c)}{\leq} \left(\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+1}^2(b)\left\|\boldsymbol{V}^\top \boldsymbol{w}\right\|_2^{2k}}{k!}\right)^{\frac{1}{2}}\mathbb{E}[\|\nabla\sigma^*(\boldsymbol{z})\|_2^2]^{\frac{1}{2}} + \left(\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+2}^2(b)\left\|\boldsymbol{V}^\top \boldsymbol{w}\right\|_2^{2k}}{k!}\right)^{\frac{1}{2}}\mathbb{E}[\sigma^*(\boldsymbol{z})_2^2]^{\frac{1}{2}}\end{aligned} \tag{C.9}$$

where we use that $\boldsymbol{V}$ is orthonormal and $\boldsymbol{w}$ is a unit vector in (a), the multi-linear property of tensors and Cauchy-Schwartz inequality for (b), and Parseval's identity for (c). We observe that for $\|\boldsymbol{V}^\top \boldsymbol{w}\|_2 > 0$ or $N \geq 0$

$$\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+1}^2(b)\left\|\boldsymbol{V}^\top \boldsymbol{w}\right\|_2^{2k}}{k!} \leq \left(\sup_{k\geq N+1}\frac{\tilde{\gamma}_{k+1}^2(b)}{k!}\right)\sum_{k\geq N+1}\left\|\boldsymbol{V}^\top \boldsymbol{w}\right\|_2^{2k} \leq \frac{\|\boldsymbol{V}^\top \boldsymbol{w}\|_2^{2(N+1)}}{1-\|\boldsymbol{V}^\top \boldsymbol{w}\|_2^2} \tag{C.10}$$

and

$$\sum_{k\geq N+1} \frac{\tilde{\gamma}_{k+2}^2(b)\left\|\boldsymbol{V}^\top \boldsymbol{w}\right\|_2^{2k}}{k!} \leq \left(\sup_{k\geq N+1}\frac{\tilde{\gamma}_{k+2}^2(b)}{(k+1)!}\right)\sum_{k\geq N+1}(k+1)\left\|\boldsymbol{V}^\top \boldsymbol{w}\right\|_2^{2k} \leq \frac{(N+2)\|\boldsymbol{V}^\top \boldsymbol{w}\|^{2(N+1)}}{(1-\|\boldsymbol{V}^\top \boldsymbol{w}\|^2)^2} \tag{C.11}$$

where we used $\sum_{k\geq 0}\frac{\tilde{\gamma}^2_{k+1}(b)}{k!}=\mathbb{E}[\phi'(Z+b)]\leq 1$ and the sum formula for $\sum_{k\geq k^\star} kz^{k+1}$. Since $\mathbb{E}[\sigma^*(\boldsymbol{z})_2^2]\leq \mathbb{E}[\|\nabla\sigma^*(\boldsymbol{z})\|_2^2]^{1/2}=C_{\sigma^*}$ and $\|\boldsymbol{V}^\top\boldsymbol{w}\|_2\leq 1$, we have

$$(\text{C.9})\leq (1+\sqrt{N+2})C_{\sigma^*}\frac{\|\boldsymbol{V}^\top\boldsymbol{w}\|^{N+1}}{1-\|\boldsymbol{V}^\top\boldsymbol{w}\|^2}. \tag{C.12}$$

For $\|\boldsymbol{V}^\top\boldsymbol{w}\|_2>0$ or $N\geq 0$ do not hold, we observe that the right-hand-side of both (C.10)- (C.11) is 1. Therefore, by the argument in (C.12), the statement follows in this case too. ■

Corollary 18 *Let $\phi_\pm$ be the functions introduced in Corollary 16. For For $N\in\mathbb{N}\cup\{-1,0\}$, $\boldsymbol{w}\in S^{d-1}$ and $b\in\mathbb{R}$, let*

$$\begin{aligned}\zeta_N^+ &:= \mathbb{E}\left[\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})\phi'_+(\langle\boldsymbol{w},\boldsymbol{x}\rangle;b)\boldsymbol{x}\right]\\ &\qquad - \boldsymbol{V}\sum_{\substack{k=0\\ k\ odd}}^{N}\frac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right]-\boldsymbol{w}\sum_{\substack{k=0\\ k\ even}}^{N}\frac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right],\\ \zeta_N^- &:= \mathbb{E}\left[\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})\phi'_-(\langle\boldsymbol{w},\boldsymbol{x}\rangle;b)\boldsymbol{x}\right]\\ &\qquad - \boldsymbol{V}\sum_{\substack{k=0\\ k\ even}}^{N}\frac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right]-\boldsymbol{w}\sum_{\substack{k=0\\ k\ odd}}^{N}\frac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right].\end{aligned}$$

We have

$$\|\zeta_N^\pm\|_2\leq(1+\sqrt{N+2})C_{\sigma^*}\begin{cases}\frac{\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^{N+1}}{1-\|\boldsymbol{V}^\top\boldsymbol{w}\|^2} & \|\boldsymbol{V}^\top\boldsymbol{w}\|_2>0 \text{ or } N\geq 0\\ 1 & \text{otherwise}\end{cases}$$

Proof The statement follows from $\mathbb{E}[\phi'_\pm(Z+b)^2]\leq 1$ and Proposition 17 (see (C.10) and (C.11)). ■

C.4. Bounding ℓ_q Norm of the Higher-Order Terms

Proposition 19 *By using the notation of Proposition 17 and Corollary 18, for $\boldsymbol{w}\in S^{d-1}$, $N\in\mathbb{N}\cup\{-1,0\}$ and $q\in[0,2)$, we have*

$$\|\zeta_N\|_q^q\vee\|\zeta_N^\pm\|_q^q\leq 2^{(q-1)\vee 0}C_{\sigma^*}^q\left[\|\boldsymbol{V}\|_{2,q}^q+(N+2)^{\frac{q}{2}}\|\boldsymbol{w}\|_q^q\right]\begin{cases}\left(\frac{\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^{N+1}}{1-\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^2}\right)^q & \|\boldsymbol{V}^\top\boldsymbol{w}\|_2>0 \text{ or } N\geq 0\\ 1 & \text{otherwise.}\end{cases}$$

Proof By Propositions 13 and 72, if $\|\boldsymbol{V}^\top\boldsymbol{w}\|_2 > 0$ or $N \geq 0$ hold, we have

$$\begin{aligned}
\|\zeta_N\|_q^q &\overset{(a)}{\leq} 2^{(q-1)\vee 0}\left(\|\boldsymbol{V}\|_{2,q}^q\left\|\sum_{k\geq N+1}\frac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\big[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\big]\right\|_2^q + \|\boldsymbol{w}\|_q^q\left|\sum_{k\geq N+1}\frac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right]\right|^q\right) \\
&\overset{(b)}{\leq} 2^{(q-1)\vee 0}\|\boldsymbol{V}\|_{2,q}^q C_{\sigma^*}^q\left(\frac{\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^{N+1}}{1-\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^2}\right)^q + 2^{(q-1)\vee 0}\|\boldsymbol{w}\|_q^q C_{\sigma^*}^q\left(\frac{\sqrt{N+2}\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^{N+1}}{1-\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^2}\right)^q \\
&= 2^{(q-1)\vee 0}C_{\sigma^*}^q\left(\frac{\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^{N+1}}{1-\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^2}\right)^q\left[\|\boldsymbol{V}\|_{2,q}^q + (N+2)^{\frac{q}{2}}\|\boldsymbol{w}\|_q^q\right],
\end{aligned}$$

where (a) follows $\|\boldsymbol{V}\boldsymbol{u}\|_q^q \leq \|\boldsymbol{V}\|_{2,q}^q\|\boldsymbol{u}\|_2^q$ and (b) follows the steps in (C.8)- (C.11). For $\|\zeta_N^{\pm}\|_q^q$, the same argument applies. if neither $\|\boldsymbol{V}^\top\boldsymbol{w}\|_2 > 0$ nor $N \geq 0$ hold, since we can replace $\frac{\sqrt{N+2}\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^{N+1}}{1-\|\boldsymbol{V}^\top\boldsymbol{w}\|_2^2}$ in (b) with 1, the statement follows in this case as well. ■

Appendix D. Concentration Bound for Empirical Gradients

In this part, we derive a concentration bound for the empirical gradient

$$g(\boldsymbol{w}, b) \coloneqq \frac{1}{n}\sum_{i=1}^n\left(y_i - \langle\hat{\boldsymbol{\mu}}|_{\mathcal{J}}, \boldsymbol{x}_i\rangle\right)\boldsymbol{x}_i\phi'\left(\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle + b\right), \tag{D.1}$$

where $\hat{\boldsymbol{\mu}} = 0$ in the single index setting and $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{j=1}^n y_j\boldsymbol{x}_j$ in the multi index setting. In the following, to avoid repetitions, we will consider (D.1) with $\phi(t) \in \{t, \mathrm{ReLU}(t)\}$ and particularly with $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{j=1}^n y_j\boldsymbol{x}_j$. Our proof will give us a bound for the $\hat{\boldsymbol{\mu}} = 0$ case as well.

To handle dependencies between $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and $\mathcal{J}$, we will consider the following process: For $\theta \coloneqq (\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}$,

$$\begin{aligned}
\boldsymbol{T}_\theta &\coloneqq g(\theta) - \mathbb{E}_{(\boldsymbol{x},y)}\left[\overline{y}\boldsymbol{x}\phi'(\langle\boldsymbol{w}, \boldsymbol{x}\rangle + b)\right] \\
&= \frac{1}{n}\sum_{i=1}^n\tilde{y}_i\boldsymbol{x}_i\phi'(\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle + b) - \mathbb{E}_{(\boldsymbol{x},y)}\left[\overline{y}\boldsymbol{x}\phi'(\langle\boldsymbol{w}, \boldsymbol{x}\rangle + b)\right],
\end{aligned}$$

where $(\boldsymbol{x}, y)$ is a generic data point that is independent of $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and

$$\tilde{y}_i = y_i - \langle\hat{\boldsymbol{\mu}}|_{\mathcal{J}}, \boldsymbol{x}_i\rangle \quad \text{and} \quad \overline{y} = y - \langle\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}}, \boldsymbol{x}\rangle\,. \tag{D.2}$$

We particularly derive a concentration bound for

$$\sup_{\substack{\mathcal{J}\subseteq[d]\\|\mathcal{J}|=M'}}\ \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}}\|\boldsymbol{T}_\theta|_{\mathcal{J}}\|_2, \tag{D.3}$$

where $M, M' \in [d]$, and the restriction sets in (D.2) and (D.3), i.e., $\mathcal{J}$, are the same. We observe that for a fixed $(\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}$,

$$\begin{aligned}
\boldsymbol{T}_\theta = &\left(\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i \phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) - \mathbb{E}\left[y\boldsymbol{x}\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\right]\right) \\
&- \left(\frac{1}{n}\sum_{i=1}^n \langle \mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}}, \boldsymbol{x}_i\rangle \boldsymbol{x}_i \phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) - \mathbb{E}_{(\boldsymbol{x},y)}\left[\langle \mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}}, \boldsymbol{x}\rangle \boldsymbol{x}\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\right]\right) \\
&- \left(\frac{1}{n}\sum_{i=1}^n \langle (\hat{\boldsymbol{\mu}} - \mathbb{E}[y\boldsymbol{x}])|_{\mathcal{J}}, \boldsymbol{x}_i\rangle \boldsymbol{x}_i \phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) - \mathbb{E}\left[\langle (\hat{\boldsymbol{\mu}} - \mathbb{E}[y\boldsymbol{x}])|_{\mathcal{J}}, \boldsymbol{x}\rangle \boldsymbol{x}\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\right]\right) \\
&- \mathbb{E}\left[\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}\boldsymbol{x}^\top\right](\hat{\boldsymbol{\mu}} - \mathbb{E}[y\boldsymbol{x}])|_{\mathcal{J}}.
\end{aligned}$$

Let

$$\begin{aligned}
\boldsymbol{Y}_\theta &\coloneqq \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i \phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) - \mathbb{E}_{(\boldsymbol{x},y)}\left[y\boldsymbol{x}\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\right], \\
\boldsymbol{\Sigma}_\theta &\coloneqq \frac{1}{n}\sum_{i=1}^n \phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b)\boldsymbol{x}_i\boldsymbol{x}_i^\top - \mathbb{E}_{\boldsymbol{x}}\left[\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}\boldsymbol{x}^\top\right].
\end{aligned}$$

Then, we can write

$$\boldsymbol{T}_\theta|_{\mathcal{J}} = \boldsymbol{Y}_\theta|_{\mathcal{J}} - \boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}} - \left(\boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}} + \mathbb{E}\left[\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}\boldsymbol{x}^\top\right]|_{\mathcal{J}\times\mathcal{J}}\right)(\hat{\boldsymbol{\mu}} - \mathbb{E}[y\boldsymbol{x}])|_{\mathcal{J}}. \tag{D.4}$$

In the following, we derive concentration bounds for $\boldsymbol{Y}_\theta$ and $\boldsymbol{\Sigma}_\theta$, which will lead us a bound for (D.4). Our proof technique relies on the use of Radamacher averages with an extension of the symmetrization lemma for the moment-generating function, which is presented as follows:

Lemma 20 *Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n \in \mathbb{R}^d$ be independent random vectors and let $\{\varepsilon_i\}_{i\in[n]}$ be iid Radamacher random variables, independent of $\{\boldsymbol{X}_i\}_{i\in[n]}$. For $\ell : \mathbb{R}^d \times S_M^{d-1} \times \mathbb{R}$, $\lambda > 0$ and $h(t) \in \{t, \exp(t)\}$, we have*

$$\mathbb{E}\left[h\left(\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\ b\in\mathbb{R}}} \frac{\lambda}{n}\sum_{i=1}^n \ell(\boldsymbol{X}_i, (\boldsymbol{w}, b)) - \mathbb{E}[\ell(\boldsymbol{X}, (\boldsymbol{w}, b))]\right)\right] \leq \mathbb{E}\left[\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\ b\in\mathbb{R}}} h\left(\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i \ell(\boldsymbol{X}_i, (\boldsymbol{w}, b))\right)\right].$$

Proof Let $Z \coloneqq \sup_{\boldsymbol{w},b} \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{X}_i, (\boldsymbol{w}, b)) - \mathbb{E}[\ell(\boldsymbol{X}, (\boldsymbol{w}, b))]$. By using Jensen's inequality, one can show that for any convex and nondecreasing function h,

$$\mathbb{E}[h(Z)] \leq \mathbb{E}\left[\sup_{\boldsymbol{w},b} h\left(\frac{2}{n}\sum_{i=1}^n \varepsilon_i \ell\left(\boldsymbol{X}_i, (\boldsymbol{w}, b)\right)\right)\right].$$

Since $t \to h(\lambda t)$, where $h(t) \in \{t, \exp(t)\}$ and $\lambda > 0$, is convex and nondecreasing, the statement follows. ■

D.1. VC Dimension of $\{\cdot \to \phi'(\langle \boldsymbol{w}, \cdot \rangle + b);\, (\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}\}$

Let $\mathcal{F}_M := \{\cdot \to \phi'(\langle \boldsymbol{w}, \cdot \rangle + b) \mid (\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}\}$. We want to bound the VC dimension of $\mathcal{F}_M$.

Proposition 21 *Let $VC(\mathcal{F}_M) = d^*$. We have $M \leq d^* \leq 6M \log\left(\frac{ed}{M}\right)$.*

Proof Let $\mathcal{F}^{(d)} := \{\cdot \to \phi'(\langle \boldsymbol{w}, \cdot \rangle + b) \mid (\boldsymbol{w}, b) \in S^{d-1} \times \mathbb{R}\}$ and $s(\mathcal{F}^{(d)}, n)$ be the shattering coefficient of $\mathcal{F}^{(d)}$. Since $VC(\mathcal{F}^{(d)}) = d + 1$, we have $M + 1 \leq d^* \leq d + 1$.

To improve the upper bound, we observe that S_M^{d-1} has $\binom{d}{M}$ different possible support, hence, we have $s(\mathcal{F}_M, n) \leq \binom{d}{M} S(\mathcal{F}^{(M)}, n)$. Then, by definition of VC dimension,

$$s(F_M, d^*) = 2^{d^*} \leq \binom{d}{M} s(\mathcal{F}^{(M)}, d^*) \overset{(a)}{\leq} \binom{d}{M} s(\mathcal{F}^{(M)}, d+1) \overset{(b)}{\leq} \binom{d}{M} \left(\frac{e(d+1)}{(M+1)}\right)^{(M+1)} \overset{(c)}{\leq} \left(\frac{ed}{M}\right)^{2M+1}.$$

where we use $d^* \leq d + 1$ in (a), Sauer's lemma in (b), and $\binom{d}{M} \leq \left(\frac{ed}{M}\right)^M$ and $(d+1)/(M+1) \leq d/M$ in (c). By observing that $e^{d^*/2} \leq 2^{d^*}$ and $4M + 2 \leq 6M$, we obtain the upper bound as well. ■

Corollary 22 *Let $n \geq d^*$. For any $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^d$, there exists $Q^x \subset S_M^{d-1} \times \mathbb{R}$ and $\pi : S_M^{d-1} \times \mathbb{R} \to Q^x$ with $|Q^x| \leq \left(\frac{en}{d^*}\right)^{d^*}$ such that for any $(\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}$, $\phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) = \phi'(\langle \pi((\boldsymbol{w}, b)), (\boldsymbol{x}_i, 1) \rangle)$ for $i = 1, \cdots, n$.*

Proof By Sauer's lemma, the image of $\Phi((\boldsymbol{w}, b)) := \big(\phi'(\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle + b), \cdots, \phi'(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b)\big)$, $(\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}$, has at most $(en/d^*)^{d^*}$ elements. We can define Q_x by mapping each $(\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}$ to a fixed $(\boldsymbol{w}', b')$ such that $\Phi((\boldsymbol{w}, b)) = \Phi((\boldsymbol{w}', b'))$. ■

D.2. Concentration for $\boldsymbol{Y}_\theta$

In this section, we derive a concentration bound for

$$\sup_{\substack{\mathcal{J} \subseteq [d] \\ |\mathcal{J}| = M'}} \sup_{\substack{\boldsymbol{w} \in S_M^{d-1} \\ b \in \mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2,$$

We will prove our bound in two steps. First, we will prove a bound for the truncated version $\boldsymbol{Y}_\theta$. In its following, we will extend that result by bounding the bias introduced by truncation.

Concentration of the truncated process: For some $R > 0$ and $\boldsymbol{v} \in S^{d-1}$ and $\theta = (\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}$, we let

$$\tilde{\boldsymbol{Y}}_{\theta,v} := \frac{1}{n} \sum_{i=1}^{n} y_i \mathbb{1}_{|y_i| \leq R} \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle \phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - \mathbb{E}\left[y \mathbb{1}_{|y| \leq R} \langle \boldsymbol{v}, \boldsymbol{x} \rangle \phi'(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\right].$$

Lemma 23 *For $\phi(t) \in \{t, ReLU(t)\}$, $n \geq d^*$ and $t \geq 0$, we have*

$$\mathbb{P}\left[\sup_{\theta \in S_M^{d-1} \times \mathbb{R}} \tilde{\boldsymbol{Y}}_{\theta,v} \geq 8R \max\{t, t^2\}\right] \leq \left(\frac{en}{d^*}\right)^{d^*} \exp\left(-nt^2\right).$$

Proof In the following, we will use that $|\phi'| \leq 1$ and $VC(\mathcal{F}_M) \leq d^*$, where d^* is defined in Proposition 21. We note that both hold for $\phi(t) \in \{t, \mathrm{ReLU}(t)\}$. Let

$$\ell\left((\boldsymbol{x}, \epsilon), (\boldsymbol{w}, b)\right) := y \mathbb{1}_{|y| \leq R} \left\langle \boldsymbol{v}, \boldsymbol{x} \right\rangle \phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b) \text{ and } \tilde{Z} := \sup_{\theta \in S_M^{d-1} \times \mathbb{R}} \tilde{\boldsymbol{Y}}_{\theta,v}.$$

By Lemma 20, for $\lambda > 0$, we have that

$$\mathbb{E}\left[\exp\left(\lambda \tilde{Z}\right)\right] \leq \mathbb{E}\left[\sup_{\substack{\boldsymbol{w} \in S_M^{d-1} \\ b \in \mathbb{R}}} \exp\left(\frac{2\lambda}{n} \sum_{i=1}^n \varepsilon_i \ell\left((\boldsymbol{x}_i, \epsilon_i), (\boldsymbol{w}, b)\right)\right)\right].$$

Let's focus on the empirical complexity. We have

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left[\sup_{\substack{\boldsymbol{w} \in S_M^{d-1} \\ b \in \mathbb{R}}} \exp\left(\frac{2\lambda}{n} \sum_{i=1}^n \varepsilon_i \ell\left((\boldsymbol{x}_i, \epsilon_i), (\boldsymbol{w}, b)\right)\right)\right] \\
&\overset{(a)}{=} \mathbb{E}_\varepsilon\left[\sup_{(\boldsymbol{w}, b) \in Q_x} \exp\left(\frac{2\lambda}{n} \sum_{i=1}^n \varepsilon_i \ell\left((\boldsymbol{x}_i, \epsilon_i), (\boldsymbol{w}, b)\right)\right)\right] \\
&\leq \sum_{(\boldsymbol{w}, b) \in Q_x} \mathbb{E}_\varepsilon\left[\exp\left(\frac{2\lambda}{n} \sum_{i=1}^n \varepsilon_i \ell\left((\boldsymbol{x}_i, \epsilon_i), (\boldsymbol{w}, b)\right)\right)\right] \\
&\overset{(b)}{=} \sum_{(\boldsymbol{w}, b) \in Q_x} \prod_{i=1}^n \mathbb{E}_\varepsilon\left[\exp\left(\frac{2\lambda}{n} \varepsilon_i \ell\left((\boldsymbol{x}_i, \epsilon_i), (\boldsymbol{w}, b)\right)\right)\right]
\end{aligned} \tag{D.5}$$

where (a) follows from Corollary 22 and (b) follows from the independence of ε_i. By using the moment generating function for Radamacher random variables, Lemma 73 and Corollary 22, we have for $\lambda \in \left[0, \frac{n}{4R}\right]$,

$$\begin{aligned}
\text{(D.5)} \leq \sum_{(\boldsymbol{w}, b) \in Q_x} \prod_{i=1}^n \exp\left(\frac{4\lambda^2}{n^2} \ell\left((\boldsymbol{x}_i, \epsilon_i), (\boldsymbol{w}, b)\right)^2\right) &\leq \sum_{(\boldsymbol{w}, b) \in Q_x} \prod_{i=1}^n \exp\left(\frac{8\lambda^2 R^2}{n^2}\right) \\
&\leq \left(\frac{en}{d^*}\right)^{d^*} \exp\left(\frac{8\lambda^2 R^2}{n}\right).
\end{aligned}$$

By Chernoff bound, the statement follows. ■

Concentration of $\boldsymbol{Y}_\theta$

Lemma 24 *Let $\phi(t) \in \{t, ReLU(t)\}$, $d \geq 4M$ and $M' \leq 2M$, and*

$$n \geq 24M \log^2\left(\frac{24dn}{M}\right) \text{ and } M \geq \log(2/\delta).$$

We have for $\delta \in (0,1]$,

$$\mathbb{P}\left[\sup_{\substack{\mathcal{J}\subseteq[d]\\|\mathcal{J}|=M'}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2 \geq K \log^{C_2}(6n/\delta)\sqrt{\frac{M\log^2\left(\frac{24dn}{M}\right)}{n}}\right] \leq \delta,$$

where K is a constant depending on (C_1, C_2, r, Δ).

Proof Let $\tilde{\boldsymbol{Y}}_\theta := \frac{1}{n}\sum_{i=1}^n y_i \mathbb{1}_{|y_i|\leq R}\boldsymbol{x}_i\phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) - \mathbb{E}\left[y\mathbb{1}_{|y|\leq R}\boldsymbol{x}\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\right]$, where $R = C_1(r+2)^{C_2}(e\log(6n/\delta))^{C_2} + \sqrt{\frac{\Delta}{e}}(e\log(6n/\delta))^{\frac{1}{2}}$. We observe that

$$\sup_{\substack{\mathcal{J}\subseteq[d]\\|\mathcal{J}|=M'}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2 \leq \underbrace{\sup_{\substack{\mathcal{J}\subseteq[d]\\|\mathcal{J}|=M'}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}} \left\|\tilde{\boldsymbol{Y}}_\theta|_{\mathcal{J}}\right\|_2}_{:=S_1} + \underbrace{\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}} \left\|\frac{1}{n}\sum_{i=1}^n y_i\mathbb{1}_{|y_i|>R}\boldsymbol{x}_i\phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b)\right\|_2}_{:=S_2}$$
$$+ \underbrace{\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}} \left\|\mathbb{E}\left[y\mathbb{1}_{|y|>R}\boldsymbol{x}\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\right]\right\|_2}_{:=S_3}.$$

For $K = \left(C_1^4(4C_2)^{4C_2}(r+2)^{4C_2} + 2\Delta^2\right)^{\frac{1}{4}}$, we have

$$\begin{aligned}
\mathbb{P}\left[\sup_{\substack{\mathcal{J}\subseteq[d]\\|\mathcal{J}|=M'}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2 \geq 16R\max\{t,t^2\} + 4K\sqrt{\tfrac{\delta}{6n}}\right] &\overset{(a)}{\leq} \mathbb{P}\left[S_1 \geq 16R\max\{t,t^2\}\right] \\
&\quad + \mathbb{P}\left[S_2 \geq \left(4 - 6^{\frac{3}{4}}\right)K\sqrt{\tfrac{\delta}{6n}}\right] \\
&\overset{(b)}{\leq} \mathbb{P}\left[S_1 \geq 16R\max\{t,t^2\}\right] + \tfrac{\delta}{2}
\end{aligned}$$

where (a) follows from Proposition 68 (since $4 > 6^{\frac{3}{4}}$), and (b) from Proposition 67.

Next, we need to establish a high probability bound via covering argument. Let $\mathcal{N}_{M'}^{1/2}$ be the minimal $1/2$-cover of $S_{M'}^{d-1}$. We have

$$S_1 = \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}} \sup_{\boldsymbol{v}\in S_{M'}^{d-1}} \left\langle \boldsymbol{v}, \tilde{\boldsymbol{Y}}_\theta\right\rangle \leq 2 \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}} \sup_{\boldsymbol{v}\in \mathcal{N}_{M'}^{1/2}} \left\langle \boldsymbol{v}, \tilde{\boldsymbol{Y}}_\theta\right\rangle, \tag{D.6}$$

where $\tilde{\boldsymbol{Y}}_{\theta,v}$ is introduced in Lemma 23. Therefore, by (D.6), we have

$$\mathbb{P}\left[S_1 \geq 16R\max\{t,t^2\}\right] \leq \sum_{\boldsymbol{v}\in\mathcal{N}_{M'}^{1/2}} \mathbb{P}\left[\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\ b\in\mathbb{R}}} \tilde{\boldsymbol{Y}}_{\theta,v} \geq 8R\max\{t,t^2\}\right] \overset{(a)}{\leq} \binom{d}{M'} 5^{M'} \left(\frac{en}{d^*}\right)^{d^*} e^{-nt^2}$$

where (a) follows from Corollary 71. Therefore, we have

$$\mathbb{P}\left[\sup_{\substack{\mathcal{J}\subseteq[d]\\ |\mathcal{J}|=M'}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\ b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2 \geq 16R\max\{t,t^2\} + 4K\sqrt{\frac{\delta}{6n}}\right] \leq \frac{\delta}{2} + \binom{d}{M'} 5^{M'} \left(\frac{en}{d^*}\right)^{d^*} e^{-nt^2}. \tag{D.7}$$

We note that

$$R \leq (C_1(r+2)^{C_2}e^{C_2} + \sqrt{\Delta e})\log^{C_2}(6n/\delta).$$

Moreover, for $d \geq 4M$ and $M' \leq 2M$, we have

$$\begin{aligned}\binom{d}{M'} 5^{M'} \left(\frac{en}{d^*}\right)^{d^*} \leq \binom{d}{2M} 5^{2M} \left(\frac{en}{M}\right)^{6M\log\left(\frac{ed}{M}\right)} &\overset{(a)}{\leq} \left(\frac{5ed}{2M}\right)^{2M} \left(\frac{en}{M}\right)^{6M\log\left(\frac{ed}{M}\right)} \\ &\leq \left(\frac{5e^2nd}{2M}\right)^{6M\log\left(\frac{ed}{M}\right)}\end{aligned}$$

where (a) follows from $\binom{d}{M} \leq \left(\frac{ed}{M}\right)^M$. Therefore,

$$\log\left[\binom{d}{M'} 5^{M'} \left(\frac{en}{d^*}\right)^{d^*}\right] \leq 6M\log\left(\frac{ed}{M}\right)\log\left(\frac{5e^2nd}{2M}\right) \leq 6M\log^2\left(\frac{24nd}{M}\right). \tag{D.8}$$

By using (D.8) and (D.7) with $t = \sqrt{\frac{6M\log^2\left(\frac{24nd}{M}\right)}{n}} + \sqrt{\frac{\log(2/\delta)}{n}} \in [0,1]$ and $u = e\log(6n/\delta)$, we obtain the statement. ■

D.3. Concentration for $\boldsymbol{\Sigma}_\theta$

In this part, we are interested in deriving a concentration bound for

$$\sup_{\substack{\mathcal{J}\subseteq[d]\\ |\mathcal{J}|=M'}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\ b\in\mathbb{R}}} \|\boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\|_2.$$

For a fixed $(\boldsymbol{w}, b) \in S_M^{d-1} \times \mathbb{R}$, by using the Rayleigh quotient formula, we can write that

$$\sup_{\substack{\mathcal{J}\subseteq[d]\\ |\mathcal{J}|=M'}} \|\boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\|_2 = \sup_{\substack{\mathcal{J}\subseteq[d]\\ |\mathcal{J}|=M'}} \sup_{\boldsymbol{v}\in S^{d-1}} |\langle \boldsymbol{v}, \boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{v}\rangle| = \sup_{\boldsymbol{v}\in S_{M'}^{d-1}} |\langle \boldsymbol{v}, \boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle|.$$

Let $\mathcal{N}_{M'}^{1/4}$ be the minimal $1/4$-cover of $S_{M'}^{d-1}$. It is easy to check that for $\boldsymbol{v} \in \mathcal{N}_{M'}^{1/4}$, we have

$$\sup_{\boldsymbol{v}\in S_{M'}^{d-1}} |\langle \boldsymbol{v}, \boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle| \leq 2 \sup_{\boldsymbol{v}\in\mathcal{N}_{M'}^{1/4}} |\langle \boldsymbol{v}, \boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle|.$$

Therefore, we have

$$\sup_{\substack{\mathcal{J}\subseteq[d]\\|\mathcal{J}|=M'}}\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}}\|\boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\|_2\leq\sup_{\boldsymbol{v}\in\mathcal{N}_{M'}^{1/4}}2\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}}|\langle\boldsymbol{v},\boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle|. \tag{D.9}$$

Since we already have a bound for the size of $\mathcal{N}_{M'}^{1/4}$, we first derive a concentration bound for $\sup_{\boldsymbol{w},b}|\langle\boldsymbol{v},\boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle|$ for a fixed $\boldsymbol{v}\in S_{M'}^{d-1}$.

Concentration for $\sup_{\boldsymbol{w},b}|\langle\boldsymbol{v},\boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle|$

Lemma 25 *For $\phi(t)\in\{t,ReLU(t)\}$, $M,\in[d]$, and for a fixed $\boldsymbol{v}\in S^{d-1}$ and $n\geq d^*$, we have that for $t\geq 0$,*

$$\mathbb{P}\left[\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}}|\langle\boldsymbol{v},\boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle|\geq 8\sqrt{2}\max\{t,t^2\}\right]\leq 2\left(\frac{en}{d^*}\right)^{d^*}\exp\left(-nt^2\right).$$

Proof We observe that

$$\langle\boldsymbol{v},\boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle=\frac{1}{n}\sum_{i=1}^n\phi'(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}_i\right\rangle^2-\mathbb{E}\left[\phi'(\langle\boldsymbol{w},\boldsymbol{x}\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}\right\rangle^2\right].$$

For

$$Z\coloneqq\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}}\frac{1}{n}\sum_{i=1}^n\phi'(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}_i\right\rangle^2-\mathbb{E}\left[\phi'(\langle\boldsymbol{w},\boldsymbol{x}\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}\right\rangle^2\right]$$

by using Lemma 20, we can write that for $\lambda\geq 0$,

$$\mathbb{E}\left[\exp(\lambda Z)\right]\leq\mathbb{E}\left[\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}}\exp\left(\frac{2\lambda}{n}\sum_{i=1}^n\varepsilon_i\phi'(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}_i\right\rangle^2\right)\right].$$

Let's look at the empirical complexity. We have

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left[\sup_{\substack{\boldsymbol{w}\in S_M^{d-1}\\b\in\mathbb{R}}}\exp\left(\frac{2\lambda}{n}\sum_{i=1}^n\varepsilon_i\phi'(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}_i\right\rangle^2\right)\right]\\
&\qquad\overset{(a)}{=}\mathbb{E}_\varepsilon\left[\sup_{(\boldsymbol{w},b)\in Q_x}\exp\left(\frac{2\lambda}{n}\sum_{i=1}^n\varepsilon_i\phi'(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}_i\right\rangle^2\right)\right]\\
&\qquad\leq\sum_{(\boldsymbol{w},b)\in Q_x}\mathbb{E}_\varepsilon\left[\exp\left(\frac{2\lambda}{n}\sum_{i=1}^n\varepsilon_i\phi'(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}_i\right\rangle^2\right)\right]\\
&\qquad\overset{(b)}{=}\sum_{(\boldsymbol{w},b)\in Q_x}\prod_{i=1}^n\mathbb{E}_\varepsilon\left[\exp\left(\frac{2\lambda}{n}\varepsilon_i\phi'(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)\left\langle\boldsymbol{v},\boldsymbol{x}_i\right\rangle^2\right)\right],
\end{aligned}\tag{D.10}$$

where (a) follows from Corollary 22 and (b) follows by independence. Let $\cosh(t) := \frac{e^t+e^{-t}}{2}$. We observe that for a fixed $i \in [n]$,

$$\mathbb{E}_\varepsilon\Big[\exp\left(\tfrac{2\lambda}{n}\varepsilon_i\phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b)\left\langle \boldsymbol{v}, \boldsymbol{x}_i\right\rangle^2\right)\Big] = \cosh\left(\tfrac{2\lambda}{n}|\phi'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b)|\left\langle \boldsymbol{v}, \boldsymbol{x}_i\right\rangle^2\right) \leq \cosh\left(\tfrac{2\lambda}{n}\left\langle \boldsymbol{v}, \boldsymbol{x}_i\right\rangle^2\right) \tag{D.11}$$

where we use $|\phi'| \leq 1$ and that cosh is increasing on $t \geq 0$. Therefore by (D.10) and (D.11), for $\lambda \in \left[0, n/4\sqrt{2}\right]$

$$\mathbb{E}\left[\exp(\lambda Z)\right] \leq \sum_{(\boldsymbol{w},b)\in Q_x}\prod_{i=1}^{n}\mathbb{E}\left[\cosh\left(\frac{2\lambda}{n}\left\langle \boldsymbol{v}, \boldsymbol{x}_i\right\rangle^2\right)\right] \leq \left(\frac{en}{d^*}\right)^{d^*}\exp\left(\frac{16\lambda^2}{n}\right),$$

where we used Lemma 73 and Corollary 22. By Chernoff's bound, the statement follows. ■

Concentration for $\boldsymbol{\Sigma}_\theta$

The next statement provides a concentration bound for (D.9).

Lemma 26 *For $\phi(t) \in \{t, ReLU(t)\}$, $M, M' \in [d]$, and for $d \geq 4M$ and $M' \leq 2M$,*

$$n \geq 24M\log^2\left(\frac{35dn}{M}\right) \quad \textit{and} \quad M \geq \log(2/\delta),$$

we have for $\delta \in (0,1]$,

$$\mathbb{P}\left[\sup_{\substack{\mathcal{J}\subseteq[d]\\|\mathcal{J}|=M'}}\ \sup_{\substack{\boldsymbol{w}\in S^{d-1}_M\\ b\in\mathbb{R}}}\|\boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\|_2 \geq K\sqrt{\frac{M\log^2\left(\frac{35dn}{M}\right)}{n}}\right] \leq \delta,$$

where K is a universal positive constant.

Proof By using (D.9) and Lemma 25, we can write that for $n \geq d^*$

$$\mathbb{P}\left[\sup_{\substack{\mathcal{J}\subseteq[d]\\|\mathcal{J}|=M'}}\ \sup_{\substack{\boldsymbol{w}\in S^{d-1}_M\\ b\in\mathbb{R}}}\|\boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\|_2 \geq 16\sqrt{2}\max\{t,t^2\}\right] \leq \sum_{\boldsymbol{v}\in\mathcal{N}^{1/4}_{M'}}\mathbb{P}\left[\sup_{\substack{\boldsymbol{w}\in S^{d-1}_M\\ b\in\mathbb{R}}}|\langle \boldsymbol{v}, \boldsymbol{\Sigma}_\theta\boldsymbol{v}\rangle| \geq 8\sqrt{2}\max\{t,t^2\}\right] \overset{(a)}{\leq} 2\binom{d}{M'}9^{M'}\left(\frac{en}{d^*}\right)^{d^*}\exp\left(-nt^2\right). \tag{D.12}$$

where (a) follows from Corollary 71. We note that $n \geq 24M\log^2\left(\frac{35dn}{M}\right) \geq 6M\log\left(\frac{ed}{M}\right) \geq d^*$ by Proposition 21. Moreover, for $d \geq 4M$ and $M' \leq 2M$, we have

$$\binom{d}{M'}9^{M'}\left(\frac{en}{d^*}\right)^{d^*} \leq \binom{d}{2M}9^{2M}(en)^{6M\log\left(\frac{ed}{M}\right)} \leq \left(\frac{9ed}{2M}\right)^{2M}(en)^{6M\log\left(\frac{ed}{M}\right)} \leq \left(\frac{9e^2nd}{2M}\right)^{6M\log\left(\frac{ed}{M}\right)}$$

where the second inequality follows from $\binom{d}{M} \leq \left(\frac{ed}{M}\right)^M$. Therefore,

$$\log\left[\binom{d}{M'} 9^{M'} \left(\frac{6n}{d^*}\right)^{d^*}\right] \leq 6M \log\left(\frac{ed}{M}\right) \log\left(\frac{9e^2nd}{2M}\right) \leq 6M \log^2\left(\frac{35nd}{M}\right). \quad \text{(D.13)}$$

By using (D.13) and (D.12) with $t = \sqrt{\frac{6M\log^2\left(\frac{35nd}{M}\right)}{n}} + \sqrt{\frac{\log(2/\delta)}{n}} \in [0,1]$, we obtain the statement. ■

D.4. Concentration for T_θ

By (D.4) and $\|\mathbb{E}[\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\boldsymbol{x}\boldsymbol{x}^\top]\|_2 \leq 1$, we have

$$\|\boldsymbol{T}_\theta|_{\mathcal{J}}\|_2 \leq \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2 + \|\boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\|_2 \|\mathbb{E}[y\boldsymbol{x}]\|_2 + \left(\|\boldsymbol{\Sigma}_\theta|_{\mathcal{J}\times\mathcal{J}}\|_2 + 1\right)\|(\hat{\boldsymbol{\mu}} - \mathbb{E}[y\boldsymbol{x}])|_{\mathcal{J}}\|_2.$$

We have the following statement.

Lemma 27 *For $\phi(t) \in \{t, ReLU(t)\}$, $M, M' \in [d]$, for $d \geq 4M$, $M' \leq 2M$*

$$n \geq 24M\log^2\left(\frac{35dn}{M}\right) \quad \textit{and} \quad M \geq \log(6/\delta),$$

we have that for $\delta \in (0,1]$

$$\mathbb{P}\left[\sup_{\substack{\mathcal{J}\subseteq[d] \\ |\mathcal{J}|=M'}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{T}_\theta|_{\mathcal{J}}\|_2 \geq K\log^{C_2}(18n/\delta)\sqrt{\frac{M\log^2\left(\frac{35dn}{M}\right)}{n}}\right] \leq \delta,$$

where K is a positive constant depending on (C_1, C_2, r, Δ).

Proof We note that Lemma 24 applies to $\phi(t) \in \{t, \mathrm{ReLU}(t)\}$. Therefore, by Lemma 24 for $\phi(t) = |t|$, and $\phi(t) = \mathrm{ReLU}(t)$, and Lemma 26, we have the statement. ■

D.5. Concentration Bound for the Empirical Gradient in the Single-Index Setting

In this part, since $r = 1$, for clarity, we use the following notation: $\sigma^* = \sum_{k\geq k^\star} \frac{\gamma_k}{k!} H_{e_k}$ and $y = \sigma^*(\langle \boldsymbol{v}, \boldsymbol{x}\rangle) + \sqrt{\Delta}\epsilon$.

Proposition 28 *We consider (D.1) with $\hat{\boldsymbol{\mu}} = 0$ and $\phi(t) \in \{t, ReLU(t)\}$. Let $j \in [2m]$ be a fixed index and $\mathcal{J}$ be any function of $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ such that $|\mathcal{J}| \leq M$ almost surely. For $d \geq 4M$,*

$$n \geq 24M\log^2\left(\frac{24dn}{M}\right) \quad \textit{and} \quad M \geq 24(1 + \log(4/\delta)),$$

the intersection of the following events holds with at least probability $1-\delta$,

$$\left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} - \frac{\gamma_{k^\star}\tilde{\gamma}_{k^\star}(b)}{(k^\star-1)!}\left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)}\right\rangle^{k^\star-1} \boldsymbol{v}|_{\mathcal{J}}\right\|_2 \leq K\left(\sqrt{\frac{M\log^2\left(\frac{24dn}{M}\right)\log^{2C_2}\left(\frac{12n}{\delta}\right)}{n}} + \left(\frac{1+\log(4/\delta)}{M}\right)^{\frac{k^\star}{2}}\right)$$

$$\left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} \right\|_2 \leq K \left(\frac{|\gamma_{k^\star} \tilde{\gamma}_{k^\star}(b)|}{(k^\star-1)!} \left(\frac{1+\log(4/\delta)}{M}\right)^{\frac{k^\star-1}{2}} + \sqrt{\frac{M \log^2\left(\frac{24dn}{M}\right) \log^{2C_2}\left(\frac{12n}{\delta}\right)}{n}} \right).$$

where $K > 0$ is a constant depending on $(C_1, C_2, k^\star, \Delta, C_{\sigma^*})$.

Proof We first observe that by Proposition 13,

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{x},y)}&\left[y\boldsymbol{x}\phi'\Big(\left\langle \boldsymbol{W}_{j*}^{(0)}, x\right\rangle + b\Big)\right]\Big|_{\mathcal{J}} \\ &= \boldsymbol{v}|_{\mathcal{J}} \sum_{k \geq k^\star - 1} \frac{\gamma_{k+1}\tilde{\gamma}_{k+1}(b)}{k!} \left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)} \right\rangle^k + \boldsymbol{W}_{j*}^{(0)} \sum_{k \geq k^\star} \frac{\gamma_k \tilde{\gamma}_{k+2}(b)}{k!} \left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)} \right\rangle^k.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\left\| g\left(\boldsymbol{W}_{j^\star}^{(0)}, b\right)\Big|_{\mathcal{J}} - \frac{\gamma_{k^\star}\tilde{\gamma}_{k^\star}(b)}{(k^\star-1)!} \left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)} \right\rangle^{k^\star-1} \boldsymbol{v}|_{\mathcal{J}} \right\|_2 \\ &\leq \left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} - \mathbb{E}_{(\boldsymbol{x},y)}\left[y\boldsymbol{x}\phi'\left(\left\langle \boldsymbol{W}_{j*}^{(0)}, x\right\rangle + b\right)\right]\Big|_{\mathcal{J}} \right\|_2 + \|\zeta_{k^\star-1}\| \\ &\leq \sup_{\substack{\mathcal{J}\subseteq[d] \\ |\mathcal{J}|=M}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2 + (1+\sqrt{k^\star+1})C_{\sigma^*} \frac{\left|\left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)}\right\rangle\right|^{k^\star}}{1-\left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)}\right\rangle^2}\end{aligned}$$

where $\zeta_{k^\star-1}$ is the higher order terms in the Hermite expansion defined in Proposition 17 and we use Proposition 17 in the third line line.

To bound the second term, we recall that $\boldsymbol{W}_{j*}^{(0)} = \frac{\boldsymbol{W}_{j*}|_{\mathcal{J}}}{\|\boldsymbol{W}_{j*}|_{\mathcal{J}}\|_2}$ where $\boldsymbol{W}_{j*} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and it is independent of $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Since $\mathcal{J}$ is independent of $\boldsymbol{W}_{j*}$, without loss of generality, we can fix a $\mathcal{J}$ with $|\mathcal{J}| = M$. By using Corollaries 57 and 58, the intersection of (i) $\sum_{i\in\mathcal{J}} \boldsymbol{W}_{ij}^2 \geq \frac{M}{2}$, (ii) $\langle \boldsymbol{v}, \boldsymbol{W}_{j*}|_{\mathcal{J}}\rangle^2 \leq 3(1+\log(4/\delta))$ holds with probability at least $1-\delta/2$. Within that event, for $M \geq 24(1+\log(4/\delta))$, we have

$$(1+\sqrt{k^\star+1})C_{\sigma^*} \frac{\left|\left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)}\right\rangle\right|^{k^\star}}{1-\left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)}\right\rangle^2} \leq 6^{\frac{k^\star+1}{2}} C_{\sigma^*}(1+\sqrt{k^\star+1}) \left(\frac{(1+\log(4/\delta))}{M}\right)^{\frac{k^\star}{2}} \tag{D.14}$$

Then, by Lemma 24, the first item in the statement follows. For the second item, by using the event used for (D.14), we have

$$\begin{aligned}\left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} \right\|_2 \leq& \left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} - \frac{\gamma_{k^\star}\tilde{\gamma}_{k^\star}(b)}{(k^\star-1)!} \left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)}\right\rangle^{k^\star-1} \boldsymbol{v}|_{\mathcal{J}} \right\|_2 \\ &+ \frac{|\gamma_{k^\star}\tilde{\gamma}_{k^\star}(b)|}{(k^\star-1)!} \left(\frac{6(1+\log(4/\delta))}{M}\right)^{\frac{k^\star-1}{2}}.\end{aligned}$$

By using the first item in the statement, the second item also follows. ■

D.6. Concentration Bound for the Empirical Gradient in the Multi-Index Setting

We first derive the Hermite expansion of $\mathbb{E}_{(\boldsymbol{x},y)}\left[\overline{y}\boldsymbol{x}\phi'(\langle \boldsymbol{w},\boldsymbol{x}\rangle + b)\right]$ (see (D.2) for its definition).

Lemma 29 *We recall that* $\phi(\cdot + b) := \sum_{k\geq 0} \frac{\tilde{\gamma}_k(b)}{k!} H_{e_k}$ *and* $\sigma^* := \sum_{k\geq 0} \frac{1}{k!}\langle \boldsymbol{T_k}, \boldsymbol{H_{e_k}}\rangle$*. For any* $\mathcal{J} \subseteq [d]$ *and any* $\boldsymbol{w} \in S^{d-1}$ *supported on* $\mathcal{J}$*, we have*

$$\mathbb{E}_{(\boldsymbol{x},y)}\left[\overline{y}\boldsymbol{x}\phi'\left(\langle \boldsymbol{w},\boldsymbol{x}\rangle + b\right)\right]\Big|_{\mathcal{J}} = \tilde{\gamma}_2(b)\boldsymbol{H}|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{w} + \boldsymbol{V}|_{\mathcal{J}}\sum_{k\geq 2}\tfrac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right]$$
$$+ \boldsymbol{w}\sum_{k\geq 2}\tfrac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right],$$

where $\boldsymbol{H}$ *is defined in* (DEF-H).

Proof We first observe that $\mathbb{E}[y\boldsymbol{x}] = \mathbb{E}[\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})\boldsymbol{x}] = \boldsymbol{V}\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}]$ and $\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}} = \boldsymbol{V}|_{\mathcal{J}}\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}]$. By Proposition 13, we have

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{x},y)}\left[\overline{y}\boldsymbol{x}\phi'\left(\langle \boldsymbol{w},\boldsymbol{x}\rangle + b\right)\right] &\overset{(a)}{=} \tilde{\gamma}_1(b)\boldsymbol{V}|_{\mathcal{J}^c}\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}] + \tilde{\gamma}_3(b)\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{V}|_{\mathcal{J}^c}\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}] + \tilde{\gamma}_2(b)\boldsymbol{H}\boldsymbol{w}\\ &+ \boldsymbol{V}\sum_{k\geq 2}\frac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right] + \boldsymbol{w}\sum_{k\geq 2}\frac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right]\end{aligned} \tag{D.15}$$

where (a) holds since $\boldsymbol{\nabla T_1} = \mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}]$, $\boldsymbol{V}\boldsymbol{\nabla T_2}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes 1}\right] = \boldsymbol{H}\boldsymbol{w}$, $\boldsymbol{T_0} = 0$, $\boldsymbol{T_1}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes 1}\right] = \langle \boldsymbol{w}, \boldsymbol{V}\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}]\rangle$. Since $\boldsymbol{w}^\top\boldsymbol{V}|_{\mathcal{J}^c}\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}] = 0$, we have

$$\begin{aligned}\text{(D.15)} &= \tilde{\gamma}_2(b)\boldsymbol{H}\boldsymbol{w} + \tilde{\gamma}_1(b)\boldsymbol{V}|_{\mathcal{J}^c}\mathbb{E}[\sigma^*(\boldsymbol{z})\boldsymbol{z}]\\ &+ \boldsymbol{V}\sum_{k\geq 2}\tfrac{\tilde{\gamma}_{k+1}(b)}{k!}\boldsymbol{\nabla T_{k+1}}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right] + \boldsymbol{w}\sum_{k\geq 2}\tfrac{\tilde{\gamma}_{k+2}(b)}{k!}\boldsymbol{T_k}\left[(\boldsymbol{V}^\top\boldsymbol{w})^{\otimes k}\right].\end{aligned}$$

Since $\boldsymbol{w}$ is supported on $\mathcal{J}$, the statement follows. ■

Proposition 30 *We consider* (D.1) *with* $\hat{\boldsymbol{\mu}} = \sum_{i=1}^n y_i\boldsymbol{x}_i$ *and* $\phi(t) \in \{t, ReLU(t)\}$*. Let* $j \in [m]$ *be a fixed index and* $\mathcal{J}$ *be any function of* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ *such that* $|\mathcal{J}| \leq M$ *almost surely. For* $d \geq 4M$*,*

$$n \geq 24M\log^2\left(\frac{35dn}{M}\right) \quad \text{and} \quad M \geq 24(r + \log(12/\delta)),$$

the intersection the following events hold with at least probability $1-\delta$*,*

$$\left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} - \tilde{\gamma}_2(b)\boldsymbol{H}|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)}\right\|_2 \leq K\left(\sqrt{\frac{M\log^2\left(\frac{35dn}{M}\right)\log^{2C_2}\left(\frac{18n}{\delta}\right)}{n}} + \frac{(r+\log(4/\delta))}{M}\right)$$

$$\left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}}\right\| \leq K\left(|\tilde{\gamma}_2(b)|\sqrt{\frac{r+\log(4/\delta)}{M}} + \sqrt{\frac{M\log^2\left(\frac{35dn}{M}\right)\log^{2C_2}\left(\frac{18n}{\delta}\right)}{n}}\right).$$

where $K > 0$ is a constant depending on $(C_1, C_2, r, \Delta, C_{\sigma^*})$.

Proof We have that

$$\begin{aligned}\left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} - \tilde{\gamma}_2(b)\boldsymbol{H}|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)}\right\|_2 &\leq \left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} - \mathbb{E}\left[\overline{y}\boldsymbol{x}\phi'\left(\left\langle \boldsymbol{W}_{j*}^{(0)}, \boldsymbol{x}\right\rangle + b\right)\right]\Big|_{\mathcal{J}}\right\|_2 \\ &+ \left\| \mathbb{E}\left[\overline{y}\boldsymbol{x}\phi'\left(\left\langle \boldsymbol{W}_{j*}^{(0)}, \boldsymbol{x}\right\rangle + b\right)\right]\Big|_{\mathcal{J}} - \tilde{\gamma}_2(b)\boldsymbol{H}|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)}\right\|_2 \\ &\overset{(a)}{\leq} \sup_{\substack{\mathcal{J}\subseteq[d] \\ |\mathcal{J}|=M}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{T}_\theta|_{\mathcal{J}}\|_2 + \|\zeta_1|_{\mathcal{J}}\|_2 \\ &\overset{(b)}{\leq} \sup_{\substack{\mathcal{J}\subseteq[d] \\ |\mathcal{J}|=M}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{T}_\theta|_{\mathcal{J}}\|_2 + 2\sqrt{3}C_{\sigma^*}\frac{\|\boldsymbol{V}^\top\boldsymbol{W}_{j*}^{(0)}\|_2^2}{1 - \|\boldsymbol{V}^\top\boldsymbol{W}_{j*}^{(0)}\|_2^2}.\end{aligned}$$

where we used Lemma 29 in (a) and Proposition 17 in (b).

We will first bound the second term. We recall that $\boldsymbol{W}_{j*}^{(0)} = \frac{\boldsymbol{W}_{j*}|_{\mathcal{J}}}{\|\boldsymbol{W}_{j*}|_{\mathcal{J}}\|_2}$ where $\boldsymbol{W}_{j*} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and it is independent of $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Since $\mathcal{J}$ is independent of $\boldsymbol{W}_{j*}$, without loss of generality, we can fix a $\mathcal{J}$ with $|\mathcal{J}| = M$. By using Corollaries 57 and 58, the intersection of (i) $\sum_{i\in\mathcal{J}} \boldsymbol{W}_{ij}^2 \geq \frac{M}{2}$, (ii) $\|\boldsymbol{V}^\top\boldsymbol{W}_{j*}|_{\mathcal{J}}\|_2^2 \leq 3(r + \log(4/\delta))$ holds with probability at least $1 - \delta/2$. Within that event, for $M \geq 24(r + \log(12/\delta))$, we have

$$2\sqrt{3}C_{\sigma^*}\frac{\|\boldsymbol{V}^\top\boldsymbol{W}_{j*}^{(0)}\|_2^2}{1 - \|\boldsymbol{V}^\top\boldsymbol{W}_{j*}^{(0)}\|_2^2} \leq 16\sqrt{3}C_{\sigma^*}\frac{(r + \log(4/\delta))}{M}. \tag{D.16}$$

Therefore, by Lemma 27, the first item follows. For the second item, we observe that

$$\left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}}\right\|_2 \leq \tilde{\gamma}_2(b)\left\|\boldsymbol{H}|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)}\right\|_2 + \left\| g\left(\boldsymbol{W}_{j*}^{(0)}, b\right)\Big|_{\mathcal{J}} - \tilde{\gamma}_2(b)\boldsymbol{H}|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)}\right\|_2$$

We have that

$$\|\boldsymbol{H}|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)}\|_2 \leq \tilde{\gamma}_2(b)\frac{\left\|\boldsymbol{V}^\top\boldsymbol{W}_{j*}^{(0)}\right\|_2}{\left\|\boldsymbol{W}_{j*}^{(0)}|_{\mathcal{J}}\right\|_2} \leq \tilde{\gamma}_2(b)\sqrt{\frac{6(r + \log(4/\delta))}{M}}.$$

where we used $\sigma_1(\boldsymbol{H}) \leq 1$ in the first step, and the event used for (D.16). Hence by the first part of the statement, the second item also follows. ■

Appendix E. Guarantee for `PruneNetwork`

We recall the following notation: For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{2m}$ and $\boldsymbol{W} \in \mathbb{R}^{2m\times d}$,

$$\begin{aligned}R_n^{\pm}(\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b}) &:= \frac{1}{2n}\sum_{i=1}^n \left(y_i - \hat{y}^{\pm}\left(\boldsymbol{x}_i; (\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b})\right)\right) \\ \hat{y}^{\pm}\left(\boldsymbol{x}; (\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b})\right) &:= \sum_{j=1}^{2m} a_j \underbrace{\left(\tfrac{\phi(\langle\tilde{\boldsymbol{e}}_l,\boldsymbol{x}\rangle + \boldsymbol{b}_j) \pm \phi(-\langle\tilde{\boldsymbol{e}}_l,\boldsymbol{x}\rangle + b_j)}{2}\right)}_{\phi_{\pm}(\langle\tilde{\boldsymbol{e}}_l,\boldsymbol{x}\rangle; b_j)}\end{aligned}$$

and the gradients of the empirical/population risks are

$$\nabla_j R_n^{\pm}(\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b}) = \frac{-a_j}{n} \sum_{i=1}^{n} \big(y_i - \hat{y}^{\pm}\left(\boldsymbol{x}_i; (\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b})\right)\big) \phi'_{\pm}(\langle \tilde{\boldsymbol{e}}_l, \boldsymbol{x}_i \rangle ; b_j) \boldsymbol{x}_i$$
$$\nabla_j R^{\pm}(\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b}) = -a_j \mathbb{E}_{(\boldsymbol{x},y)} \left[\big(y - \hat{y}^{\pm}\left(\boldsymbol{x}; (\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b})\right)\big) \phi'_{\pm}\left(\langle \tilde{\boldsymbol{e}}_l, \boldsymbol{x} \rangle ; b_j\right) \boldsymbol{x}\right].$$

Finally, we recall that

$$\|\nabla R_n(\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b})\|_F^2 = \sum_{j=1}^{m} \|\nabla_j R_n(\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b})\|_2^2 \text{ and } \|\nabla R_n^{\pm}(\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b})\|_F^2 = \sum_{j=1}^{m} \|\nabla_j R_n^{\pm}(\boldsymbol{a}, \tilde{\boldsymbol{e}}_l, \boldsymbol{b})\|_2^2.$$

E.1. Auxiliary Results

We have the following statement:

Proposition 31 *Let* $\overline{\gamma}_k^2 \coloneqq \frac{1}{m} \sum_{j=1}^{m} \tilde{\gamma}_k^2(b_j^{(0)})$*. For any* $\mathcal{J} \subseteq [d]$*, we have*

1. *For the single-index setting and* $k^\star > 1$*,*

$$\left[\left(\frac{\overline{\gamma}_{k^\star} |\gamma_{k^\star}|}{(k^\star - 1)!}\right)^{\frac{2}{k^\star - 1}} - 8 \left(\frac{c\sqrt{2} C_{\sigma^*}}{1 - c^2}\right)^{\frac{2}{k^\star - 1}}\right] \|\boldsymbol{v}|_{\mathcal{J}^c}\|_2^2 \leq \frac{m^{\frac{-1}{k^\star - 1}}}{c^2} \sum_{i \in \mathcal{J}^c} \|\nabla R^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\|_F^{\frac{2}{k^\star - 1}}.$$

where the statement with ∇R^+ *holds for even* $k^\star$*, and* ∇R^- *holds for odd* $k^\star$*.*

2. *For the multi-index setting, we have*

$$\left[\overline{\gamma}_2^2 \sigma_r^2(\boldsymbol{H}) - 16 \left(\frac{c C_{\sigma^*}}{1 - c^2}\right)^2\right] \|\boldsymbol{V}|_{\mathcal{J}^c}\|_F^2 \leq \frac{m^{-1}}{c^2} \sum_{i \in \mathcal{J}^c} \|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\|_F^2.$$

Proof We first observe that by (2.1), we have $\hat{y}^{\pm}\big(\boldsymbol{x}; (\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\big) = 0$. Therefore,

$$\nabla_j R^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) = -a_j^{(0)} \mathbb{E}_{(\boldsymbol{x},y)} \left[\sigma^*(\boldsymbol{V}^\top \boldsymbol{x}) \phi'_{\pm}(\langle \tilde{\boldsymbol{e}}_i, \boldsymbol{x} \rangle ; b_j^{(0)}) \boldsymbol{x}\right]. \tag{E.1}$$

Moreover, we observe that by (2.1), $\overline{\gamma}_k^2 = \frac{1}{m} \sum_{j=1}^{m} \tilde{\gamma}_k^2(b_j^{(0)})$.

1. We will prove this item only for even $k^\star > 1$. The proof for the odd case is identical when $(+)$ signs are replaced with $(-)$. We have

$$\begin{aligned}
&\Big(\frac{\tilde{\gamma}_{k^\star}(b_j^{(0)}) \gamma_{k^\star}}{(k^\star - 1)!} (c \boldsymbol{v}_i)^{k^\star - 1}\Big)^2 \\
&\qquad \overset{(a)}{\leq} 2 \left\| \mathbb{E}_{(\boldsymbol{x},y)} \left[\sigma^*(\langle \boldsymbol{v}, \boldsymbol{x} \rangle) \phi'_+(\langle \tilde{\boldsymbol{e}}_l, \boldsymbol{x} \rangle ; b_j^{(0)}) \boldsymbol{x}\right] - \frac{\tilde{\gamma}_{k^\star}(b_j^{(0)}) \gamma_{k^\star} \langle \boldsymbol{v}, \tilde{\boldsymbol{e}}_i \rangle^{k^\star - 1}}{(k^\star - 1)!} \boldsymbol{v} \right\|_2^2 \\
&\qquad + 2 \left\| \nabla_j R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) \right\|_2^2 \\
&\qquad \overset{(b)}{\leq} 2(1 + \sqrt{k^\star + 1})^2 C_{\sigma^*}^2 \frac{c^{2k^\star} |\boldsymbol{v}_i|^{2k^\star}}{(1 - c^2)^2} + 2 \left\| \nabla_j R^-(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) \right\|_2^2
\end{aligned}$$

where (a) follows from (E.1), (b) follows from Corollary 18. By summing each side over $j \in [2m]$ and dividing by $1/2m$, we get

$$\left(\frac{\overline{\gamma}_{k^\star}\gamma_{k^\star}}{(k^\star-1)!}c^{k^\star-1}\boldsymbol{v}_i^{k^\star-1}\right)^2 \le 2(1+\sqrt{k^\star+1})^2 C_{\sigma^*}^2 \frac{c^{2k^\star}|\boldsymbol{v}_i|^{2k^\star}}{(1-c^2)^2} + \frac{2}{2m}\left\|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_F^2.$$

By taking $\frac{1}{(k^\star-1)}$th power of each sides, we get

$$\begin{aligned}\left(\frac{\overline{\gamma}_{k^\star}|\gamma_{k^\star}|}{(k^\star-1)!}\right)^{\frac{2}{k^\star-1}} c^2\boldsymbol{v}_i^2 &\overset{(a)}{\le} 2^{\frac{1}{k^\star-1}}(1+\sqrt{k^\star+1})^{\frac{2}{k^\star-1}}\left(C_{\sigma^*}\frac{c|\boldsymbol{v}_i|}{1-c^2}\right)^{\frac{2}{k^\star-1}} c^2\boldsymbol{v}_i^2 \\ &\quad + m^{\frac{-1}{k^\star-1}}\left\|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_F^{\frac{2}{k^\star-1}} \\ &\overset{(b)}{\le} 2^{\frac{1}{k^\star-1}} 8\left(C_{\sigma^*}\frac{c}{1-c^2}\right)^{\frac{2}{k^\star-1}} c^2\boldsymbol{v}_i^2 \\ &\quad + m^{\frac{-1}{k^\star-1}}\left\|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_F^{\frac{2}{k^\star-1}}.\end{aligned}$$

where (a) follows from Proposition 72 and (b) holds since $|\boldsymbol{v}_i| \le 1$ and $(1+\sqrt{k^\star+1})^{\frac{2}{k^\star-1}}$ is decreasing for $k^\star \ge 2$. Then, we get

$$\left[\left(\frac{\overline{\gamma}_{k^\star}|\gamma_{k^\star}|}{(k^\star-1)!}\right)^{\frac{2}{k^\star-1}} - 2^{\frac{1}{k^\star-1}}8\left(C_{\sigma^*}\frac{c}{1-c^2}\right)^{\frac{2}{k^\star-1}}\right]\boldsymbol{v}_i^2 \le \frac{m^{\frac{-1}{k^\star-1}}}{c^2}\left\|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_F^{\frac{2}{k^\star-1}}.$$

By summing each sides over $i \in \mathcal{J}^c$, we have the statement.

2. By observing that $c\boldsymbol{H}_{i*} = \boldsymbol{H}\tilde{\boldsymbol{e}}_i$, we have

$$\begin{aligned}&\|\tilde{\gamma}_2(b_j^{(0)})c\boldsymbol{H}_{i*}\|_2^2 \\ &\overset{(a)}{\le} 2\left\|\mathbb{E}_{(\boldsymbol{x},y)}\left[\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})\phi'_+(\langle\tilde{\boldsymbol{e}}_i,\boldsymbol{x}\rangle\,; b_j^{(0)})\boldsymbol{x}\right] - \tilde{\gamma}_2(b_j^{(0)})\boldsymbol{H}\tilde{\boldsymbol{e}}_i\right\|_2^2 + 2\left\|\nabla_j R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_2^2 \\ &\overset{(b)}{\le} 16C_{\sigma^*}^2\left(\frac{c}{1-c^2}\right)^2 c^2\|\boldsymbol{V}_{i*}\|_2^2 + 2\left\|\nabla_j R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_2^2.\end{aligned}$$

where (a) follows from (E.1), and (b) holds since Corollary 18 and $\|\boldsymbol{V}_{i*}\|_2 \le 1$. By summing each side over $j \in [2m]$ and dividing by $1/2m$, we get

$$\overline{\gamma}_2^2 c^2\|\boldsymbol{H}_{i*}\|_2^2 \le 16C_{\sigma^*}^2\left(\frac{c}{1-c^2}\right)^2 c^2\|\boldsymbol{V}_{i*}\|_2^2 + 2(2m)^{-1}\left\|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_F^2.$$

Therefore, we have

$$\left[\overline{\gamma}_2^2\sigma_r^2(\boldsymbol{H}) - 16C_{\sigma^*}^2\left(\frac{c}{1-c^2}\right)^2\right]\|\boldsymbol{V}_{i*}\|_2^2 \le \frac{m^{-1}}{c^2}\left\|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_F^2.$$

By summing each sides over $i \in \mathcal{J}^c$, we have the statement.

■

Proposition 32 *For this statement, by abusing the notation, we use $0^0 = 1$. Let*

$$\tilde{R}_i^{\pm} := \frac{1}{2m}\sum_{j=1}^{2m} \|\widetilde{\nabla}_j R_n^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) - \nabla_j R^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\|_2^2,$$

where $\widetilde{\nabla}_j R_n^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) := \nabla_j R_n^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})|_{top(M)}$*,*

$$\tilde{M} := M \log^2\left(\frac{35nd}{M}\right) \;\; \textit{and} \;\; C_q := 8q(2-q)^{\frac{2-q}{q}}.$$

For $d \geq 4M$, $n \geq 24\tilde{M}$ and $M \geq \log(2/\delta)$, each of the following items holds with probability at least $1-\delta$:

1. For the single-index setting with $k^\star \geq 1$, we have

$$\max_{i\in[d]} \tilde{R}_i^{\pm} \leq \begin{cases} \dfrac{K\tilde{M}\log^{2C_2}\left(\frac{12nd}{\delta}\right)}{n} & q = 0, M \geq \|\boldsymbol{v}\|_0 + 2 \\ \dfrac{K\tilde{M}\log^{2C_2}\left(\frac{12nd}{\delta}\right)}{n} + \dfrac{C_q\left(\frac{c^{(k^\star-1)}C_{\sigma^*}}{1-c^2}\right)^2 |\boldsymbol{v}_i|^{2(k^\star-1)}\left[\|\boldsymbol{v}\|_q^2 \vee k^\star 2^{\frac{2}{q}}\right]}{M^{\frac{2}{q}-1}} & q \in (0,2). \end{cases}$$

2. For the multi-index setting, we have

$$\max_{i\in[d]} \tilde{R}_i^{\pm} \leq \begin{cases} \dfrac{K\tilde{M}\log^{2C_2}\left(\frac{12nd}{\delta}\right)}{n} & q = 0, M \geq \|\boldsymbol{V}\|_{2,0} + 2 \\ \dfrac{K\tilde{M}\log^{2C_2}\left(\frac{12nd}{\delta}\right)}{n} + \dfrac{C_q\left(\frac{C_{\sigma^*}}{1-c^2}\right)^2 (c\|\boldsymbol{V}_{i*}\|_2)^{1\pm1}\left[\|\boldsymbol{V}\|_{2,q}^2 \vee 2^{\frac{2}{q}+1}\right]}{M^{\frac{2}{q}-1}} & q \in (0,2). \end{cases}$$

Here, K is a positive constant depending on (C_1, C_2, r, Δ).

Proof By Lemma 69, we have

$$\begin{aligned} \|\widetilde{\nabla}_j R_n^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) - \nabla_j R^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\|_2^2 \\ \leq 5 \sup_{\substack{J\subseteq[d] \\ |J|=2M}} \left\| \left(\nabla_j R_n^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) - \nabla_j R^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right)\big|_{\mathcal{J}}\right\|_2^2 \\ + 4\|\nabla_j R^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) - \nabla_j R^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})|_{\mathrm{top}(M)}\|_2^2. \end{aligned} \tag{E.2}$$

For any $\mathcal{J} \subseteq [d]$ with $|\mathcal{J}| = 2M$, by using Jensen's inequality, we can show that

$$\left\| \left(\nabla_j R_n^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)}) - \nabla_j R^{\pm}(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right)\big|_{\mathcal{J}}\right\|_2^2 \leq \sup_{\substack{J\subseteq[d] \\ |J|=2M}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2^2. \tag{E.3}$$

By (E.2) and (E.3), we have for any $i \in [d]$,

$$\tilde{R}_i^{\pm} \le 5 \sup_{\substack{J\subseteq[d] \\ |J|=2M}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2^2 + \frac{4}{2m}\sum_{j=1}^{2m}\left\|\mathbb{E}\left[\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})\phi'_{\pm}(\langle\tilde{\boldsymbol{e}}_i,\boldsymbol{x}\rangle\,;b_j^{(0)})\boldsymbol{x}\right] - \mathbb{E}\left[\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})\phi'_{\pm}(\langle\tilde{\boldsymbol{e}}_i,\boldsymbol{x}\rangle\,;b_j^{(0)})\boldsymbol{x}\right]\Big|_{\text{top}(M)}\right\|_2^2.$$

If $q = 0$ and $M \ge \|\boldsymbol{V}\|_{2,0} + 2$, the statement follows for each item by Proposition 19. For $q > 0$, we have the following:

1. We consider $k^\star \ge 1$ and even. We have

$$\begin{aligned}
\tilde{R}_i^{\pm} &\overset{(a)}{\le} 5 \sup_{\substack{J\subseteq[d] \\ |J|=2M}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2^2 \\
&\quad + \frac{2q\left(1-\frac{q}{2}\right)^{\frac{2-q}{q}} M^{\frac{-2}{q}+1}}{2m}\sum_{j=1}^{2m}\left\|\mathbb{E}_{(\boldsymbol{x},y)}\left[\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})\phi'_{\pm}(\langle\tilde{\boldsymbol{e}}_i,\boldsymbol{x}\rangle\,;b_j^{(0)})\boldsymbol{x}\right]\right\|_q^2 \\
&\overset{(b)}{\le} 5 \sup_{\substack{J\subseteq[d] \\ |J|=2M}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2^2 \\
&\quad + 2q\left(1-\frac{q}{2}\right)^{\frac{2-q}{q}} M^{\frac{-2}{q}+1} 4^{\frac{(q-1)}{q}\vee 0} 2^{\frac{2}{q}-1\vee 0}\left(\frac{c^{k^\star-1}C_{\sigma^*}|\boldsymbol{v}_i|^{k^\star-1}}{1-c^2}\right)^2\left[\|\boldsymbol{V}\|_{2,q}^2 + k^\star\|\tilde{\boldsymbol{e}}_l\|_q^2\right] \\
&\le 5 \sup_{\substack{J\subseteq[d] \\ |J|=2M}} \sup_{\substack{\boldsymbol{w}\in S_M^{d-1} \\ b\in\mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2^2 + C_q M^{\frac{-2}{q}+1}\left(\frac{c^{k^\star-1}C_{\sigma^*}|\boldsymbol{v}_i|^{k^\star-1}}{1-c^2}\right)^2\left[\|\boldsymbol{V}\|_{2,q}^2 \vee k^\star 2^{\frac{2}{q}}\right]
\end{aligned}$$

   where we used Lemma 70 for (a), and Proposition 19 with $N = k^\star - 2$ and Proposition 72 for (b). By using Lemma 24 with $\frac{\delta}{2d}$ (for $i \in [d]$ and $(\pm)$ cases), we have the result.

2. By using $k^\star = 1$ for $(-)$ and $k^\star = 2$ for $(+)$ in the proof of first item, one can prove this item as well.

■

E.1.1. CONCENTRATION FOR $\overline{\gamma}_k$

Proposition 33 *Let $m = \Theta(d^\varepsilon)$ where $\varepsilon > 0$ is a small constant, $Z_i \sim_{iid} \mathcal{N}(0,1)$ for $i \in [m]$, and let $\tilde{\gamma}_k(\cdot)$ be as in (C.5). For any $u \in \mathbb{N}$, we have with probability at least $1 - d^{-u}$*

$$\frac{1}{m}\sum_{i=1}^{m}\tilde{\gamma}_k(Z_i)^2 \ge c_k(k-1)!$$

for d larger than a constant depending on (k, u, ε).

Proof For $p \geq 1$, by Jensen's inequality, we have $\mathbb{E}[|\tilde{\gamma}_k^2(Z) - \mathbb{E}[\tilde{\gamma}_k^2(Z)]|^p]^{1/p} \leq 2\mathbb{E}[\tilde{\gamma}_k^{2p}(Z)]^{1/p}$. For $k \geq 2$,

$$2\mathbb{E}[\tilde{\gamma}_k^{2p}(Z)]^{1/p} = \frac{2}{2\pi}\mathbb{E}[e^{-pZ^2} H_{e_{k-2}}^{2p}(Z)]^{1/p} \leq \frac{1}{\pi}\mathbb{E}[H_{e_{k-2}}^{2p}(Z)]^{1/p} \overset{(a)}{\leq} \frac{(2p-1)^{k-2}}{\pi}(k-2)!,$$

where we use Lemma 63 for (a). Therefore, if $Y_m := \sum_{i=1}^m \tilde{\gamma}_k(Z_i)^2 - \mathbb{E}[\tilde{\gamma}_k^2(Z)]$ and $K_p := \frac{(2p-1)^{k-2}}{\pi}(k-2)!$, by Lemma 49, we have

$$\mathbb{E}[Y_m^{2p}]^{1/2p} \leq C\Big[\sqrt{pK_2}\sqrt{m} + pm^{1/2p}K_p\Big] \Rightarrow \mathbb{P}\left[\left|\frac{1}{m}Y_m\right| \geq eC\left(\sqrt{\frac{pK_2}{m}} + \frac{pm^{1/2p}K_p}{m}\right)\right] \leq e^{-p}.$$

By using $p = u\log d$ and hiding all of the constants with k in C_k, we have for $k \geq 1$

$$\mathbb{P}\left[\left|\frac{1}{m}Y_m\right| \geq C_k\sqrt{\frac{(u\log d)^{(k-1)\vee 1}}{m}}\right] \leq d^u.$$

Therefore, with probability $1 - d^u$, we have

$$\frac{1}{m}\sum_{i=1}^m \tilde{\gamma}_k(Z_i)^2 \geq \mathbb{E}[\tilde{\gamma}_k(Z)^2] - C_k\sqrt{\frac{(u\log d)^{2(k-1)\vee 1}}{m}} \overset{(a)}{\geq} \frac{1}{2}\mathbb{E}[\tilde{\gamma}_k(Z)^2].$$

where for (a), we assume that d is larger than a constant depending on (k, u, ε). Since $\mathbb{E}[\tilde{\gamma}_k(Z)^2] \geq c_k(k-1)!$, where c_k is some k-dependent constant, the statement follows. ■

E.2. Main Results

Lemma 34 (Single-Index Setting) *Consider the single index setting. For $u \in \mathbb{N}$ and a small constant $\varepsilon > 0$, let*

$$m = \Theta(d^\varepsilon), \quad d \geq d(\gamma_{k^\star}, k^\star, u, \varepsilon) \vee 4M \text{ and } c \leq \frac{1}{\log d},$$

and $\rho_1, \rho_2 \geq 1$, where $d(\gamma_{k^\star}, k^\star, u, \varepsilon)$ is a constant depending on $(\gamma_{k^\star}, k^\star, u, \varepsilon)$. There exists a constant $K > 0$ that depends on $(C_1, C_2, \Delta, k^\star, C_{\sigma^*})$ such that if

$$n \geq \frac{KM^{k^\star}\log^2\left(\frac{35nd}{M}\right)\log^{2C_2}\left(18nd^{u+1}\right)\left(\rho_1\log^{\rho_2} d\right)^{k^\star}}{c^{2(k^\star-1)}}$$

$$M \geq \log(4nd^u) \vee \begin{cases} (\|\boldsymbol{v}\|_0 + 2) & q = 0 \\ (2-q)\left[\left(\|\boldsymbol{v}\|_q^2 \vee k^\star 2^{\frac{2}{q}}\right)\frac{q}{2}\left(\rho_1\log^{\rho_2} d\right)^{k^\star}\right]^{\frac{q}{2-q}} & q \in (0,2) \end{cases} \tag{E.4}$$

with probability at least $1 - 4d^{-u}$, Algorithm 1 returns $\mathcal{J} \subseteq [d]$ such that

$$\|\boldsymbol{v}|_{\mathcal{J}^c}\|_2^2 \leq K\frac{\gamma_{k^\star}^{-\left(\frac{2}{k^\star-1}\wedge 2\right)}}{\rho_1\log^{\rho_2} d}.$$

Proof We choose any $u \in \mathbb{N}$. We consider the intersection of the following events:

C.1 There exists $j \in [m]$ such that $b_j^{(0)} \geq 0$.

C.2 $\sup_{\substack{J \subseteq [d] \\ |J| = 2M}} \sup_{\substack{\boldsymbol{w} \in S_M^{d-1} \\ b \in \mathbb{R}}} \|\boldsymbol{Y}_\theta|_{\mathcal{J}}\|_2^2 \leq \frac{K\tilde{M}\log^{2C_2}(6nd^u)}{n}$

C.3 Proposition 32 holds with $\delta = d^{-u}$.

C.4 Proposition 33 holds with $\delta = d^{-u}$.

It is easy to verify that the intersection of (C.1)-(C.4) holds with probability at least $1 - 4d^{-u}$ when d is larger than a constant depending on $(k^\star, u, \epsilon)$. We consider $k^\star = 1$ and $k^\star > 1$ cases separately. For $k^\star = 1$, let $\tilde{\mathcal{J}}$ be the set of indices added in Line 3. For $j \in [m]$ with $b_j^{(0)} \geq 0$, we have

$$\begin{aligned}\left\|\frac{1}{2}\gamma_1 \boldsymbol{v}|_{\mathcal{J}^c}\right\|_2^2 \overset{(a)}{\leq} \left\|\tilde{\gamma}_1(b_j^{(0)})\gamma_1\boldsymbol{v}|_{\tilde{\mathcal{J}}^c}\right\|_2^2 &\overset{(b)}{=} \|\nabla_j R^-(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_d, \boldsymbol{b}^{(0)}) - \nabla_j R^-(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_d, \boldsymbol{b}^{(0)})|_{\tilde{\mathcal{J}}}\|_2^2 \\ &\overset{(c)}{\leq} \|\nabla_j R^-(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_d, \boldsymbol{b}^{(0)}) - \widetilde{\nabla}_j R_n^-(\tilde{\boldsymbol{e}}_d)\|_2^2 \end{aligned} \tag{E.5}$$

where we use $\mathcal{J} \supseteq \tilde{\mathcal{J}}$ and $b_j^{(0)} \geq 0$ (see (C.5)) in (a), $\nabla_j R^-(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_d, \boldsymbol{b}^{(0)}) = -a_j^{(0)}\tilde{\gamma}_1(b_j^{(0)})\gamma_1\boldsymbol{v}$ (since $\boldsymbol{V}_{d*} = 0$) in (b), and that $\left\|\boldsymbol{x}|_{\tilde{\mathcal{J}}^c}\right\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}|_{\tilde{\mathcal{J}}}\|_2$ in (c). By using (C.3) with $k^\star = 1$, we have

$$(\text{E.5}) \leq \begin{cases} \dfrac{K\tilde{M}\log^{2C_2}(12nd^{1+u})}{n} & q = 0, M \geq \|\boldsymbol{v}\|_0 + 2 \\[2ex] \dfrac{K\tilde{M}\log^{2C_2}(12nd^{1+u})}{n} + C_q C_{\sigma^*}^2 \dfrac{\left(\frac{1}{1-c^2}\right)^2 \left[\|\boldsymbol{v}\|_q^2 \vee 2^{\frac{2}{q}}\right]}{M^{\frac{2}{q}-1}} & q \in (0, 2). \end{cases}$$

By (E.4), the statement follows for $k^\star = 1$.
For $k^\star > 1$ and even , we assume d is high enough that

$$c \leq \frac{1}{4} \text{ and } \left[\left(\frac{\overline{\gamma}_{k^\star}|\gamma_{k^\star}|}{(k^\star - 1)!}\right)^{\frac{2}{k^\star - 1}} - 8\left(\frac{\sqrt{2}c}{1 - c^2}\right)^{\frac{2}{k^\star - 1}} C_{\sigma^*}^{\frac{2}{k^\star - 1}}\right] \geq \left(\frac{1}{2}c_{k^\star}\gamma_{k^\star}^2\right)^{\frac{1}{k^\star - 1}}, \tag{E.6}$$

where $c_{k^\star}$ is the constant in Proposition 33. Let

$$\begin{aligned} \boldsymbol{u} &:= 1/\sqrt{2m}\,(\|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_1, \boldsymbol{b}^{(0)})\|_F, \cdots, \|\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_d, \boldsymbol{b}^{(0)})\|_F) \\ \tilde{\boldsymbol{u}} &:= 1/\sqrt{2m}\,(\|\widetilde{\nabla} R^+(\tilde{\boldsymbol{e}}_1)\|_F, \|\widetilde{\nabla} R^+(\tilde{\boldsymbol{e}}_2)\|_F, \cdots, \|\widetilde{\nabla} R^+(\tilde{\boldsymbol{e}}_d)\|_F). \end{aligned} \tag{E.7}$$

In the following, we will first bound $\sum_{j \in \mathcal{J}^c} \boldsymbol{u}_j^{\frac{2}{k^\star - 1}}$, and then use Proposition 31 with (E.6) to prove our statement. Let $\tilde{\mathcal{J}}$ be the set of indices added on Line 4. By using Lemma 69, we can write

$$\sum_{j \in \mathcal{J}^c} \boldsymbol{u}_j^{\frac{2}{k^\star - 1}} \leq \|\boldsymbol{u} - \boldsymbol{u}|_{\tilde{\mathcal{J}}}\|_{\frac{2}{k^\star - 1}}^{\frac{2}{k^\star - 1}} \leq \|\boldsymbol{u} - \tilde{\boldsymbol{u}}|_{\text{top}(M)}\|_{\frac{2}{k^\star - 1}}^{\frac{2}{k^\star - 1}} \tag{E.8}$$

$$\leq 4\|\boldsymbol{u} - \boldsymbol{u}|_{\text{top}(M)}\|_{\frac{2}{k^\star - 1}}^{\frac{2}{k^\star - 1}} + 5 \sup_{\substack{\mathcal{I} \subseteq [d] \\ |\mathcal{I}| = 2M}} \sum_{i \in \mathcal{I}} |\boldsymbol{u}_i - \tilde{\boldsymbol{u}}_i|^{\frac{2}{k^\star - 1}}. \tag{E.9}$$

Moreover, by Corollary 18 (with $N = k^\star - 2$) and $c \leq 1/4$,

$$\boldsymbol{u}_i^{\frac{2}{k^\star-1}} = \left\|(1/\sqrt{2m})\nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\right\|_F^{\frac{2}{k^\star-1}} \leq (1+\sqrt{k^\star})^{\frac{2}{k^\star-1}} C_{\sigma^*}^{\frac{2}{k^\star-1}} \frac{c^2}{(1-c^2)^{\frac{2}{k^\star-1}}} |\boldsymbol{v}_i|^2$$

$$\leq 12 C_{\sigma^*}^{\frac{2}{k^\star-1}} c^2 |\boldsymbol{v}_i|^2, \tag{E.10}$$

where we use that $(1+\sqrt{k^\star})^{\frac{2}{k^\star-1}}$ is non-increasing for $k^\star \geq 2$ in the last step. By Lemma 70, we have

$$\|\boldsymbol{u} - \boldsymbol{u}|_{\text{top}(M)}\|_{\frac{2}{k^\star-1}}^{\frac{2}{k^\star-1}} \leq 12 c^2 C_{\sigma^*}^{\frac{2}{k^\star-1}} \|\boldsymbol{v} - \boldsymbol{v}|_{\text{top}(M)}\|_2^2 \leq 12 c^2 C_{\sigma^*}^{\frac{2}{k^\star-1}} \begin{cases} 0 & q = 0, M \geq \|\boldsymbol{v}\|_0 + 2 \\ \dfrac{\left(1-\frac{q}{2}\right)^{\frac{2-q}{q}} \frac{q}{2}\|\boldsymbol{v}\|_q^2}{M^{\frac{2}{q}-1}} & q \in (0,2) \end{cases}$$

$$\leq \frac{12 c^2 C_{\sigma^*}^{\frac{2}{k^\star-1}}}{\rho_1 \log^{\rho_2} d}, \tag{E.11}$$

where we used (E.4). Moreover, we have

$$\sup_{\substack{\mathcal{I}\subseteq[d] \\ |\mathcal{I}|=2M}} \sum_{i\in\mathcal{I}} |\boldsymbol{u}_i - \tilde{\boldsymbol{u}}_i|^{\frac{2}{k^\star-1}} \leq \sup_{\substack{\mathcal{I}\subseteq[d] \\ |\mathcal{I}|=2M}} \sum_{i\in\mathcal{I}} (2m)^{\frac{-1}{k^\star-1}} \|\widetilde{\nabla} R_n^+(\tilde{\boldsymbol{e}}_i) - \nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\|_F^{\frac{2}{k^\star-1}}, \tag{E.12}$$

where by (C.3), we have

$$\forall i \in [d];\ (2m)^{-1}\|\widetilde{\nabla} R_n^+(\tilde{\boldsymbol{e}}_i) - \nabla R^+(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_i, \boldsymbol{b}^{(0)})\|_F^2$$

$$\leq \begin{cases} \dfrac{K\tilde{M}\log^{2C_2}(12nd^{1+u})}{n} & q = 0, M \geq \|\boldsymbol{v}\|_0 + 2 \\ \dfrac{K\tilde{M}\log^2\left(\frac{35nd}{M}\right)\log^{2C_2}(12nd^{1+u})}{n} + \dfrac{C_q \frac{C_{\sigma^*}^2 c^{2(k^\star-1)}}{(1-c^2)^2}|\boldsymbol{v}_i|^{2(k^\star-1)}\left[\|\boldsymbol{v}\|_q^2 \vee k^\star 2^{\frac{2}{q}}\right]}{M^{\frac{2}{q}-1}} & q \in (0,2). \end{cases} \tag{E.13}$$

Therefore, by (E.4), we have (E.12) $\leq \frac{c^2\tilde{K}}{\rho_1 \log^{\rho_2} d}$, where $\tilde{K}$ depends on $(C_1, C_2, \Delta, k^\star, C_{\sigma^*})$. By (E.9) and (E.11), the statement follows. ■

Lemma 35 (Multi-Index Setting) *Consider the multi-index setting. For $u \in \mathbb{N}$ and a small constant $\varepsilon > 0$, let*

$$m = \Theta(d^\varepsilon), \quad d \geq d(\sigma_r(\boldsymbol{H}), u, \varepsilon) \vee 4M \ \text{ and } \ c \leq \frac{1}{\log d},$$

and $\rho_1, \rho_2 \geq 1$. There exists a constant $K > 0$ that depends on $(C_1, C_2, \Delta, r, C_{\sigma^*})$ such that if

$$n \geq \frac{KM^2\log^2\left(\frac{35nd}{M}\right)\log^{2C_2}\left(18nd^{u+1}\right)\left(\rho_1\log^{\rho_2} d\right)}{c^2}$$

$$M \geq \log(4nd^u) \vee \begin{cases} (\|\boldsymbol{V}\|_{2,0} + 2) & q = 0 \\ (2-q)\left[\left(\|\boldsymbol{V}\|_{2,q}^2 \vee 2^{\frac{2}{q}+1}\right)\frac{q}{2}\left(\rho_1\log^{\rho_2} d\right)\right]^{\frac{q}{2-q}} & q \in (0,2) \end{cases} \tag{E.14}$$

with probability at least $1 - 4d^{-u}$*, Algorithm 1 returns* $\mathcal{J} \subseteq [d]$ *such that*

$$\|\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}\|_2^2 \vee \|\boldsymbol{V}|_{\mathcal{J}^c}\|_F^2 \leq \frac{K\sigma_r^{-2}(\boldsymbol{H})}{\rho_1 \log^{\rho_2} d}.$$

Proof We will follow the same arguments in the proof of Lemma 34. We choose any $u \in \mathbb{N}$. We consider the intersection of (C.1)-(C.4) above, which holds with probability at least $1 - 4d^{-u}$.
For $\|\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}\|_2^2$, let $\tilde{\mathcal{J}}$ be the set of indices added in Line 3. For $j \in [m]$ with $b_j^{(0)} \geq 0$, we have

$$\left\|\frac{1}{2}\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}\right\|_2^2 \overset{(a)}{\leq} \left\|\tilde{\gamma}_1(b_j^{(0)})\mathbb{E}[y\boldsymbol{x}]|_{\tilde{\mathcal{J}}^c}\right\|_2^2 \overset{(b)}{=} \|\nabla_j R^-(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_d, \boldsymbol{b}^{(0)}) - \nabla_j R^-(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_d, \boldsymbol{b}^{(0)})|_{\tilde{\mathcal{J}}}\|_2^2, \tag{E.15}$$

where we use $\mathcal{J} \supseteq \tilde{\mathcal{J}}$ and $b_j^{(0)} \geq 0$ in (a) (see (C.5)), $\nabla_j R^-(\boldsymbol{a}^{(0)}, \tilde{\boldsymbol{e}}_d, \boldsymbol{b}^{(0)}) = -a_j^{(0)}\tilde{\gamma}_1(b_j^{(0)})\gamma_1\boldsymbol{v}$ (since $\boldsymbol{V}_{d*} = 0$) in (b). By (C.3), we have

$$\text{(E.15)} \leq \begin{cases} \dfrac{K\tilde{M}\log^{2C_2}\left(12nd^{1+u}\right)}{n} & q = 0, M \geq \|\boldsymbol{V}\|_{2,0} + 2 \\ \dfrac{K\tilde{M}\log^{2C_2}\left(12nd^{1+u}\right)}{n} + C_q C_{\sigma^*}^2 \dfrac{\left(\frac{1}{1-c^2}\right)^2 \left[\|\boldsymbol{V}\|_{2,q}^2 \vee 2^{\frac{2}{q}}\right]}{M^{\frac{2}{q}-1}} & q \in (0, 2). \end{cases}$$

By (E.14), the statement follows for $\|\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}\|_2^2$.

For $\|\boldsymbol{V}|_{\mathcal{J}^c}\|_F^2$, we assume d is high enough that

$$c \leq \frac{1}{4} \text{ and } \left[\overline{\gamma}_2^2\sigma_r^2(\boldsymbol{H}) - 16\left(\frac{c}{1-c^2}\right)^2 C_{\sigma^*}^2\right] \geq \frac{1}{2}c_2\sigma_r^2(\boldsymbol{H}). \tag{E.16}$$

where c_2 is the constant in Proposition 33 for $k = 2$. Let $\boldsymbol{u}$ and $\tilde{\boldsymbol{u}}$ be the vectors defined in (E.7) and let $\tilde{\mathcal{J}}$ be the set of indices added on Line 4. By following the arguments in (E.8)-(E.9) with $k^\star = 2$, we can write

$$\sum_{j \in \mathcal{J}^c} \boldsymbol{u}_j^2 \leq 4\|\boldsymbol{u} - \boldsymbol{u}|_{\text{top}(M)}\|_2^2 + 5 \sup_{\substack{\mathcal{I} \subseteq [d] \\ |\mathcal{I}| = 2M}} \sum_{i \in \mathcal{I}} |\boldsymbol{u}_i - \tilde{\boldsymbol{u}}_i|^2$$

For $\boldsymbol{v} := (\|\boldsymbol{V}_{1*}\|_2, \cdots, \|\boldsymbol{V}_{d*}\|_2)$, by following the arguments in (E.10) and (E.11), we can write that

$$\|\boldsymbol{u} - \boldsymbol{u}|_{\text{top}(M)}\|_2^2 \leq 12c^2C_{\sigma^*}^2\|\boldsymbol{v} - \boldsymbol{v}|_{\text{top}(M)}\|_2^2 \leq 12c^2C_{\sigma^*}^2 \begin{cases} 0 & q = 0, M \geq \|\boldsymbol{V}\|_{2,0} + 2 \\ \dfrac{\left(1 - \frac{q}{2}\right)^{\frac{2-q}{q}} \frac{q}{2}\|\boldsymbol{V}\|_{2,q}^2}{M^{\frac{2}{q}-1}} & q \in (0, 2) \end{cases}$$

$$\leq \frac{6c^2C_{\sigma^*}^2}{\rho_1 \log^{\rho_2} d}.$$

Moreover, by following the arguments in (E.12) and (E.13), we can show that

$$\sup_{\substack{\mathcal{I}\subseteq[d]\\|\mathcal{I}|=2M}} \sum_{i\in\mathcal{I}} |\boldsymbol{u}_i - \tilde{\boldsymbol{u}}_i|^2 \leq \begin{cases} \dfrac{KM^2\log^2\left(\frac{35nd}{M}\right)\log^{2C_2}(12nd^{1+u})}{n} & q=0, M\geq \|\boldsymbol{V}\|_{2,0}+2 \\ \dfrac{KM^2\log^2\left(\frac{35nd}{M}\right)\log^{2C_2}(12nd^{1+u})}{n} + \dfrac{rC_q\frac{C_{\sigma^*}^2c^2}{1-c^2}\left[\|\boldsymbol{V}\|_{2,q}^2 \vee 2^{\frac{2}{q}+1}\right]}{M^{\left(\frac{2}{q}-1\right)}} & q\in(0,2) \end{cases}$$

$$\leq \frac{2c^2}{\rho_1\log^{\rho_2} d} + \frac{32rC_{\sigma^*}^2c^2}{\rho_1\log^{\rho_2} d} \tag{E.17}$$

By the arguments between (E.16)-(E.17), the statement follows. ■

Appendix F. Feature Learning

F.1. Additional Notation and Terminology

In the following, we will use SI for the single-index setting and MI for the multi-index setting. In the following, we assume $|\mathcal{J}|\leq M$ and ignore the constants. For SI, we consider a polynomial link function $\sigma^*:\mathbb{R}\to\mathbb{R}$ such that $\sigma^*(t)=\sum_{k\leq p}c_kt^k$. For MI, we consider a polynomial link function $\sigma^*:\mathbb{R}^r\to\mathbb{R}$ and $\tilde{\sigma}^*(\boldsymbol{z})=\sigma^*(\boldsymbol{z})-\langle\mathbb{E}[y\boldsymbol{x}],\boldsymbol{z}\rangle=\sum_{k\leq p}\left\langle\tilde{\boldsymbol{T}}_{\boldsymbol{k}},\boldsymbol{z}^{\otimes k}\right\rangle$.

Henceforth, $\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)$ is a random vector independent of the remaining random variable unless otherwise stated. Let $\boldsymbol{w}_{\mathcal{J}}:=\frac{\boldsymbol{w}|_{\mathcal{J}}}{\|\boldsymbol{w}|_{\mathcal{J}}\|_2}$. Let $vec(\boldsymbol{T})$ denotes the vectorized version of the tensor $\boldsymbol{T}$ and

$$\boldsymbol{s}_{\mathcal{J}} := \begin{cases}\langle\boldsymbol{v},\boldsymbol{w}_{\mathcal{J}}\rangle^{k^\star-1} & \text{SI}\\ \boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{w}_{\mathcal{J}} & \text{MI}\end{cases}$$

$$z_k(\boldsymbol{s}_{\mathcal{J}}) := \begin{cases} 0 & \boldsymbol{s}_{\mathcal{J}}=0\\ c_k\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{s}_{\mathcal{J}}^{2k}]^{-1}\boldsymbol{s}_{\mathcal{J}}^k & \text{SI and } \boldsymbol{s}_{\mathcal{J}}\neq 0\\ \left\langle vec(\tilde{\boldsymbol{T}}_{\boldsymbol{k}}), \mathbb{E}\left[vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})^\top\right]^+ vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})\right\rangle & \text{MI and } \boldsymbol{s}_{\mathcal{J}}\neq 0\end{cases}$$

where A^+ denotes the pseudoinverse of A. We will use

$$\beta(b_l^{(0)}) := \begin{cases}\frac{\gamma_{k^\star}\tilde{\gamma}_{k^\star}(b_l^{(0)})}{(k^\star-1)!} & \text{SI}\\ \tilde{\gamma}_2(b_l^{(0)}) & \text{MI}\end{cases} \quad \text{and} \quad N^\tau := \sum_{l=1}^N \mathbb{1}_{|\beta(b_l^{(0)})|\geq\tau},$$

where τ will be specified later.

F.2. Auxiliary Results

Lemma 36 ((Damian et al., 2022, Lemma 9) with explicit constants) *Let $a\sim$ Unif$(\{-1,1\})$ and $b\sim\mathcal{N}(0,1)$. Then for any $k\geq 0$, there exists $v_k(a,b)$ such that for $|x|\leq 1$,*

$$\mathbb{E}\left[v_k(a,b)\phi(at+b)\right]=t^k \quad \textit{and} \quad \sup_{a,b}|v_k(a,b)|\leq 6\sqrt{2}(k+1)^2.$$

Proof By following the constants in (Damian et al., 2022, Lemma 9), we have the statement. ∎

Lemma 37 ((Damian et al., 2022, Lemma 21) with explicit constants) *Let $\sigma^* : \mathbb{R}^r \to \mathbb{R}$ be a polynomial of degree-p such that $\mathbb{E}[\sigma^*(\boldsymbol{z})^2] \leq 1$. There exists symmetric $\tilde{\boldsymbol{T}}_{\boldsymbol{0}}, \cdots, \tilde{\boldsymbol{T}}_{\boldsymbol{p}}$ such that $\sigma^*(\boldsymbol{z}) = \sum_{k=0}^{p} \left\langle \tilde{\boldsymbol{T}}_{\boldsymbol{k}}, \boldsymbol{z}^{\otimes k} \right\rangle$ where*

$$\|\tilde{\boldsymbol{T}}_{\boldsymbol{k}}\|_F^2 \leq \frac{2e^k}{k!}(e\sqrt{r})^{\lfloor \frac{p-k}{2} \rfloor}.$$

Consequently, we have $\sum_{k=0}^{p} \|\tilde{\boldsymbol{T}}_{\boldsymbol{k}}\|_F (k+1)^2 \leq C(e\sqrt{r})^{\frac{p}{4}}$, where $C > 0$ is a universal constant.

Proof Let $\sigma^*(\boldsymbol{z}) = \sum_{j=0}^{p} \frac{1}{j!} \langle \boldsymbol{T}_{\boldsymbol{j}}, \boldsymbol{H}_{\boldsymbol{e_k}} \rangle$. Then,

$$\tilde{\boldsymbol{T}}_{\boldsymbol{k}} k! = \nabla^k \sigma^*(0) = \sum_{j=0}^{p-k} \frac{1}{j!} \boldsymbol{\nabla^k T_{j+k}}[\boldsymbol{H}_{\boldsymbol{e_k}}(0)] \overset{(a)}{=} \sum_{\substack{j=0 \\ j \text{ even}}}^{p-k} \frac{(-1)^{j/2}(j-1)!!}{j!} \boldsymbol{\nabla^k T_{j+k}}[sym(\boldsymbol{I}_r^{\otimes \frac{j}{2}})]$$

where (a) follows by Lemma 12 and since $\boldsymbol{\nabla^k T_{j+k}}$ is symmetric by Lemma 7. Therefore,

$$\|\tilde{\boldsymbol{T}}_{\boldsymbol{k}} k!\|_F \overset{(a)}{\leq} \sum_{\substack{j=0 \\ j \text{ even}}}^{p-k} \frac{(j-1)!!}{j!} \|\boldsymbol{T}_{\boldsymbol{j+k}}\|_F \|sym(\boldsymbol{I}_r^{\otimes \frac{j}{2}})\|_F \overset{(b)}{\leq} \sum_{\substack{j=0 \\ j \text{ even}}}^{p-k} \frac{(j-1)!!}{j!} r^{\frac{j}{4}} \|\boldsymbol{T}_{\boldsymbol{j+k}}\|_F.$$

where (a) follows Cauchy-Schwartz inequality and Lemma 7, and (b) follows (Damian et al., 2023, Lemma 3). Therefore,

$$\|\tilde{\boldsymbol{T}}_{\boldsymbol{k}} k!\|_F^2 \overset{(a)}{\leq} \sum_{\substack{j=0 \\ j \text{ even}}}^{p-k} \frac{\|\boldsymbol{T}_{\boldsymbol{j+k}}\|_F^2}{(j+k)!} \sum_{\substack{j=0 \\ j \text{ even}}}^{p} \left(\frac{(j-1)!!}{j!}\right)^2 r^{\frac{j}{2}} (j+k)! \overset{(b)}{\leq} \sum_{\substack{j=0 \\ j \text{ even}}}^{p-k} \left(\frac{(j-1)!!}{j!}\right)^2 r^{\frac{j}{2}} (j+k)!$$
$$\overset{(c)}{\leq} k! \sum_{\substack{j=0 \\ j \text{ even}}}^{p-k} \binom{j+k}{k} r^{\frac{j}{2}}. \tag{F.1}$$

where (a) follows from Cauchy-Schwartz inequality, (b) follows $\mathbb{E}[\sigma^*(\boldsymbol{z})^2] \leq 1$, and (c) follows $(j-1)!!^2 \leq j!$. Therefore,

$$\text{(F.1)} \overset{(a)}{\leq} k! e^k \sum_{\substack{j=0 \\ j \text{ even}}}^{p-k} (e^2 r)^{\frac{j}{2}} = k! e^k \sum_{j=0}^{\lfloor \frac{p-k}{2} \rfloor} (e\sqrt{r})^j \overset{(b)}{\leq} 2k! e^k (e\sqrt{r})^{\lfloor \frac{p-k}{2} \rfloor}.$$

where (a) follows $\binom{j+k}{k} \leq e^{j+k}$. For the second part of the statement, let $\sup_{k \geq 0} \frac{2e^k (k+1)^4}{k!} = C < \infty$ (as $k!$ grows faster than $e^k(k+1)^4$). We have

$$\sum_{k=0}^{p} \|\tilde{\boldsymbol{T}}_{\boldsymbol{k}}\|_F (k+1)^2 \leq \sum_{k=0}^{p} \left(\frac{2e^k (k+1)^4}{k!}\right)^{1/2} (e\sqrt{r})^{\frac{p-k}{4}} \leq C^{1/2} \sum_{k=0}^{p} (e\sqrt{r})^{\frac{p-k}{4}} \leq \tilde{C}(e\sqrt{r})^{\frac{p}{4}}.$$

∎

Proposition 38 *We consider MI (i.e., $\boldsymbol{s}_{\mathcal{J}} = \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{w}_{\mathcal{J}}$). For $k \in \mathbb{N}$ and $d \geq 2k$, we have*

$$\inf_{\substack{\boldsymbol{T_k}:(\mathbb{R}^r)^{\otimes k}\to\mathbb{R} \\ \boldsymbol{T_k} \text{ is symmetric} \\ \|\boldsymbol{T_k}\|_F=1}} \left\langle vec(\boldsymbol{T_k}), \mathbb{E}_{\boldsymbol{w}}[vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})^\top]vec(\boldsymbol{T_k})\right\rangle \geq k!\frac{\sigma_r^{2k}(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})}{\mathbb{E}\left[\|\boldsymbol{w}|_{\mathcal{J}}\|_2^{2k}\right]}.$$

Proof Let $\boldsymbol{T_k} : (\mathbb{R}^r)^{\otimes k} \to \mathbb{R}$ be a symmetric tensor with $\|\boldsymbol{T_k}\|_F^2 = 1$. We have

$$\left\langle vec(\boldsymbol{T_k}), \mathbb{E}[vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})^\top]vec(\boldsymbol{T_k})\right\rangle = \mathbb{E}\left[\|\boldsymbol{w}|_{\mathcal{J}}\|_2^{2k}\right]^{-1}\mathbb{E}\left[\left\langle \boldsymbol{T_k}, (\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{w}|_{\mathcal{J}})^{\otimes k}\right\rangle^2\right], \tag{F.2}$$

where we use that $w/\|\boldsymbol{w}\|_2$ and $\|\boldsymbol{w}\|_2$ are independent. Let $\hat{\boldsymbol{T}}_{\boldsymbol{k}} : (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}$ such that

$$\hat{\boldsymbol{T}}_{\boldsymbol{k}}[\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k] = T_k[\boldsymbol{D}\boldsymbol{V}|_{\mathcal{J}}^\top \boldsymbol{u}_1, \cdots, \boldsymbol{D}\boldsymbol{V}|_{\mathcal{J}}^\top \boldsymbol{u}_k].$$

By using Lemma 8 and (Damian et al., 2022, Lemma 23), we have (F.2) $\geq k!\|\hat{\boldsymbol{T}}_{\boldsymbol{k}}\|_F^2\mathbb{E}\left[\|\boldsymbol{w}|_{\mathcal{J}}\|_2^{2k}\right]^{-1} \geq k!\sigma_r^{2k}(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})\mathbb{E}\left[\|\boldsymbol{w}|_{\mathcal{J}}\|_2^{2k}\right]^{-1}$. ■

Lemma 39 *There exists $\tau > 0$ (that depends on $(k^\star, \gamma_{k^\star})$ for SI and universal for MI) such that for $b \sim \mathcal{N}(0,1)$, we have*

$$\mathbb{P}\left[|\beta(b)| \geq \tau\right] \geq \frac{2}{3} \;\; \textit{and} \;\; \mathbb{P}\left[\frac{N^\tau}{N} \geq \frac{1}{3}\right] \geq 1 - \exp\left(-\frac{2N}{9}\right).$$

Proof In the following, we will prove an anti-concentration result for $\tilde{\gamma}_k(b)$, $k \in \mathbb{N}$. Note that by scaling the $k = k^\star$ case with $|\gamma_{k^\star}|$, the statement can be extended to SI. MI immediately follows from the $k = 2$ case.
For $k = 1$, since $\tilde{\gamma}_k(b) \sim$ Unif$[0,1]$, if we take $\tau = 1/3$, we have the first statement. For $k = 2$, since $\tilde{\gamma}_k(b) = \frac{e^{\frac{-b^2}{2}}}{\sqrt{2\pi}}$, if we choose $\tau = \frac{1}{e\sqrt{2\pi}}$, we have

$$\mathbb{P}\left[\tilde{\gamma}_k(b) \geq \tau\right] = \mathbb{P}\left[|b| \leq \sqrt{2}\right] \overset{(a)}{\geq} 1 - \frac{1}{e\sqrt{2}} \geq \frac{2}{3}.$$

where we use $\mathbb{P}[|b| \geq t] \leq \frac{e^{-t^2/2}}{t}$ for (a). For $k \geq 3$, we have

$$|\tilde{\gamma}_k(b)| \leq \frac{1/(e^2\sqrt{2\pi})}{(2C)^{k-2}}\left(\frac{\varepsilon}{k-2}\right)^{k-2}\sqrt{(k-2)!} \Rightarrow |b| \geq 2 \text{ OR } |H_{e_{k-2}}(-b)| \leq \left(\frac{\varepsilon/2C}{(k-2)}\right)^{k-2}\sqrt{(k-2)!},$$

where C is the constant appeared in (Carbery and Wright, 2001, Theorem 8). Therefore, if we choose

$$\tau = \frac{1/(e^2\sqrt{2\pi})}{(2C)^{k-2}}\left(\frac{\varepsilon}{k-2}\right)^{k-2}\frac{\sqrt{(k-2)!}}{(k-1)!},$$

by (Carbery and Wright, 2001, Theorem 8), we have

$$\mathbb{P}\left[|\tilde{\gamma}_k(b)| \le \tau\right] \le \mathbb{P}\left[|b| \ge 2\right] + \mathbb{P}\left[H^2_{e_{k-2}}(-b) \le \frac{1}{C^{2k-4}}\left(\frac{\varepsilon}{2k-4}\right)^{2k-4}(k-2)!\right] \le \frac{1}{2e^2} + \varepsilon.$$

By choosing $\varepsilon = \frac{1}{6}$, we have the first part of the statement for $k \ge 3$ as well. The second part follows from Hoeffding's inequality and the result in first part.

■

F.2.1. LEMMAS FOR MOMENTS

Lemma 40 *For any event E,*

$$\begin{aligned} SI: \quad & \left|\mathbb{E}_{\boldsymbol{w}}\left[z_k(\boldsymbol{w}_{\mathcal{J}})\boldsymbol{s}^k_{\mathcal{J}}\left\langle \boldsymbol{v}, \boldsymbol{x}_i\right\rangle^k \mathbb{1}_E\right]\right| \le |c_k| 9^{k(k^\star-1)} |\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle|^k \mathbb{P}[E]^{1/2} \\ MI: \quad & \left|\mathbb{E}_{\boldsymbol{w}}\left[z_k(\boldsymbol{w}_{\mathcal{J}})\left\langle \boldsymbol{s}_{\mathcal{J}}, \boldsymbol{V}^\top \boldsymbol{x}_i\right\rangle^k \mathbb{1}_E\right]\right| \le \frac{2^k}{(4k)^{1/4}} \frac{\sigma_1^k(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})}{\sigma_r^k(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})} \|\tilde{\boldsymbol{T}}_{\boldsymbol{k}}\|_F \|\boldsymbol{V}^\top \boldsymbol{x}_i\|_2^k \mathbb{P}[E]^{1/4}. \end{aligned}$$

Proof For SI:

$$\begin{aligned} \left|\mathbb{E}_{\boldsymbol{w}}\left[z_k(\boldsymbol{w}_{\mathcal{J}})\boldsymbol{s}^k_{\mathcal{J}}\left\langle \boldsymbol{v}, \boldsymbol{x}_i\right\rangle^k \mathbb{1}_E\right]\right| & \overset{(a)}{\le} |c_k| |\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle|^k \mathbb{E}_{\boldsymbol{w}}\left[\left(\frac{\boldsymbol{s}^{2k}_{\mathcal{J}}}{\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{s}^{2k}_{\mathcal{J}}]}\right)^2\right]^{1/2} \mathbb{P}[E]^{1/2} \\ & \overset{(b)}{\le} |c_k| |\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle|^k 9^{k(k^\star-1)} \mathbb{P}[E]^{1/2}, \end{aligned}$$

where we used Cauchy-Schwartz inequality for (a) and Lemma 63 for (b).
For MI: By using Cauchy-Schwartz inequality,

$$\mathbb{E}_{\boldsymbol{w}}\left[z_k(\boldsymbol{w}_{\mathcal{J}})\left\langle \boldsymbol{s}_{\mathcal{J}}, \boldsymbol{V}^\top \boldsymbol{x}_i\right\rangle^k \mathbb{1}_E\right] \le \mathbb{E}_{\boldsymbol{w}}\left[z_k^2(\boldsymbol{w}_{\mathcal{J}})\right]^{1/2} \mathbb{E}\left[\left\langle \boldsymbol{s}_{\mathcal{J}}, \boldsymbol{V}^\top \boldsymbol{x}_i\right\rangle^{4k}\right]^{1/4} \mathbb{P}[E]^{1/4}. \quad \text{(F.3)}$$

We have

$$\begin{aligned} \mathbb{E}\left[\left\langle \boldsymbol{s}_{\mathcal{J}}, \boldsymbol{V}^\top \boldsymbol{x}_i\right\rangle^{4k}\right]^{1/4} & = \|(\boldsymbol{H}\boldsymbol{x}_i)|_{\mathcal{J}}\|_2^{4k}(4k-1)!!\mathbb{E}\left[\|\boldsymbol{w}|_{\mathcal{J}}\|_2^{4k}\right]^{-1} \\ & \le \sigma_1^{4k}(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})\|\boldsymbol{V}^\top \boldsymbol{x}_i\|_2^{4k}(4k-1)!!\mathbb{E}_{\boldsymbol{w}}\left[\|\boldsymbol{w}|_{\mathcal{J}}\|_2^{4k}\right]^{-1}, \end{aligned} \quad \text{(F.4)}$$

where we used $(\boldsymbol{H}\boldsymbol{x}_i)|_{\mathcal{J}} = \boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{x}_i$ in the last step. Moreover, we have

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{w}}\left[z_k^2(\boldsymbol{w}_{\mathcal{J}})\right] \\ & = \mathbb{E}_{\boldsymbol{w}}\left[\left\langle vec(\tilde{\boldsymbol{T}}_{\boldsymbol{k}}), \mathbb{E}\left[vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})^\top\right]^+ vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})\right\rangle \left\langle vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k}), \mathbb{E}\left[vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})^\top\right]^+ vec(\tilde{\boldsymbol{T}}_{\boldsymbol{k}})\right\rangle\right] \\ & = \left\langle vec(\tilde{\boldsymbol{T}}_{\boldsymbol{k}}), \mathbb{E}\left[vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})^\top\right]^+ vec(\tilde{\boldsymbol{T}}_{\boldsymbol{k}})\right\rangle \\ & \le \frac{\mathbb{E}\left[\|\boldsymbol{w}|_{\mathcal{J}}\|_2^{2k}\right]}{k!\sigma_r^{2k}(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})} \end{aligned} \quad \text{(F.5)}$$

where we used Proposition 38 in the last line. By using (F.4) and (F.5), we have

$$(F.3) \leq \left(\frac{(4k-1)!!}{k!k!}\right)^{1/4} \frac{\sigma_1^k(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})}{\sigma_r^k(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})} \|\boldsymbol{V}^\top \boldsymbol{x}_i\|_2^k \mathbb{P}[E]^{1/4} \leq \frac{2^k}{(4k)^{1/4}} \frac{\sigma_1^k(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})}{\sigma_r^k(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})} \|\boldsymbol{V}^\top \boldsymbol{x}_i\|_2^k \mathbb{P}[E]^{1/4},$$

where we use Stirling's formula in the last step. ■

F.3. Approximation of the target

We define

$$h(\boldsymbol{w}, a^{(0)}, b^{(1)}, b_l^{(0)}) := \sum_{k=0}^{p} \frac{v_k(a^{(0)}, b^{(1)})}{\eta^k \beta^k(b_l^{(0)})} z_k(\boldsymbol{s}_{\mathcal{J}}) \mathbb{1}_E,$$

where

$$E \equiv \begin{cases} |\boldsymbol{s}_{\mathcal{J}}| \leq \frac{1}{\eta\tau} \text{ AND } \|\boldsymbol{v}|_{\mathcal{J}^c}\|_2^2 \leq \frac{1}{4} \text{ AND } |\beta(b_l^{(0)})| \geq \tau \text{ AND } \max_{i\in[n]} \eta |\beta(b_l^{(0)}) \boldsymbol{s}_{\mathcal{J}} \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle| \leq 1 & \text{SI} \\ \|\boldsymbol{s}_{\mathcal{J}}\|_2 \leq \frac{1}{\eta\tau} \text{ AND } \|\boldsymbol{V}|_{\mathcal{J}^c}\|_F^2 \leq \frac{1}{4} \text{ AND } |\beta(b_l^{(0)})| \geq \tau \text{ AND } \max_{i\in[n]} \eta |\beta(b_l^{(0)}) \langle \boldsymbol{s}_{\mathcal{J}}, \boldsymbol{V}^\top \boldsymbol{x}_i \rangle| \leq 1 & \text{MI} \end{cases}$$

Lemma 41 *Let us have iid $\{b_l^{(0)}\}_{l\in[N]}$. We assume that: For SI, $M \geq 2p(k^\star - 1)$, $N^\tau > 0$ and $\|\boldsymbol{v}|_{\mathcal{J}^c}\|_2^2 \leq \frac{1}{4}$. For MI, $M \geq 2p$, $N^\tau > 0$ and $\|\boldsymbol{V}|_{\mathcal{J}^c}\|_F^2 \leq \frac{1}{4}$. Then, there exists a constant $C_{k^\star} > 0$ depending on $k^\star$, and a universal constant $\tilde{C} > 0$ such that the following holds:*

For SI:

$$\text{(i)} \left| \mathbb{E}_{(\boldsymbol{w}, a^{(0)}, b^{(1)})} \left[\frac{1}{N^\tau} \sum_{l=1}^{N} h(\boldsymbol{w}, a^{(0)}, b^{(1)}, b_l^{(0)}) \phi\left(a^{(0)} \eta \beta(b_l^{(0)}) \boldsymbol{s}_{\mathcal{J}} \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle + b^{(1)}\right)\right] - \sigma^*(\langle \boldsymbol{v}, \boldsymbol{x}_i \rangle) \right|$$

$$\leq C_{k^\star} e^{\frac{p}{4}} \left(\max_{k\leq p} |\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle|^k\right) \mathbb{P}_{\boldsymbol{w}} \left[|\boldsymbol{s}_{\mathcal{J}}| \geq \frac{1}{\eta\tau} \textit{ OR } \max_{i\in[n]} |\boldsymbol{s}_{\mathcal{J}} \langle \boldsymbol{v}, \boldsymbol{x}_i\rangle| > \frac{1}{\eta k^\star}\right]^{\frac{1}{2}} \tag{F.6}$$

$$\text{(ii)} \quad |h(\boldsymbol{w}, a^{(0)}, b^{(1)}, b^{(0)})| \leq \tilde{C} e^{\frac{p}{4}} \max_{k\leq p} \frac{M^{k(k^\star-1)}}{\eta^{2k}\tau^{2k}}. \tag{F.7}$$

For MI:

$$\text{(i)} \left| \mathbb{E}_{(\boldsymbol{w}, a^{(0)}, b^{(1)})} \left[\frac{1}{N^\tau} \sum_{l=1}^{N} h(\boldsymbol{w}, a^{(0)}, b^{(1)}, b_l^{(0)}) \phi\left(a^{(0)} \eta \beta(b_l^{(0)}) \langle \boldsymbol{s}_{\mathcal{J}}, \boldsymbol{V}^\top \boldsymbol{x}_i \rangle + b^{(1)}\right)\right] - \tilde{\sigma}^*(\boldsymbol{V}^\top \boldsymbol{x}_i) \right|$$

$$\leq C_{k^\star} (e\sqrt{r})^{\frac{p}{4}} \left(\frac{\sigma_1(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})}{\sigma_r(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})}\right)^p \left(\max_{k\leq p} \|\boldsymbol{V}^\top \boldsymbol{x}_i\|_2^k\right) \mathbb{P}_{\boldsymbol{w}} \left[\|\boldsymbol{s}_{\mathcal{J}}\|_2 \geq \frac{1}{\eta\tau} \textit{ OR } \max_{i\in[n]} |\langle \boldsymbol{s}_{\mathcal{J}}, \boldsymbol{V}^\top \boldsymbol{x}_i\rangle| > \frac{1}{\eta}\right]^{\frac{1}{4}} \tag{F.8}$$

$$\text{(ii)} \; |h(\boldsymbol{w}, a^{(0)}, b^{(1)}, b^{(0)})| \leq \tilde{C} (e\sqrt{r})^{\frac{p}{4}} \max_{k\leq p} \frac{M^k}{\eta^{2k}\tau^{2k}\sigma_r^{2k}(\boldsymbol{D})}. \tag{F.9}$$

Proof We start with SI. Fix an $k \le p$ and $l \in [N]$. We have

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{w},a^{(0)},b^{(1)})}\left[\frac{1}{N^\tau}\sum_{l=1}^{N} h(\boldsymbol{w},a^{(0)},b^{(1)},b_l^{(0)})\phi\left(a^{(0)}\eta\beta(b_l^{(0)})\boldsymbol{s}_{\mathcal{J}}\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle + b^{(1)}\right)\right] \\
&\overset{(a)}{=} \mathbb{1}_{|\beta(b_l^{(0)})|\ge\tau}\mathbb{E}_{\boldsymbol{w}}\left[\mathbb{1}_E z_k(\boldsymbol{s}_{\mathcal{J}})\boldsymbol{s}_{\mathcal{J}}^k\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle^k\right] \\
&\overset{(b)}{=} \mathbb{1}_{|\beta(b_l^{(0)})|\ge\tau}\left(c_k\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle^k - \mathbb{E}_{\boldsymbol{w}}\left[\mathbb{1}_{E^c} z_k(\boldsymbol{s}_{\mathcal{J}})\boldsymbol{s}_{\mathcal{J}}^k\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle^k\right]\right)
\end{aligned} \tag{F.10}$$

where we use Lemma 36 in (a) and the definition of z_k and $\|\boldsymbol{v}|_{\mathcal{J}}\|_2^2 > 0$ in (b). Therefore, we have

$$\begin{aligned}
&\left|\mathbb{E}_{(\boldsymbol{w},a^{(0)},b^{(1)})}\left[\frac{1}{N^\tau}\sum_{l=1}^{N} h(\boldsymbol{w},a^{(0)},b^{(1)},b_l^{(0)})\phi\left(a^{(0)}\eta\beta(b_l^{(0)})\boldsymbol{s}_{\mathcal{J}}\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle - b^{(1)}\right)\right] - \sigma^*(\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle)\right| \\
&\le \left|\sum_{k=0}^{p}\sum_{l=1}^{N}\frac{\mathbb{1}_{|\beta(b_l^{(0)})|\ge\tau}}{N^\tau}\mathbb{E}_{\boldsymbol{w}}\left[\mathbb{1}_{E^c} z_k(\boldsymbol{s}_{\mathcal{J}})\boldsymbol{s}_{\mathcal{J}}^k\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle^k\right]\right| \\
&\overset{(a)}{\le} \left(\max_{k\le p}|\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle|^k\right)\mathbb{P}[E^c]^{1/2}\sum_{k=0}^{p}|c_k|9^{k(k^\star-1)}
\end{aligned}$$

where we use Lemma 40 for (a). By Lemma 37, we have

$$\sum_{k=0}^{p}|c_k|9^{k(k^\star-1)} \overset{(a)}{\le} \sum_{k=0}^{p}\frac{\sqrt{2}9^{kk^\star}}{\sqrt{k!}}e^{\frac{p-k}{4}} \le Ce^{\frac{9^{2k^\star}}{2}}e^{\frac{p}{4}}. \tag{F.11}$$

where (a) follows $9 \ge \sqrt{e}$. By observing that $|\beta(b_l^{(0)})| \le k^\star$ and $E^c \Rightarrow \max_{i\in[n]}|\boldsymbol{s}_{\mathcal{J}}\left\langle \boldsymbol{v},\boldsymbol{x}_i\right\rangle| > \frac{1}{\eta k^\star}\ OR\ |\boldsymbol{s}_{\mathcal{J}}| \ge \frac{1}{\eta\tau}$, we have (F.6). For (F.7), by Lemma 36, we have

$$\frac{|v_k(a^{(0)},b^{(1)})|}{\eta^k\beta^k(b^{(0)})} \le \frac{6\sqrt{2}(k+1)^2}{\eta^k\tau^k}. \tag{F.12}$$

Moreover,

$$|z_k(\boldsymbol{s}_{\mathcal{J}})|\mathbb{1}_E \overset{(a)}{\le} \frac{|c_k|}{\eta^k\tau^k}\frac{1}{\mathbb{E}_{\boldsymbol{w}}\left[\boldsymbol{s}_{\mathcal{J}}^{2k}\right]} \overset{(b)}{\le} \frac{|c_k|}{\eta^k\tau^k}\frac{4^{k(k^\star-1)}M^{k(k^\star-1)}}{\left(2k(k^\star-1)\right)!!} \overset{(c)}{\le} \frac{e^2|c_k|}{\eta^k\tau^k}M^{k(k^\star-1)},$$

where we use $E \Rightarrow |\boldsymbol{s}_{\mathcal{J}}| \le \frac{1}{\eta\tau}$ for (a), $\|\boldsymbol{v}|_{\mathcal{J}^c}\|_2^2 \le \frac{1}{4}$ and $M \ge 2p(k^\star-1)$ for (b), and $\frac{4^{k(k^\star-1)}}{\left(2k(k^\star-1)\right)!!} = \frac{2^{k(k^\star-1)}}{\left(k(k^\star-1)\right)!} \le e^2$ for (c). Therefore,

$$|h(\boldsymbol{w},a^{(0)},b^{(1)},b^{(0)})| \le \sum_{k=0}^{p}\frac{M^{k(k^\star-1)}}{\eta^{2k}\tau^{2k}}6e^2\sqrt{2}(k+1)^2|c_k| \overset{(a)}{\le} \tilde{C}e^{\frac{p}{4}}\max_{k\le p}\frac{M^{k(k^\star-1)}}{\eta^{2k}\tau^{2k}},$$

where we used Lemma 37 for (a). For MI, by adjusting the arguments between (F.10)-(F.11) by using the bounds for MI proven above, we can obtain (F.8). For (F.9), we observe that

$$|z_k(\boldsymbol{s}_{\mathcal{J}})|\mathbb{1}_E \overset{(a)}{\leq} \|\boldsymbol{s}_{\mathcal{J}}\|_2^k \Big\| \mathbb{E}\left[vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})vec(\boldsymbol{s}_{\mathcal{J}}^{\otimes k})^\top\right]^+ vec(\tilde{\boldsymbol{T}}_{\boldsymbol{k}}) \Big\|_2 \mathbb{1}_E \overset{(b)}{\leq} \frac{1}{\eta^k \tau^k} \frac{\mathbb{E}\left[\|\boldsymbol{w}|_{\mathcal{J}}\|_2^{2k}\right]}{k!\sigma_r^{2k}(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})} \|\tilde{\boldsymbol{T}}_{\boldsymbol{k}}\|_F \mathbb{1}_E$$
$$\overset{(c)}{\leq} \frac{e^4}{\eta^k \tau^k} \frac{M^k}{\sigma_r^{2k}(\boldsymbol{D})} \|\tilde{\boldsymbol{T}}_{\boldsymbol{k}}\|_F \mathbb{1}_E$$

where we used Cauchy Schwartz inequality for (a), Proposition 38 and $E \Rightarrow \|\boldsymbol{s}_{\mathcal{J}}\|_2 \leq \frac{1}{\eta\tau}$ for (b), and $E \Rightarrow \|\boldsymbol{V}|_{\mathcal{J}^c}\|_F \leq \frac{1}{2}$, $M \geq 2p$, and $\frac{4^k}{k!} \leq e^4$ for (c). By (F.12) and Lemma 37, we have

$$|h(\boldsymbol{w}, a^{(0)}, b^{(1)}, b^{(0)})| \leq \sum_{k \leq p} \frac{M^k 6\sqrt{2}e^4(k+1)^2}{\eta^{2k}\tau^{2k}\sigma_r^{2k}(\boldsymbol{D})} \|\tilde{\boldsymbol{T}}_{\boldsymbol{k}}\|_F \leq \tilde{C}(e\sqrt{r})^{\frac{p}{4}} \max_{k \leq p} \frac{M^k}{\eta^{2k}\tau^{2k}\sigma_r^{2k}(\boldsymbol{D})}.$$

■

F.4. Empirical Approximation

For the following theorem, we introduce:

$$\|X\|_{\psi_2} \coloneqq \inf\left\{t > 0 \mid \mathbb{E}_{(\boldsymbol{w},\boldsymbol{a}^{(0)},\boldsymbol{b}^{(1)})}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\}.$$

For the following, let us assume that we have i.i.d. $\{(\boldsymbol{w}_j, a_j^{(0)}, b_j^{(1)}, b_j^{(0)})\}_{j\in[m]}$ and for $B, N \in \mathbb{N}$, let $m = B \cdot N$. We will double index parameters as $\boldsymbol{w}_{jl} = \boldsymbol{w}_{(j-1)N+l}$, $j \in [B]$ and $l \in [N]$. Recall that

$$h(\boldsymbol{w}, a^{(0)}, b^{(1)}, b_l^{(0)}) \coloneqq \sum_{k=0}^{p} \frac{v_k(a^{(0)}, b^{(1)})}{\eta^k \beta^k(b_l^{(0)})} z_k(\boldsymbol{s}_{\mathcal{J}})\mathbb{1}_E$$

We let

$$Y_{jl} \coloneqq \begin{cases} h(\boldsymbol{w}_{jl}, a_{jl}^{(0)}, b_{jl}^{(1)}, b_{jl}^{(0)})\phi\left(a_{jl}^{(0)}\eta\beta\big(b_{jl}^{(0)}\big)\left\langle \boldsymbol{v}, \frac{\boldsymbol{w}_{jl}|_{\mathcal{J}}}{\|\boldsymbol{w}_{jl}|_{\mathcal{J}}\|_2}\right\rangle^{k^\star-1}\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle + b_{jl}^{(1)}\right) & \text{SI} \\ h(\boldsymbol{w}_{jl}, a_{jl}^{(0)}, b_{jl}^{(1)}, b_{jl}^{(0)})\phi\left(a_{jl}^{(0)}\eta\beta\big(b_{jl}^{(0)}\big)\left\langle \boldsymbol{D}\boldsymbol{V}^\top \frac{\boldsymbol{w}_{jl}|_{\mathcal{J}}}{\|\boldsymbol{w}_{jl}|_{\mathcal{J}}\|_2}, \boldsymbol{V}^\top \boldsymbol{x}_i\right\rangle + b_{jl}^{(1)}\right) & \text{MI} \end{cases}$$

Moreover let $Y_j \coloneqq \frac{1}{N^\tau}\sum_{l=1}^N Y_{jl}$ and $N_j^\tau \coloneqq \sum_{l=1}^N \mathbb{1}_{|\beta(b_{jl}^{(0)})| \geq \tau}$. We have the following statement:

Lemma 42 *We assume that: For SI, $M \geq 2p(k^\star - 1)$, and $N_j^\tau > N/3$. For MI: $M \geq 2p$, and $N_j^\tau > N/3$. Then, there exists a universal constant $\tilde{C} > 0$ such that*

$$\|Y_j - \mathbb{E}_{(\boldsymbol{w},\boldsymbol{a}^{(0)},\boldsymbol{b}^{(1)})}[Y_j]\|_{\psi_2} \leq \tilde{C} \begin{cases} \frac{e^{\frac{p}{4}}}{\sqrt{N}} \max_{k \leq p} \frac{M^{k(k^\star-1)}}{\eta^{2k}\tau^{2k}} & \textit{SI} \\ \frac{(e\sqrt{r})^{\frac{p}{4}}}{\sqrt{N}} \max_{k\leq p} \frac{M^k}{\eta^{2k}\tau^{2k}\sigma_r^{2k}(\boldsymbol{D})} & \textit{MI}. \end{cases}$$

Proof For both SI and MI, there exists a universal $C > 0$ such that we have

$$\left\| Y_j - \mathbb{E}_{(\boldsymbol{w},a^{(0)},b^{(1)})}[Y_j] \right\|_{\psi_2}^2 = \left\| \frac{1}{N^\tau} \sum_{l=1}^{N} Y_{jl} - \mathbb{E}_{(\boldsymbol{w},\boldsymbol{a}^{(0)},\boldsymbol{b}^{(1)})}[Y_{jl}] \right\|_{\psi_2}^2 \le C \sum_{l=1}^{N} \|Y_{jl}\|_{\psi_2}^2. \quad \text{(F.13)}$$

Since $\phi(t)^2 \le t^2$, for SI, we have

$$\begin{aligned} \|Y_{jl}\|_{\psi_2} &\le \left\| h(\boldsymbol{w}_{jl}, a_{jl}^{(0)}, b_{jl}^{(1)}, b_{jl}^{(0)}) \left(a_{jl}^{(0)} \eta \beta(b_{jl}^{(0)}) \left\langle \boldsymbol{v}, \frac{\boldsymbol{w}_{jl}|_{\mathcal{J}}}{\|w_{jl}|_{\mathcal{J}}\|_2} \right\rangle^{k^\star - 1} \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle + b_{jl}^{(1)} \right) \right\|_{\psi_2} \\ &\overset{(a)}{\le} \tilde{C} e^{\frac{p}{4}} \max_{k \le p} \frac{M^{k(k^\star - 1)}}{\eta^{2k} \tau^{2k}}, \end{aligned}$$

where (a) follows by the definition of E and $\|b_{jl}^{(1)}\|_{\psi_2} \le 3$. For MI, we have

$$\begin{aligned} \|Y_{jl}\|_{\psi_2} &\overset{(a)}{=} \left\| h(\boldsymbol{w}_{jl}, a_{jl}^{(0)}, b_{jl}^{(1)}, b_{jl}^{(0)}) \phi \left(\boldsymbol{a}_{jl}^{(0)} \eta \beta(b_{jl}^{(0)}) \left\langle \boldsymbol{D} \boldsymbol{V}^\top \frac{\boldsymbol{w}_{jl}|_{\mathcal{J}}}{\|\boldsymbol{w}_{jl}|_{\mathcal{J}}\|_2}, \boldsymbol{V}^\top \boldsymbol{x}_i \right\rangle + b_{jl}^{(1)} \right) \right\|_{\psi_2} \\ &\overset{(b)}{\le} \tilde{C} (e\sqrt{r})^{\frac{p}{4}} \max_{k \le p} \frac{M^k}{\eta^{2k} \tau^{2k} \sigma_r^{2k}(\boldsymbol{D})}, \end{aligned}$$

where (a) follows from $\phi(t)^2 \le t^2$, (b) follows by the definition of E and $\|\boldsymbol{b}^{(1)}\|_{\psi_2} \le 3$. By (F.13) and $N_j^\tau > N/3$, the statement follows. ∎

Let $poly(\cdot)$ a polynomial respectively, depending on $(p, k^\star, \gamma_{k^\star})$ for SI, and $(p, r, \sigma_1(\boldsymbol{D})/\sigma_r(\boldsymbol{D}))$ for MI, which will be defined later (see (F.14)). We define the following event:

$$\widetilde{E} \equiv \begin{cases} \left| \frac{1}{B} \sum_{j=1}^{B} Y_j - \sigma^*(\langle \boldsymbol{v}, \boldsymbol{x}_i \rangle) \right| \ge \frac{poly[\log n, \log d^u] \log^{\frac{1}{2}}\left(\frac{2n}{\delta}\right)}{\sqrt{m}} + \frac{1}{n} & \text{SI} \\ \left| \frac{1}{B} \sum_{j=1}^{B} Y_j - \tilde{\sigma}^*(\boldsymbol{V}^\top \boldsymbol{x}_i) \right| \ge \frac{poly[\log n, \log d^u] \log^{\frac{1}{2}}\left(\frac{2n}{\delta}\right)}{\sqrt{m}} + \frac{1}{n} & \text{MI} \end{cases}$$

Lemma 43 *There exists a constant $C > 0$ depending on $(k^\star, \gamma_{k^\star})$ for SI and r for MI such that if we have*

For SI:

1. $\max_{i \in [n]} |\langle \boldsymbol{v}, \boldsymbol{x}_i \rangle| \le \sqrt{3} \sqrt{1 + \log(4nd^u)}$.
2. $\eta = \frac{1}{C} \frac{1}{\tau \sqrt{1 + \log(4nd^u)}} \left(\frac{M}{1 + \log(P)} \right)^{\frac{k^\star - 1}{2}}$ *where* $P = n^2 \left[C \left(1 + \log\left(4nd^u\right) \right) \right]^p$.
3. $M \ge 2p(k^\star - 1) \vee 16 \log(P)$
4. $\|\boldsymbol{v}|_{\mathcal{J}^c}\|_2^2 \le 1/4$
5. $N_j^\tau \ge N/3$ *for all* $j \in [B]$

For MI:

1. $\max_{i \in [n]} \|\boldsymbol{V}^\top \boldsymbol{x}_i\| \le \sqrt{3} \sqrt{r + \log(4nd^u)}$.
2. $\eta = \frac{1}{C} \frac{1}{\tau \sigma_1(D) \sqrt{r + \log(4nd^u)}} \left(\frac{M}{r + \log(P)} \right)^{\frac{1}{2}}$ *where* $P = n^4 \left[C \left(r + \log\left(4nd^u\right) \right) \right]^{2p}$.
3. $M \ge 2p \vee 16 \log(P)$
4. $\|\boldsymbol{V}|_{\mathcal{J}^c}\|_F^2 \le 1/4$
5. $N_j^\tau \ge N/3$ *for all* $j \in [B]$

then, the following holds:

– $\max_{k\le p} \frac{e^{\frac{p}{4}} M^{k(k^\star-1)}}{\eta^{2k}\tau^{2k}} \le C^{2p} e^{\frac{p}{4}} (1+\log(4nd^u))^p (1+\log(P))^{p(k^\star-1)}$

– $\mathbb{P}_{(\boldsymbol{w},a^{(0)},b^{(1)})}[\widetilde{E}] \le \delta$

– $\max_{k\le p} \frac{(e\sqrt{r})^{\frac{p}{4}} M^k}{\eta^{2k}\tau^{2k}\sigma_r^{2k}(\boldsymbol{D})} \le C^{2p} (e\sqrt{r})^{\frac{p}{4}} \left(\frac{\sigma_1(\boldsymbol{D})}{\sigma_r(\boldsymbol{D})}\right)^{2p} (r+\log(4nd^u))^p (1+\log(P))^p$

– $\mathbb{P}_{(\boldsymbol{w},a^{(0)},b^{(1)})}[\widetilde{E}] \le \delta$

Proof For SI, we have

$$\begin{aligned}
\max_{k\le p} \frac{e^{\frac{p}{4}} M^{k(k^\star-1)}}{\eta^{2k}\tau^{2k}} &= e^{\frac{p}{4}} \left(\max_{k\le p} C^k (1+\log(4nd^u))^k (1+\log(P))^{k(k^\star-1)}\right) \\
&= C^{2p} e^{\frac{p}{4}} (1+\log(4nd^u))^p (1+\log(P))^p
\end{aligned}$$

For MI, we have

$$\begin{aligned}
&(e\sqrt{r})^{\frac{p}{4}} \left(\max_{k\le p} \frac{M^k}{\eta^{2k}\tau^{2k}\sigma_r^{2k}(\boldsymbol{D})}\right) \\
&\quad = (e\sqrt{r})^{\frac{p}{4}} \left(\frac{\sigma_1(\boldsymbol{D})}{\sigma_r(\boldsymbol{D})}\right)^{2p} \left(\max_{k\le p} C^{2k} (r+\log(4nd^u))^k (1+\log(P))^k\right) \\
&\quad = C^{2p} (e\sqrt{r})^{\frac{p}{4}} \left(\frac{\sigma_1(\boldsymbol{D})}{\sigma_r(\boldsymbol{D})}\right)^{2p} (r+\log(4nd^u))^p (r+\log(P))^p
\end{aligned}$$

Let

$$poly(\log n, \log d^u) \ge \begin{cases} C^p e^{\frac{p}{4}} (1+\log(4nd^u))^p (1+\log(P))^p & \text{SI} \\ C^p (e\sqrt{r})^{\frac{p}{4}} \left(\frac{\sigma_1(\boldsymbol{D})}{\sigma_r(\boldsymbol{D})}\right)^{2p} (r+\log(4nd^u))^p (1+\log(P))^p & \text{MI.} \end{cases} \tag{F.14}$$

By Lemma 42, for both SI and MI, we have

$$\mathbb{P}_{(\boldsymbol{w},\boldsymbol{a}^{(0)},\boldsymbol{b}^{(1)})}\left[\left|\frac{1}{B}\sum_{j=1}^{B} Y_j - \mathbb{E}_{(\boldsymbol{w},a^{(0)},b^{(1)})}[Y_j]\right| \ge \underbrace{poly(\log n, \log d^u)\sqrt{\frac{\log(2/\delta)}{m}}}_{:=A_1}\right] \le \delta.$$

By Lemma 41, we have

$$\begin{aligned}
&\text{SI}: \left|\mathbb{E}_{(\boldsymbol{w},\boldsymbol{a}^{(1)},\boldsymbol{b}^{(1)})}[Y_j] - \sigma^*(\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle)\right| \\
&\quad \le \underbrace{C_{k^\star} e^{\frac{p}{4}} \left(\max_{k\le p} |\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle|_2^k\right) \mathbb{P}_{\boldsymbol{w}}\left[|\boldsymbol{s}_{\mathcal{J}}| \ge \frac{1}{\eta\tau} \text{ OR } \max_{i\in[n]} |\boldsymbol{s}_{\mathcal{J}} \langle \boldsymbol{v}, \boldsymbol{x}_i\rangle| > \frac{1}{\eta k^\star}\right]^{\frac{1}{2}}}_{:=A_2}
\end{aligned}$$

$$\begin{aligned}
&\text{MI}: \left|\mathbb{E}_{(\boldsymbol{w},\boldsymbol{a}^{(0)},\boldsymbol{b}^{(1)})}[Y_j] - \tilde{\sigma}^*(\boldsymbol{V}^\top \boldsymbol{x}_i)\right| \\
&\le \underbrace{C_{k^\star} (e\sqrt{r})^{\frac{p}{4}} \left(\frac{\sigma_1(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})}{\sigma_r(\boldsymbol{V}|_{\mathcal{J}}\boldsymbol{D})}\right)^{p} \left(\max_{k\le p} \|\boldsymbol{V}^\top \boldsymbol{x}_i\|_2^k\right) \mathbb{P}_{\boldsymbol{w}}\left[\|\boldsymbol{s}_{\mathcal{J}}\|_2 \ge \frac{1}{\eta\tau} \text{OR} \max_{i\in[n]} \left|\langle \boldsymbol{s}_{\mathcal{J}}, \boldsymbol{V}^\top \boldsymbol{x}_i\rangle\right| > \frac{1}{\eta}\right]^{\frac{1}{4}}}_{:=A_2}
\end{aligned}$$

Therefore, for both SI and MI, we have

$$\mathbb{P}_{(\boldsymbol{w},\boldsymbol{a}^{(0)},\boldsymbol{b}^{(1)})}\left[\left|\frac{1}{B}\sum_{j=1}^{B}Y_j-\tilde{\sigma}^*(\boldsymbol{V}^\top\boldsymbol{x}_i)\right|\geq A_1+A_2\right]\leq\delta$$

For SI, by Lemmas 59 and 60, we have

$$\mathbb{P}\left[|\boldsymbol{s}_{\mathcal{J}}|\geq\frac{1}{\eta\tau}\right]\overset{(a)}{\leq}\frac{2}{P}\quad\text{and}\quad\mathbb{P}\left[\max_{i\in[n]}|\langle\boldsymbol{v},\boldsymbol{x}_i\rangle||\boldsymbol{s}_{\mathcal{J}}|\geq\frac{1}{\eta k^\star}\right]\overset{(b)}{\leq}\frac{2}{P},$$

where we choose $C\geq 1\vee\frac{k^\star\sqrt{3}}{\tau}6^{\frac{k^\star-1}{2}}$ for (a) and (b). Therefore, by choosing $C\geq 3\sqrt{e}(2C_{k^\star})^{2/p}$, we have

$$A_2\leq 2C_{k^\star}e^{\frac{p}{4}}\left(\sqrt{3}\sqrt{1+\log(4nd^u)}\right)^p\frac{1}{\sqrt{P}}\leq\frac{1}{n}.$$

For MI, the same argument with its corresponding bounds applies. ■

F.5. Concentration Bound for a Desirable Event

Corollary 44 *We fix $u\in\mathbb{N}$. For any $\varepsilon>0$, if*

$$m=\Theta(d^\varepsilon),\ \ d\geq O(M)\ \ \textit{and}\ \ c=\frac{1}{\log d},$$

n and M are chosen as in Lemmas 34 and 35 for SI and MI respectively, and

$$\eta=\frac{1}{\tau C}\begin{cases}\frac{1}{\sqrt{1+\log(4nd^u)}}\left(\frac{M}{1+\log(P)}\right)^{\frac{k^\star-1}{2}} & \textit{SI}\\ \frac{1/\sigma_1(H)}{\sqrt{r+\log(4nd^u)}}\left(\frac{M}{r+\log(P)}\right) & \textit{MI}\end{cases}\quad\textit{where}\ \ P=\begin{cases}n^2\left[C\left(1+\log\left(4nd^u\right)\right)\right]^p, & \textit{SI}\\ n^4\left[C\left(r+\log\left(4nd^u\right)\right)\right]^{2p}, & \textit{MI}\end{cases}\tag{F.15}$$

and C is the constant appeared in Lemma 43, we have with probability at least $1-(16+6m)d^{-u}$, the intersection of the

C.1 $\max_{j\in[2m]}\|\boldsymbol{W}_{j*}^{(1)}\|_2\leq\tilde{O}(1)$

C.2 $\|\hat{\boldsymbol{\mu}}|_{\mathcal{J}}\|_2\leq 1+O\left(\frac{1}{\sqrt{M}}\right)$

C.3 $\|\boldsymbol{b}^{(1)}\|_2^2\leq 4m$ *and* $\|\boldsymbol{b}^{(1)}\|_4^4\leq 6m$ *and* $\|\boldsymbol{b}^{(1)}\|_\infty\leq\tilde{O}(1)$

C.4 There exists $\hat{\mathbf{a}}\in\mathbb{R}^{2m}$ *such that*

$$\|\hat{\mathbf{a}}\|_2^2\leq\begin{cases}O\left(\frac{(1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{m}\right) & \textit{SI}\\ O\left(\frac{(r+\log(4nd^u))^{2p}(r+\log(P))^{2p}}{m}\right) & \textit{MI},\end{cases}$$

and

C.5 $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}(\boldsymbol{x}_i;(\hat{\boldsymbol{a}},\boldsymbol{W}^{(1)},\boldsymbol{b}^{(1)}))\right)^2 \le \Delta\mathbb{E}[\epsilon^2] + \tilde{O}\left(\frac{1}{m}+\frac{1}{\sqrt{n}}+\frac{1}{M}\right)$

$$+\begin{cases} O\left(\frac{1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{\rho_1\log^{\rho_2} d}\right) & \textit{SI} \\ O\left(\frac{r+\log(4nd^u))^{2p}(r+\log(P))^{2p}}{\rho_1\log^{\rho_2} d}\right) & \textit{MI} \end{cases}$$

where O suppresses constants, and $\tilde{O}$ suppresses constants and Poly $[\log n, \log d]$ *depending on the problem parameters* [3].

Let $\tau > 0$ be the values defined in Lemma 39, $N = \lfloor\sqrt{m}\rfloor$, and let

$$\hat{\boldsymbol{a}}_j := \frac{h(\boldsymbol{W}_{j*}, a_j^{(0)}, b_j^{(0)}, b_j^{(1)})}{BN_j^\tau}.$$

Moreover let

$$\tilde{y}_i := \begin{cases} \sum_{j=1}^{2m}\hat{\boldsymbol{a}}_j\phi\left(a_j^{(0)}\left\langle \boldsymbol{v}, \boldsymbol{W}_{j*}^{(0)}\right\rangle^{k^\star-1}\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle - b_j^{(1)}\right) & \text{SI} \\ \langle\hat{\boldsymbol{\mu}}|_{\mathcal{J}}, \boldsymbol{x}_i\rangle + \sum_{j=1}^{2m}\hat{\boldsymbol{a}}_j\phi\left(a_j^{(0)}\eta\beta(b_j^{(0)})\left\langle H\boldsymbol{W}_{j*}^{(0)}, \boldsymbol{x}_i\right\rangle - b_j^{(1)}\right) & \text{MI} \end{cases}$$

$$\hat{y}_i := \begin{cases} \left\langle \hat{\boldsymbol{a}}, \phi(\boldsymbol{W}^{(1)}\boldsymbol{x}_i + \boldsymbol{b}^{(1)})\right\rangle & \text{SI} \\ \langle\hat{\boldsymbol{\mu}}|_{\mathcal{J}}, \boldsymbol{x}_i\rangle + \left\langle \hat{\boldsymbol{a}}, \phi(\boldsymbol{W}^{(1)}\boldsymbol{x}_i + \boldsymbol{b}^{(1)})\right\rangle & \text{MI.} \end{cases}$$

We consider the intersection of the following events:

E.1 $N_j^\tau \ge N/3$ for all $j \in [B]$

E.2 For SI Proposition 28, for MI Proposition 30 holds for all $j \in [2m]$ with $\delta = d^{-u}$

E.3 For SI: $\|\boldsymbol{v}|_{\mathcal{J}^c}\|_2^2 \le O\left(\frac{1}{\rho_1\log^{\rho_2} d}\right)$. For MI: $\|\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}\|_2^2 \vee \|\boldsymbol{V}|_{\mathcal{J}^c}\|_2^2 \le O\left(\frac{1}{\rho_1\log^{\rho_2} d}\right)$.

E.4 We have

$$\max_{i\in[n]}|\langle \boldsymbol{v}, \boldsymbol{x}_i\rangle| \le \sqrt{3}\sqrt{1+\log(4nd^u)} \text{ and } \max_{i\in[n]}\|\boldsymbol{V}^\top\boldsymbol{x}_i\|_2 \le \sqrt{3}\sqrt{r+\log(4nd^u)},$$

for SI and MI respectively.

E.5 $\|\hat{\boldsymbol{a}}\|_2^2 \le \begin{cases} O\left(\frac{(1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{m}\right) & \text{SI} \\ O\left(\frac{(r+\log(4nd^u))^{2p}(r+\log(P))^{2p}}{m}\right) & \text{MI,} \end{cases}$

E.6 $\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \hat{y}_i)^2 \le \tilde{O}\left(\frac{1}{M}\right) + \begin{cases} O\left(\frac{(1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{\rho_1\log^{\rho_2} d}\right) & \text{SI} \\ O\left(\frac{(r+\log(4nd^u))^{2p}(r+\log(P))^{2p-1}}{\rho_1\log^{\rho_2} d}\right) & \text{MI,} \end{cases}$

E.7 For MI: $\frac{1}{n}\sum_{i=1}^{n}(\langle\mathbb{E}[y\boldsymbol{x}], \boldsymbol{x}_i\rangle - \langle\hat{\boldsymbol{\mu}}|_{\mathcal{J}}, \boldsymbol{x}_i\rangle)^2 \le O\left(\frac{1}{\rho_1\log^{\rho_2} d} + \frac{1}{M}\right)$

3. Specifically, $(k^\star, \gamma_{k^\star}, u, p, \varepsilon, \alpha, C_1, C_2, C_{\sigma^*}, \Delta)$ for SI, $(\sigma_1(H), \sigma_r(H), u, p, \varepsilon, \alpha, C_1, C_2, C_{\sigma^*}, r, \Delta)$ for MI.

Lemma 45 *With the choice of parameters in Corollary 44, the intersection of* (E.1)-(E.7) *holds with probability at least* $1-(11+4m)d^{-u}$.

Proof Since $N=\lfloor\sqrt{m}\rfloor$, by using Lemma 39 and union bound, we can show that (E.1) holds with probability at least $1-\Theta(d^{\varepsilon/2})\exp\left(-\Theta(d^{\varepsilon/2})\right)\geq 1-d^{u}$ for large enough d depending on (u,ε). Since with a sufficiently large constant factor, M satisfies the condition in Propositions 28 and 30, we have (E.2) holds with probability at least $1-2md^{-u}$. By Lemmas 34, 35 and the choice of parameters, we can show that (E.3) holds with probability at least $1-4d^{-u}$. By Corollary 58 we have that (E.4) holds with probability at least $1-d^{-u}$.
For (E.5), by Lemmas 41 and 43, we have

$$|\hat{\boldsymbol{a}}_j|\leq\begin{cases}O\left(\frac{N}{N_j^{\tau}}\frac{\tilde{C}}{m}\max_{k\leq p}\frac{M^{k(k^\star-1)}}{\eta^{2k}\tau^{2k}}\right), & \text{SI}\\ O\left(\frac{N}{N_j^{\tau}}\frac{\tilde{C}}{m}\max_{k\leq p}\frac{M^{k}}{\eta^{2k}\tau^{2k}\sigma_r^{2k}(H)}\right), & \text{MI}\end{cases}\leq\begin{cases}O\left(\frac{(1+\log(4nd^u))^p(1+\log(P))^{p(k^\star-1)}}{m}\right), & \text{SI}\\ O\left(\frac{(r+\log(4nd^u))^p(r+\log(P))^{p}}{m}\right) & \text{MI}\end{cases}$$

Hence, (E.5) follows. For the following, we additionally consider the intersection of the following events:

$\widetilde{\text{E}}.1$ Lemma 24 holds for $\phi(t)=t$ with $\delta=d^{-u}$.

$\widetilde{\text{E}}.2$ Lemma 26 holds for $\phi(t)=t$ with $\delta=d^{-u}$.

$\widetilde{\text{E}}.3$ Lemma 60 holds for all $\boldsymbol{W}_{j*}^{(0)}$, $j\in[2m]$, with $\delta=d^{-u}$.

$\widetilde{\text{E}}.4$ For SI, Lemma 61 holds for $\mathcal{A}=\left\{\frac{\boldsymbol{v}|_{\mathcal{J}^c}}{\|\boldsymbol{v}|_{\mathcal{J}^c}\|_2},\ |\mathcal{J}|\leq M\right\}$ with $\delta=d^{-u}$.

$\widetilde{\text{E}}.5$ For MI, Lemma 61 holds for $\mathcal{A}=\left\{\frac{\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}}{\|\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}\|_2},\ |\mathcal{J}|\leq M\right\}$ and conditioned on $\boldsymbol{W}$ (see (INIT)), holds for $\mathcal{A}=\left\{\frac{\boldsymbol{H}|_{\mathcal{J}^c\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)}}{\left\|\boldsymbol{H}|_{\mathcal{J}^c\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)}\right\|_2},\ |\mathcal{J}|\leq M\right\}$ each with $\delta=d^{-u}$.

Note that the intersection of the given events holds with probability at least $1-5d^{-u}-2md^{-u}$. For (E.6), we observe that $\boldsymbol{W}_{j*}^{(1)}=\eta a_j^{(0)}g(\boldsymbol{W}_{j*}^{(0)},b_j^{(0)})|_{\mathcal{J}}$, where g is defined in (D.1). By Cauchy-Schwartz and triangle inequalities, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i-\hat{y}_i)^2\\ &\leq 2\eta^2\|\hat{\boldsymbol{a}}\|_2^2\begin{cases}\sum_{j=1}^{2m}\frac{1}{n}\sum_{i=1}^{n}\left(\left\langle g(\boldsymbol{W}_{j*}^{(0)},b_j^{(0)})|_{\mathcal{J}}-\beta(b_j^{(0)})\left\langle\boldsymbol{v},\boldsymbol{W}_{j*}^{(0)}\right\rangle^{k^\star-1}\boldsymbol{v}|_{\mathcal{J}},\boldsymbol{x}_i\right\rangle\right)^2 & \text{SI}\\ \quad+\sum_{j=1}^{2m}\frac{1}{n}\sum_{i=1}^{n}\left(\beta(b_j^{(0)})\left\langle\boldsymbol{v},\boldsymbol{W}_{j*}^{(0)}\right\rangle^{k^\star-1}\langle\boldsymbol{v}|_{\mathcal{J}^c},\boldsymbol{x}_i\rangle\right)^2 & \\ \sum_{j=1}^{2m}\frac{1}{n}\sum_{i=1}^{n}\left(\left\langle g(\boldsymbol{W}_{j*}^{(0)},b_j^{(0)})|_{\mathcal{J}}-\beta(b_j^{(0)})\boldsymbol{H}|_{\mathcal{J}\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)},\boldsymbol{x}_i\right\rangle\right)^2 & \text{MI}\\ \quad+\sum_{j=1}^{2m}\frac{1}{n}\sum_{i=1}^{n}\left(\beta(b_j^{(0)})\left\langle\boldsymbol{H}|_{\mathcal{J}^c\times\mathcal{J}}\boldsymbol{W}_{j*}^{(0)},\boldsymbol{x}_i\right\rangle\right)^2 & \end{cases}\end{aligned}\tag{F.16}$$

Hence,

$$
\text{(F.16)} \overset{(a)}{\le} 4m\eta^2\|\hat{\boldsymbol{a}}\|_2^2 \begin{cases} O\left(\frac{M\log^2\left(\frac{24dn}{M}\right)\log^{2C_2}(12nd^u)}{n} + \left(\frac{1+\log(4d^u)}{M}\right)^{k^\star}\right) & \text{SI} \\ \quad +O\left(\frac{(1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{\rho_1\log^{\rho_2} d}\right) & \\ O\left(\frac{M\log^2\left(\frac{35dn}{M}\right)\log^{2C_2}(18nd^u)}{n} + \left(\frac{r+\log(4d^u)}{M}\right)^{2}\right) & \text{MI} \\ \quad +O\left(\frac{(r+\log(4nd^u))^{2p}(r+\log(P))^{2p}}{\rho_1\log^{\rho_2} d}\right) & \end{cases}
$$

$$
\overset{(b)}{\le} \tilde{O}\left(\tfrac{1}{M}\right) + \begin{cases} O\left(\frac{(1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{\rho_1\log^{\rho_2} d}\right) & \text{SI} \\ O\left(\frac{(r+\log(4nd^u))^{2p}(r+\log(P))^{2p-1}}{\rho_1\log^{\rho_2} d}\right) & \text{MI,} \end{cases}
$$

where we use (E.2), and ($\widetilde{\text{E}}$.2)-($\widetilde{\text{E}}$.5) for (a) and (E.5) and (F.15) for (b). Lastly,

$$
\frac{1}{n}\sum_{i=1}^{n}(\langle \mathbb{E}[y\boldsymbol{x}], \boldsymbol{x}_i\rangle - \langle \hat{\boldsymbol{\mu}}|_{\mathcal{J}}, \boldsymbol{x}_i\rangle)^2 \le 2\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top|_{\mathcal{J}\times\mathcal{J}}\right\|_2 \|(\hat{\boldsymbol{\mu}} - \mathbb{E}[y\boldsymbol{x}])|_{\mathcal{J}}\|_2^2 + \frac{2}{n}\sum_{i=1}^{n}\langle \mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}, \boldsymbol{x}_i\rangle^2
$$

$$
\overset{(a)}{\le} O\left(\frac{M\log^2\left(\frac{24dn}{M}\right)\log^{2C^2}(6nd^u)}{n} + \|\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}^c}\|_2^2\right),
$$

where we used ($\widetilde{\text{E}}$.1)- ($\widetilde{\text{E}}$.2) for (a). By (E.3), (E.7) follows. ■

Proof [Proof of Corollary 44] We assume the intersection of (E.1)-(E.7) and ($\widetilde{\text{E}}$.1)- ($\widetilde{\text{E}}$.5) holds. By recalling that $\boldsymbol{W}_{j*}^{(1)} = a_j^{(0)}\eta g\left(\boldsymbol{W}_{j*}^{(0)}, b_j^{(0)}\right)$, we have

$$
\|\boldsymbol{W}_{j*}^{(1)}\|_2 = \eta\left\|g\left(\boldsymbol{W}_{j*}^{(0)}, b_j^{(0)}\right)\right\|_2 \overset{(a)}{=} \eta\begin{cases} O\left(\left(\frac{1+\log(4d^u)}{M}\right)^{\frac{k^\star-1}{2}} + \sqrt{\frac{M\log^2\left(\frac{24dn}{M}\right)\log^{2C_2}(12nd^u)}{n}}\right) & \text{SI} \\ O\left(\left(\frac{r+\log(4d^u)}{M}\right)^{\frac{1}{2}} + \sqrt{\frac{M\log^2\left(\frac{35dn}{M}\right)\log^{2C_2}(18nd^u)}{n}}\right) & \text{MI} \end{cases}
$$

$$
\le \tilde{O}(1),
$$

where we use (E.2) in (a).

For (C.2), for SI $\hat{\boldsymbol{\mu}} = 0$, therefore, the statement is trivial in this case. For MI, by ($\widetilde{\text{E}}$.1), we can write

$$
\|\hat{\boldsymbol{\mu}}|_{\mathcal{J}}\| \le \|(\hat{\boldsymbol{\mu}} - \mathbb{E}[y\boldsymbol{x}])|_{\mathcal{J}}\|_2 + \|\mathbb{E}[y\boldsymbol{x}]|_{\mathcal{J}}\|_2 \overset{(a)}{\le} 1 + O\left(\sqrt{\frac{M\log^2\left(\frac{24dn}{M}\right)\log^{2C_2}(6nd^u)}{n}}\right)
$$

where (a) follows since $\|\mathbb{E}[y\boldsymbol{x}]\|_2 \le 1$.

For (C.3), by using Lemma 56, we have with probability $1 - d^{-u}$, for d is large enough

$$
\|\boldsymbol{b}^{(1)}\|_2^2 \le 2m + 2\sqrt{2m\log d^u} + 2\log d^u \le 3m.
$$

Moreover, by Lemma 63, we observe that $\mathbb{E}\left[\left(\frac{1}{2m}\sum_{j=1}^{2m} b_j^4 - 3\right)^p\right]^{1/p} \leq \frac{p^2\mathbb{E}[b_1^8]}{\sqrt{m}}$. Therefore, with probability $1-d^{-u}$, for d is large enough

$$\frac{1}{2m}\sum_{j=1}^{2m} b_j^4 - 3 \leq \frac{e\log^2 d^u \mathbb{E}[b_1^8]}{\sqrt{m}} \Rightarrow \|\boldsymbol{b}^{(1)}\|_4^4 \leq 7m$$

Moreover, by using standard Gaussian concentration with union bound, we have with probability $1-2md^{-u}$, $\|\boldsymbol{b}^{(1)}\|_\infty \leq \sqrt{\log(d^u)}$. (C.4) directly follows from (E.5).
For (C.5) in SI, we have

$$\frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}(\boldsymbol{x}_i;(\hat{\boldsymbol{a}},\boldsymbol{W}^{(1)},\boldsymbol{b}^{(1)})))^2 \leq \frac{1}{n}\sum_{i=1}^n(\sigma^*(\langle\boldsymbol{v},\boldsymbol{x}_i\rangle)-\hat{y}_i)^2 + \frac{\sqrt{\Delta}}{n}\sum_{i=1}^n(\sigma^*(\langle\boldsymbol{v},\boldsymbol{x}_i\rangle)-\hat{y}_i)\epsilon_i$$

$$+\frac{\Delta}{n}\sum_{i=1}^n \epsilon_i^2$$

By using $\delta = d^{-u}$ in Lemma 43 and (E.6), we have with probability at least $1-d^{-u}$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n(\sigma^*(\langle\boldsymbol{v},\boldsymbol{x}_i\rangle)-\hat{y}_i)^2 &\leq \frac{2}{n}\sum_{i=1}^n(\sigma^*(\langle\boldsymbol{v},\boldsymbol{x}_i\rangle)-\tilde{y}_i)^2 + \frac{2}{n}\sum_{i=1}^n(\tilde{y}_i-\hat{y}_i)^2\\ &\leq \tilde{O}\left(\frac{1}{m}+\frac{1}{n}+\frac{1}{M}\right) + O\left(\frac{(1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{\rho_1\log^{\rho_2} d}\right).\end{aligned}$$

Since ϵ_i has 1-Subgaussian norm, we have with probability at least $1-2d^{-u}$,

$$\begin{aligned}&\frac{\sqrt{\Delta}}{n}\sum_{i=1}^n(\sigma^*(\langle\boldsymbol{v},\boldsymbol{x}_i\rangle)-\tilde{y}_i)\epsilon_i \leq \sqrt{\frac{\Delta\log(2d^u)}{n}}\left(\frac{1}{n}\sum_{i=1}^n(\sigma^*(\langle\boldsymbol{v},\boldsymbol{x}_i\rangle)-\hat{y}_i)^2\right)^{1/2}\\ &\frac{1}{n}\sum_{i=1}^n\epsilon_i^2 - \mathbb{E}\epsilon_i^2 \leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{F.17}$$

Therefore, (C.5) follows for SI. For MI,

$$\frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}(\boldsymbol{x}_i;(\hat{\boldsymbol{a}},\boldsymbol{W}^{(1)},\boldsymbol{b}^{(1)})))^2 \leq \frac{1}{n}\sum_{i=1}^n(\sigma^*(\boldsymbol{V}^\top\boldsymbol{x}_i)-\hat{y}_i)^2 + \frac{\sqrt{\Delta}}{n}\sum_{i=1}^n(\sigma^*(\boldsymbol{V}^\top\boldsymbol{x}_i)-\hat{y}_i)\epsilon_i$$

$$+\frac{\Delta}{n}\sum_{i=1}^n \epsilon_i^2$$

We observe that

$$\begin{aligned}(\sigma^*(\boldsymbol{V}^\top\boldsymbol{x}_i)-\hat{y}_i)^2 &\leq 2(\sigma^*(\boldsymbol{V}^\top\boldsymbol{x}_i)-\tilde{y}_i)^2 + 2(\tilde{y}_i-\hat{y}_i)^2\\ &\leq 4\left(\tilde{\sigma}^*(\boldsymbol{V}^\top\boldsymbol{x}_i) - \sum_{j=1}^{2m}\hat{\boldsymbol{a}}_j\phi\left(a_j^{(0)}\eta\beta(b_j^{(0)})\left\langle H\boldsymbol{W}_{j*}^{(0)},\boldsymbol{x}_i\right\rangle - b_j^{(1)}\right)\right)^2\\ &\quad + 4\left(\langle\mathbb{E}[y\boldsymbol{x}],\boldsymbol{x}_i\rangle - \langle\hat{\boldsymbol{\mu}}|_{\mathcal{J}},\boldsymbol{x}_i\rangle\right)^2 + 2(\tilde{y}_i-\hat{y}_i)^2\end{aligned}$$

Therefore, by using $\delta = d^{-u}$ in Lemma 43 and by (E.6) and (E.7), we have with probability $1-d^{-u}$

$$\frac{1}{n}\sum_{i=1}^{n}(\sigma^*(\boldsymbol{V}^\top \boldsymbol{x}_i) - \hat{y}_i)^2 \leq O\left(\frac{(r+\log(4nd^u))^{2p}\,(r+\log(P))^{2p}}{\rho_1 \log^{\rho_2} d}\right) + \tilde{O}\left(\frac{1}{m}+\frac{1}{M}+\frac{1}{n}\right).$$

By the same argument in (F.17), (C.5) holds for MI as well. ■

F.6. Main Result

Theorem 46 (Restatement of Theorems 4 and 5) *Under the parameter choice given in Corollary 44, for* $\lambda_t = \frac{m}{\rho_1 \log^{\rho_2} d}$, $\eta_t = \frac{1}{\tilde{O}(m)+\lambda}$ *and* $T = \tilde{O}\left(\rho_1 \log^{\rho_2} d\right)$, *Algorithm 2 guarantees that with probability at least* $1-(18+6m)d^{-u}$*, we have*

$$\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(y - \hat{y}(\boldsymbol{x};(\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, b^{(1)}))\right)^2\right] \leq \Delta\mathbb{E}[\epsilon^2] + \tilde{O}\left(\frac{1}{m}+\frac{1}{M}+\sqrt{\frac{M\log\left(\frac{35d}{M}\right)}{n}}\right)$$
$$+\begin{cases} O\left(\frac{1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{\rho_1\log^{\rho_2} d}\right) & \textit{SI} \\ O\left(\frac{r+\log(4nd^u))^{2p}(r+\log(P))^{2p}}{\rho_1\log^{\rho_2} d}\right) & \textit{MI}\end{cases}$$

where O suppresses constants, and $\tilde{O}$ suppresses constants and Poly $[\log n, \log d]$ *depending on the problem parameters.*

Proof In the following, we assume that (C.1)-(C.5) in Corollary 44 hold. We will prove the statement for SI and will sketch the proof for MI, since the arguments are the same except a few minor steps. Recall that $R_n((\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{a}, \phi(\boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{b})\rangle\right)^2$. We consider

$$\boldsymbol{a}^* := \min_{\boldsymbol{a}\in\mathbb{R}^{2m}} R_n((\boldsymbol{a}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) + \lambda\frac{\|\boldsymbol{a}\|_2^2}{2} \quad \text{where} \quad \lambda = \frac{m}{\rho_1 \log^{\rho_2} d}. \tag{F.18}$$

We observe that

$$\frac{\lambda\|\boldsymbol{a}^*\|_2^2}{2} \leq R_n((\hat{\boldsymbol{a}}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) + \lambda\frac{\|\hat{\boldsymbol{a}}\|_2^2}{2} \Rightarrow$$
$$\|\boldsymbol{a}^*\|_2^2 \leq \frac{2}{\lambda}R_n((\hat{\boldsymbol{a}}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) + \|\hat{\boldsymbol{a}}\|_2^2 \leq O\left(\frac{(1+\log(4nd^u))^{2p}\,(1+\log(P))^{2p(k^\star-1)}}{m}\right), \tag{F.19}$$

and

$$R_n((\boldsymbol{a}^*, \boldsymbol{W}^{(1)}\boldsymbol{b}^{(1)})) \leq R_n((\hat{\boldsymbol{a}}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) + \lambda\frac{\|\hat{\boldsymbol{a}}\|_2^2}{2} \Rightarrow$$
$$R_n((\boldsymbol{a}^*, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) \leq \Delta\mathbb{E}[\epsilon^2] + O\left(\frac{(1+\log(4nd^u))^{2p}(1+\log(P))^{2p(k^\star-1)}}{\rho_1\log^{\rho_2} d}\right) + \tilde{O}\left(\frac{1}{m}+\frac{1}{\sqrt{n}}+\frac{1}{M}\right) \tag{F.20}$$

Moreover, we observe that

$$\nabla_a^2 R_n((\boldsymbol{a}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) = \lambda \boldsymbol{I}_{2m} + \frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{W}^{(1)}\boldsymbol{x}_i + \boldsymbol{b}^{(1)})\phi(\boldsymbol{W}^{(1)}\boldsymbol{x}_i + \boldsymbol{b}^{(1)})^\top$$

$$\Rightarrow \|\nabla_a^2 R_n((\boldsymbol{a}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\|_2 \le \lambda + \frac{1}{n}\sum_{i=1}^{n}\left\|\phi(\boldsymbol{W}^{(1)}\boldsymbol{x}_i + \boldsymbol{b}^{(1)})\right\|_2^2$$

We have

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\phi(\boldsymbol{W}^{(1)}\boldsymbol{x}_i + \boldsymbol{b}^{(1)})\right\|_2^2 \le \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{W}^{(1)}\boldsymbol{x}_i + \boldsymbol{b}^{(1)}\right\|_2^2$$
$$\le 2\sum_{j=1}^{2m}\|\boldsymbol{W}_{j*}^{(1)}\|_2^2 \left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top|_{\mathcal{J}\times\mathcal{J}}\right\|_2 + 2\sum_{j=1}^{2m}(b_j^{(1)})^2 \overset{(a)}{\le} \tilde{O}(m).$$

where we use (C.1) and (C.3) for (a) .

Therefore, (F.18) is a λ-strongly convex and $\left(\tilde{O}(m)+\lambda\right)$- smooth problem. By using $\eta_t = \frac{1}{\tilde{O}(m)+\lambda}$, we can approximate to a^* by $\frac{1}{nm}$ in $T = \tilde{O}(\rho_1 \log^{\rho_2} d)\log(nm) = \tilde{O}(\rho_1 \log^{\rho_2} d)$ iteration of gradient descent, i.e., $\|a^{(T)} - a^*\|_2^2 \le \frac{1}{nm}$ (Bubeck, 2015, Theorem 3.10). We have

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(y - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2\right] \\
&\le \mathbb{E}_{(\boldsymbol{x},y)}\left[\left(y - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{*}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2\right] \\
&+ 2\mathbb{E}\left[\left(y - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{*}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2\right]^{\frac{1}{2}}\mathbb{E}_{\boldsymbol{x}}\left[\left(\hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{*}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2\right]^{\frac{1}{2}} \\
&+ \mathbb{E}_{\boldsymbol{x}}\left[\left(\hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{*}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2\right].
\end{aligned} \tag{F.21}$$

For the last term,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}\left[\left(\hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{*}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})) - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{(T)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2\right] \\
&\qquad\le \|\boldsymbol{a}^* - \boldsymbol{a}^{(T)}\|_2^2 \mathbb{E}_{\boldsymbol{x}}\left[\left\|\phi(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})\right\|_2^2\right] \\
&\qquad\le \|\boldsymbol{a}^* - \boldsymbol{a}^{(T)}\|_2^2 \sum_{j=1}^{2m}\|\boldsymbol{W}_{j*}^{(1)}\|_2^2 + (b_j^{(1)})^2 \\
&\qquad\le \tilde{O}\left(1/n\right).
\end{aligned}$$

For the first term, for $C > 0$ and the event $E_C \equiv \left|\sigma^*(\boldsymbol{V}^\top\boldsymbol{x}) - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^*, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right| > C$, we have

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(y - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{*}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2\right] \\
&\le \mathbb{E}\left[\left(y - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{*}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2 \wedge C^2\right] + \mathbb{E}\left[\left(y - \hat{y}(\boldsymbol{x}; (\boldsymbol{a}^{*}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}))\right)^2 \mathbb{1}_{E_C}\right].
\end{aligned}$$

Here,

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(y-\hat{y}(\boldsymbol{x};(\boldsymbol{a}^*,\boldsymbol{W}^{(1)},\boldsymbol{b}^{(1)}))\right)^2\mathbb{1}_{E_C}\right]\\
&\leq\left(\mathbb{E}[y^4]^{1/4}+\mathbb{E}[\hat{y}(\boldsymbol{x};(\boldsymbol{a}^*,\boldsymbol{W}^{(1)},\boldsymbol{b}^{(1)}))^4]^{\frac{1}{4}}\right)^2\mathbb{P}_{\boldsymbol{x}}\Big[|\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})-\hat{y}(\boldsymbol{x};(\boldsymbol{a}^*,\boldsymbol{W}^{(1)},\boldsymbol{b}^{(1)}))|>C\Big]^{\frac{1}{2}}\\
&\leq\tilde{O}(1)\mathbb{P}_{\boldsymbol{x}}\left[|\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})-\hat{y}(\boldsymbol{x};(\boldsymbol{a}^*,\boldsymbol{W}^{(1)},\boldsymbol{b}^{(1)}))|>C\right]^{\frac{1}{2}}, \qquad \text{(F.22)}
\end{aligned}$$

where we use Lemma 75, and $\|\boldsymbol{a}^*\|_2^2\leq\tilde{O}(1/m)$, $\|\boldsymbol{b}^{(1)}\|_2^2\leq 4m$, and $\|\boldsymbol{W}_{j*}^{(1)}\|_2\leq\tilde{O}(1)$ in the last line. By choosing

$$C:=\|\boldsymbol{a}^*\|_2\sqrt{\|\boldsymbol{b}^{(1)}\|_2^2+\|\boldsymbol{W}^{(1)}\|_F^2}+(\|\boldsymbol{a}^*\|_2\|\boldsymbol{W}^{(1)}\|_F)\sqrt{2\log(4n)}+3C_1(2e\log 6n)^{C_2}\leq\tilde{O}\left(1\right),$$

by Lemma 76, we have (F.22) $\leq\tilde{O}\left(1/\sqrt{n}\right)$. On the other hand, by (F.19) and (F.20), we have with probability at least $1-d^{-u}$,

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(y-\hat{y}(\boldsymbol{x};(\boldsymbol{a}^*,\boldsymbol{W}^{(1)}\boldsymbol{b}^{(1)}))\right)^2\wedge C^2\right]\\
&\leq\Delta\mathbb{E}[\epsilon^2]+O\left(\frac{(1+\log(4nd^u))^{2p}\,(1+\log(P))^{2p(k^\star-1)}}{\rho_1\log^{\rho_2}d}\right)+\tilde{O}\left(\frac{1}{m}+\frac{1}{M}+\sqrt{\frac{M\log\left(\frac{6d}{M}\right)}{n}}\right). \qquad \text{(F.23)}
\end{aligned}$$

By (F.21)-(F.23), the statement follows for SI.

For MI, we observe that the setting is identical except that here we have $\hat{\boldsymbol{\mu}}|_{\mathcal{J}}$. By observing that $\|\hat{\boldsymbol{\mu}}|_{\mathcal{J}}\|_2\leq\tilde{O}(1)$ (by (C.2) in Corollary 44), we can adjust the steps between (F.21)-(F.23) to prove the statement for MI. ■

Appendix G. Lower bounds for CSQ methods

Correlational Statistical Query (CSQ) algorithms are a family learners that can access data using queries $h:\mathbb{R}^d\to\mathbb{R}$ with $\mathbb{E}_{\boldsymbol{x}}[h(\boldsymbol{x})^2]\leq 1$ and returns $\mathbb{E}_{(\boldsymbol{x},y)}[h(\boldsymbol{x})y]$ within an error margin τ. In our setting, since $y=\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})+\sqrt{\Delta}\epsilon$, where ϵ is independent zero-mean noise, the query returns a value in $\mathbb{E}_{\boldsymbol{x}}[h(\boldsymbol{x})\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})]+[-\tau,+\tau]$. An instance of a CSQ algorithm is gradient descent on the population square loss with added noise in the gradients. In this part, we give a lower bound on the CSQ complexity of learning a function in

$$\mathcal{F}_{r,k}:=\left\{\boldsymbol{x}\to\frac{1}{\sqrt{rk!}}\sum_{j=1}^r H_{e_k}(\langle\boldsymbol{V}_{*j},\boldsymbol{x}\rangle)\mid\boldsymbol{V}\in\mathbb{R}^{d\times r},\ \boldsymbol{V}^\top\boldsymbol{V}=\boldsymbol{I}_r,\ \|\boldsymbol{V}\|_{2,q}^q\leq r^{\frac{q}{2}}d^{\alpha\left(1-\frac{q}{2}\right)}\right\},$$

when $\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)$. Here, H_{e_k} denotes the kth Hermite polynomial (see Definition 1), and we use the convention $\|\boldsymbol{V}\|_{2,0}^0:=\|\boldsymbol{V}\|_{2,0}$.

For notational convenience, in the following, "d is large enough" means that $d\geq d^*(r,q,\alpha,k)$, where $d^*(r,q,\alpha,k)$ is a constant depending on the problem parameters (r,q,α,k). Without loss of generality, we can assume all d^*'s are the same since if not, we can take their maximum. We will use $\gtrsim$, $\lesssim$, and $\Omega(\cdot)$, to suppress constants depending on (r, q, α, k) in inequalities and lower bounds. We will use $\widetilde{O}(\cdot)$ to suppress the aforementioned constants and the logarithmic terms in d in upper bounds.[4] The main theorem of this section is as follows:

Theorem 47 (Restatement of Theorem 2) *Consider $\mathcal{F}_{r,k}$ with some $q \in [0, 2)$ and $\alpha \in (0, 1)$. If d is large enough, any CSQ algorithm for $\mathcal{F}_{r,k}$ that guarantees error $\varepsilon = \Omega(1)$ requires either queries of accuracy, i.e., $\tau = \widetilde{O}\left(d^{\left(\alpha\wedge\frac{1}{2}\right)\frac{-k}{2}}\right)$ or super-polynomially many queries in d.*

To prove our lower bound, we will use the argument in (Damian et al., 2022, Lemma 2), for which we need to create a large family of functions with a small average correlation. With the following lemma, we construct such a function class.

Lemma 48 *Let $q \in [0, 2)$, $\alpha \in (0, 1)$, $r \in \mathbb{N}$. When d is large enough, for any $c, k \geq 1$, we can find a set of orthonormal matrices $\mathcal{V} \subseteq \mathbb{R}^{d\times r}$ such that*

– $|\mathcal{V}| \gtrsim \exp\left(\Omega(d^\alpha)\right) \wedge c^r d^k$,

– $\max_{\boldsymbol{V}\in\mathcal{V}} \|\boldsymbol{V}\|_{2,q}^q \leq r^{\frac{q}{2}} d^{\alpha\left(1-\frac{q}{2}\right)}$,

– $\max_{\substack{\boldsymbol{V}^{(1)},\boldsymbol{V}^{(2)}\in\mathcal{V}\\ \boldsymbol{V}^{(1)}\neq\boldsymbol{V}^{(2)}}} \frac{1}{r}\sum_{i,j=1}^r \left|\left\langle \boldsymbol{V}_{*i}^{(1)}, \boldsymbol{V}_{*j}^{(2)}\right\rangle\right|^k \lesssim \frac{\log^k(cd^k)}{d^{k\left(\alpha\wedge\frac{1}{2}\right)}}$.

Proof Let $\tilde{d} = \left\lfloor \frac{d}{r} \right\rfloor$ and $s = \left\lfloor \frac{2d^\alpha}{3^{\frac{2}{2-q}} r} \right\rfloor$. When d is large enough, $\frac{\tilde{d}}{2} \geq s \geq 64$. Hence, by Corollary 55, we can find a set $\mathcal{U} \subseteq S^{\tilde{d}-1}$ such that

– $|\mathcal{U}| \geq \frac{1}{3}\min\{e^{\frac{s}{16}}, cr^k\tilde{d}^k\} \geq \frac{1}{6}\min\left\{\exp\left[\frac{d^\alpha/16}{3^{\frac{2}{2-q}} r}\right], cd^k\right\}$, where the second inequality holds when d is large enough.

– $\max_{\boldsymbol{x}\in\mathcal{U}} \|\boldsymbol{x}\|_q^q \leq \frac{r^{\frac{q}{2}} d^{\alpha\left(1-\frac{q}{2}\right)}}{r}$,

– $\max_{\substack{\boldsymbol{x},\boldsymbol{y}\in\mathcal{U}\\ \boldsymbol{x}\neq\boldsymbol{y}}} |\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq 8Ce\frac{\log(cr^k\tilde{d}^k)}{\min\{\sqrt{\tilde{d}}, s\}} \leq 16Ce3^{\frac{2}{2-q}} r\frac{\log(cd^k)}{\min\left\{d^{1/2}, d^\alpha\right\}}$, where the second inequality holds when d is large enough.

Hence, we can partition $\mathcal{U}$ into r equally sized mutually exclusive sets, and for using a vector from each set, we can form a set of orthonormal matrices $\mathcal{V} \subset \mathbb{R}^{d\times r}$ such that

– $|\mathcal{V}| \geq \frac{1}{(6r)^r}\min\left\{\exp\left[\frac{d^\alpha/16}{3^{\frac{2}{2-q}}}\right], c^r d^{rk}\right\}$.

– $\max_{\boldsymbol{V}\in\mathcal{V}} \|\boldsymbol{V}\|_{2,q}^q \leq r^{\frac{q}{2}} d^{\alpha\left(1-\frac{q}{2}\right)}$,

– $\max_{\substack{\boldsymbol{V}^{(1)},\boldsymbol{V}^{(2)}\in\mathcal{V}\\ \boldsymbol{V}^{(1)}\neq\boldsymbol{V}^{(2)}}} \frac{1}{r}\sum_{i,j=1}^r \left|\left\langle \boldsymbol{V}_{*i}^{(1)}, \boldsymbol{V}_{*j}^{(2)}\right\rangle\right|^k \leq \frac{(16rCe)^k 3^{\frac{2k}{2-q}} \log^k(cd^k)}{\min\left\{d^{k/2}, d^{\alpha k}\right\}}$.

■

4. Here, one might be concerned by the possibility of trivial bounds when $q = 0$. Although, our notation does not exclude such problematic cases, we will use our notation for the sake of readability as such problematic cases do not appear in our proof.

PROOF OF THEOREM 47

Proof [Proof of Theorem 47] Let Q represents the number of queries. We consider polynomial queries, i.e., $Q \leq d^C$ for some $C \in \mathbb{N}$. Let $h_{e_k} := \frac{1}{\sqrt{k!}} H_{e_k}$ be the normalized kth Hermite polynomial. By Lemma 48, we can construct the following function class which is a subset of $\mathcal{F}_{r,k}$:

$$\mathcal{F}_q := \left\{ \frac{1}{\sqrt{r}} \sum_{j=1}^{r} h_{e_k}(\langle \boldsymbol{V}_{*j}, \boldsymbol{x} \rangle) \;\middle|\; V \in \mathcal{V} \right\} \text{ and } \boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d),$$

where $\|\boldsymbol{V}\|_{2,q}^q \leq r^{\frac{q}{2}} d^{\alpha(1-\frac{q}{2})}$, for $\alpha \in (0,1)$, $|\mathcal{V}| \geq \Omega\left(\exp\left(\Omega(d^\alpha)\right) \wedge d^C d^k\right)$, where we used $c = d^C$. We observe that for any different $f, \tilde{f} \in \mathcal{V}$, we have

$$\mathbb{E}[f(\boldsymbol{x})^2] = 1 \text{ and } \mathbb{E}[f(\boldsymbol{x})\tilde{f}(\boldsymbol{x})] \leq \varepsilon \lesssim \frac{\log^k(d)}{d^{k\left(\alpha \wedge \frac{1}{2}\right)}}$$

Therefore, by (Damian et al., 2022, Lemma 2), to get a population loss $\mathbb{E}[(f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2] \leq 2 - 2\varepsilon$

$$\tau^2 \lesssim \frac{d^C}{\exp\left(\Omega(d^\alpha)\right) \wedge d^{Cr} d^k} + \frac{\log^k(d)}{d^{k\left(\alpha \wedge \frac{1}{2}\right)}} \lesssim \frac{\log^k(d)}{d^{k\left(\alpha \wedge \frac{1}{2}\right)}} \tag{G.1}$$

where we use $d^{Cr+k} \leq \exp(\Omega(d^\alpha))$ for d is large enough in the first line. We observe that for d large enough, $\varepsilon \leq 1$. By taking the square root of both sides in (G.1), we obtain the statement. ■

G.1. Lemmas for Lower Bounds

G.1.1. PRELIMINARIES

In this section, we will use Rosenthal-Buckholder inequality and Chernoff-Hoeffding bound given as follows.

Lemma 49 ((Pinelis, 1994, Theorem 5.2) (and see (Damian et al., 2023, Lemma 22))) *Let $\{Y_i\}_{i=0}^n$ be a martingale with martingale difference sequence $\{X_i\}_{i=1}^n$ where $X_i = Y_i - Y_{i-1}$. Let*

$$\langle Y_n \rangle = \sum_{i=1}^{n} \mathbb{E}[|X_i|^2 | \mathcal{F}_{i-1}]$$

denote the predictable quadratic variation. Then, there exists an absolute constant C such that for all $p \geq 2$

$$\|Y_n\|_p \leq C \left[\sqrt{p} \|\langle Y_n \rangle^{1/2}\|_p + p n^{1/p} \max_i \|X_i\|_p \right].$$

Lemma 50 (Chernoff-Hoeffding Bound) *Let $X_1, \cdots, X_n \sim_{iid} Ber(p)$, where $p \in (0, \frac{1}{2}]$ We have*

$$\mathbb{P}\left[\left| \frac{1}{n} \sum_{i=1}^{n} (X_i - p) \right| \geq \frac{p}{2} \right] \leq 2 \exp\left(\frac{-pn}{16} \right).$$

G.1.2. LEMMAS FOR LOWER BOUNDS

For the following, we define a probability distribution P_s, parametrized by $s \in [d]$, as follows: For $\boldsymbol{x} := (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_d)^\top$,

$$\boldsymbol{x} \sim P_s \text{ if } \boldsymbol{x}_i \sim_{iid} \begin{cases} \frac{1}{\sqrt{s}} & \text{wp } \frac{s}{2d} \\ \frac{-1}{\sqrt{s}} & \text{wp } \frac{s}{2d} \\ 0 & \text{wp } 1 - \frac{s}{d} \end{cases}, \quad \text{for } i = 1, \cdots, d.$$

Lemma 51 *Let $\boldsymbol{x}, \boldsymbol{y} \sim_{iid} P_s$. For $s \in [d]$ and $p \geq 2$, we have*

$$\mathbb{P}\left[|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \geq Ce\left(\sqrt{\frac{p}{d}} + \frac{p}{\sqrt{d}}\left(\frac{s^2}{d}\right)^{\frac{1}{p}-\frac{1}{2}}\right)\right] \leq e^{-p}. \tag{G.2}$$

Proof For any $i \in [d]$, note that $\mathbb{E}[\boldsymbol{x}_i] = 0$ and $\mathbb{E}\left[|\boldsymbol{x}_i|^p\right] = \frac{s}{d} s^{-p/2}$,. Therefore, by independence, we have $\mathbb{E}\left[|\boldsymbol{x}_i \boldsymbol{y}_i|^p\right] = s^{2-p}/d^2$. By following the notation in Lemma 49, we let $Y_0 := 0$ and $Y_d := \sum_{i=1}^d \boldsymbol{x}_i \boldsymbol{y}_i$, where $X_i = Y_i - Y_{i-1} = \boldsymbol{x}_i \boldsymbol{y}_i$. We have $\|X_i\|_p = \mathbb{E}\left[|\boldsymbol{x}_i \boldsymbol{y}_i|^p\right]^{1/p} = s^{2/p-1} d^{-2/p}$, and by the independence of $\boldsymbol{x}$ and y, $\langle Y_d \rangle = 1/d$. Hence, by Lemma 49, for $p \geq 2$,

$$\|Y_d\|_p \leq C\left[\sqrt{\frac{p}{d}} + \frac{p}{\sqrt{d}}\left(\frac{s^2}{d}\right)^{\frac{1}{p}-\frac{1}{2}}\right].$$

The statement follows by Markov's inequality. ■

Corollary 52 *By Lemma 51, for $s \in [d]$ and $p \geq 2$, we have*

$$\mathbb{P}\left[|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \geq 2Ce\frac{p}{\min\{\sqrt{d}, s\}}\right] \leq e^{-p}.$$

Proof The statement immediately follows from (G.2). ■

Lemma 53 *Let $\boldsymbol{x} \sim P_s$. For $d \geq 2s$, we have $\mathbb{P}\left[\left|\|\boldsymbol{x}\|_0 - s\right| \geq \frac{s}{2}\right] \leq 2e^{\frac{-s}{16}}$.*

Proof Note that $\mathbb{1}_{\boldsymbol{x}_i \neq 0} \sim Ber(\frac{s}{d})$ and $\|\boldsymbol{x}\|_0 = \sum_{i=1}^d \mathbb{1}_{\boldsymbol{x}_i \neq 0}$. Since $d \geq 2s$, by using Lemma 50, we have

$$\mathbb{P}\left[\left|\frac{1}{d}\sum_{i=1}^d \left(\mathbb{1}_{\boldsymbol{x}_i \neq 0} - \frac{s}{d}\right)\right| \geq \frac{s}{2d}\right] \leq 2e^{\frac{-s}{16}},$$

which is equivalent to the statement. ■

Lemma 54 *Fix any $q \in [0,2)$. For any $s \le \frac{d}{2}$, let $\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(n)} \sim_{iid} P_s$. For any $c, k \ge 1$, we let*

$$\varepsilon := 8Ce\frac{\log(cd^k)}{\min\{\sqrt{d}, s\}}.$$

For $s \ge 5$, we have

$$\mathbb{P}\left[\max_{i\in[n]}\left\|\frac{\boldsymbol{x}^{(i)}}{\|\boldsymbol{x}^{(i)}\|_2}\right\|_q^q \le 3\left(\tfrac{s}{2}\right)^{\frac{2-q}{2}} \text{ AND } \max_{\substack{i,j\in[n]\\ i\neq j}}\left|\left\langle\frac{\boldsymbol{x}^{(i)}}{\|\boldsymbol{x}^{(i)}\|_2}, \frac{\boldsymbol{x}^{(j)}}{\|\boldsymbol{x}^{(j)}\|_2}\right\rangle\right| \le \varepsilon\right] \ge 1 - 2ne^{\frac{-s}{16}} - \frac{n^2}{c^2 d^{2k}}.$$

Proof We observe that

$$\max_{i\in[n]} |\|\boldsymbol{x}^{(i)}\|_0 - s| \le \frac{s}{2} \text{ AND } \max_{\substack{i,j\in[n]\\ i\neq j}}\left|\left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right\rangle\right| \le \frac{\varepsilon}{2} \tag{G.3}$$

$$\Rightarrow \max_{i\in[n]} |\|\boldsymbol{x}^{(i)}\|_0 - s| \le \frac{s}{2} \text{ AND } \max_{\substack{i,j\in[n]\\ i\neq j}}\left|\left\langle\frac{\boldsymbol{x}^{(i)}}{\|\boldsymbol{x}^{(i)}\|_2}, \frac{\boldsymbol{x}^{(j)}}{\|\boldsymbol{x}^{(j)}\|_2}\right\rangle\right| \le \varepsilon$$

$$\Rightarrow \max_{i\in[n]}\left\|\frac{\boldsymbol{x}^{(i)}}{\|\boldsymbol{x}^{(i)}\|_2}\right\|_q^q \le 2^{\frac{q}{2}-1} 3 s^{\frac{2-q}{2}} \text{ AND } \max_{\substack{i,j\in[n]\\ i\neq j}}\left|\left\langle\frac{\boldsymbol{x}^{(i)}}{\|\boldsymbol{x}^{(i)}\|_2}, \frac{\boldsymbol{x}^{(j)}}{\|\boldsymbol{x}^{(j)}\|_2}\right\rangle\right| \le \varepsilon$$

where the second line holds since $\|\boldsymbol{x}^{(i)}\|_0 \ge s/2$ implies $\|\boldsymbol{x}^{(i)}\|_2^2 \ge 1/2$ and the last statement holds since $3s/2 \ge \|\boldsymbol{x}^{(i)}\|_0 \ge s/2$ implies $\|\boldsymbol{x}^{(i)}\|_2 \ge 1/\sqrt{2}$ and $\|\boldsymbol{x}^{(i)}\|_q^q \le \frac{3}{2}s^{\frac{2-q}{2}}$. In the following, we will lower bound (G.3). Since $d \ge 2s$, by Lemma 53, we have

$$\mathbb{P}\left[\max_{i\in[n]}|\|\boldsymbol{x}^{(i)}\|_0 - s| > \frac{s}{2}\right] \le \sum_{i\in[n]}\mathbb{P}\left[|\|\boldsymbol{x}^{(i)}\|_0 - s| \ge \frac{s}{2}\right] \le 2n\exp\left(\frac{-s}{16}\right). \tag{G.4}$$

Moreover, for any $i \neq j \in [n]$,

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right\rangle\right| \ge \frac{\varepsilon}{2}\right] = \mathbb{P}\left[\left|\left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right\rangle\right| \ge 4Ce\frac{\log(cd^k)}{\min\{\sqrt{d}, s\}}\right] \le \frac{1}{c^2 d^{2k}}.$$

where the last step follows Corollary 52, since for $s \ge 5$, we have $d \ge 10$ and $\log(cd^k) \ge 2$ for $c, k \ge 1$. Therefore,

$$\mathbb{P}\left[\max_{\substack{i,j\in[n]\\ i\neq j}}\left|\left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right\rangle\right| > \frac{\varepsilon}{2}\right] \le \frac{n^2}{c^2 d^{2k}}. \tag{G.5}$$

By lower bounding (G.3) with (G.4) and (G.5), we obtain the result. ■

Corollary 55 *For any $q \in [0,2)$ and $64 \le s \le \frac{d}{2}$ and $k, c \ge 1$, there exists a set $\mathcal{U} \subseteq S^{d-1}$ such that*

– $|\mathcal{U}| \geq \frac{1}{3}\min\{e^{\frac{s}{16}}, cd^k\}$,

– $\max_{\boldsymbol{x}\in\mathcal{U}}\|\boldsymbol{x}\|_q^q \leq 3\left(\frac{s}{2}\right)^{\frac{2-q}{2}}$,

– $\max_{\substack{x,y\in\mathcal{U}\\ \boldsymbol{x}\neq\boldsymbol{y}}}|\langle\boldsymbol{x},\boldsymbol{y}\rangle| \leq \varepsilon$, *where* ε *is defined in Lemma 54.*

Proof Consider Lemma 54 with $q \in [0, 2)$, $5 \leq s \leq \frac{d}{2}$, $k, c \geq 1$, and $n = \lceil\frac{1}{3}\min\{e^{\frac{s}{16}}, cd^k\}\rceil$. We observe that the probability of the event in Lemma 54 is nonzero. Hence, there exists such $\mathcal{U}$ as a subset of the normalized versions of the support of P_s. ■

Appendix H. Miscellaneous

H.1. Laurent-Massart Lemma and Its Corollaries

Lemma 56 (Laurent-Massart Lemma) *Let X be a chi-square with N degrees of freedom. For any $t > 0$,*

$$\text{(i) } \mathbb{P}\left[X - N \geq 2\sqrt{Nt} + 2t\right] \leq e^{-t} \text{ and (ii) } \mathbb{P}\left[X - N \leq -2\sqrt{Nt}\right] \leq e^{-t}.$$

Corollary 57 *Let $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. For $d \geq 16\log(1/\delta)$, we have with probability at least $1-\delta$, $\|\boldsymbol{w}\|_2^2 \geq \frac{d}{2}$.*

Proof By Lemma 56, with probability at least $1-\delta$, for $d \geq 16\log(1/\delta)$, $\|\boldsymbol{w}\|_2^2 = \sum_{i=1}^d \boldsymbol{w}_i^2 \geq d - 2\sqrt{d\log(1/\delta)} \geq \frac{d}{2}$. ■

Corollary 58 *For $r \leq d_1 \wedge d_2$, let $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$ be a rank-r matrix. For $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_{d_2})$, we have*

$$\mathbb{P}\left[\|\boldsymbol{A}\boldsymbol{w}\|_2^2 \geq 3\|\boldsymbol{A}\|_2^2(r + \log(1/\delta))\right] \leq \delta.$$

Proof Since $\boldsymbol{A}$ is rank-r, by using SVD, we can write that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{L}^\top$ where $\boldsymbol{U} \in \mathbb{R}^{d_1\times r}$ and $\boldsymbol{L} \in \mathbb{R}^{d_2\times r}$ are orthonormal, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ is diagonal. For $\tilde{\boldsymbol{w}} := \boldsymbol{L}^\top\boldsymbol{w}$, we have $\|\boldsymbol{A}\boldsymbol{x}\|_2^2 =^d \|\boldsymbol{\Sigma}\tilde{\boldsymbol{w}}\|_2^2 \leq \|\boldsymbol{A}\|_2^2\|\tilde{\boldsymbol{w}}\|_2^2$. By using Lemma 56, we have with probability at least $1-\delta$, $\|\boldsymbol{A}\|_2^2\|\tilde{\boldsymbol{w}}\|_2^2 \leq \|\boldsymbol{A}\|_2^2(r + 2\sqrt{r\log(1/\delta)} + 2\log(1/\delta))$. By observing that $\left(r + 2\sqrt{r\log(3/\delta)} + 2\log(3/\delta)\right) \leq 3(r + \log(3/\delta))$, we prove the statement. ■

Lemma 59 *Suppose we have $\{c_1, \cdots, c_r\} \subset \mathbb{R}$ and an orthonormal $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r\} \subset \mathbb{R}^d$. For $k \in \mathbb{N}$ and $\delta \in (0, 1]$, if $\max_{i\in[n]}\|\boldsymbol{V}^\top\boldsymbol{x}_i\|_2 \leq C_{\mathcal{D}}$ and $M \geq 16\log(2/\delta)$ hold, then*

$$\mathbb{P}_{\boldsymbol{w}}\left[\max_{i\in[n]}\left|\sum_{l=1}^r c_l\left\langle\boldsymbol{v}_l, \boldsymbol{x}_i\right\rangle\left\langle\boldsymbol{v}_l, \boldsymbol{w}_{\mathcal{J}}\right\rangle^{k-1}\right| > C_{\mathcal{D}}\max_{l\leq r}|c_l|\left(\frac{6\left(r + \log(2/\delta)\right)}{M}\right)^{\frac{k-1}{2}}\;\middle|\;\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n\right] \leq \delta.$$

Proof By assumption, we have

$$\max_{i\in[n]}\left|\sum_{l=1}^{r} c_l \langle \boldsymbol{v}_l, \boldsymbol{x}_i\rangle \langle \boldsymbol{v}_l, \boldsymbol{w}_{\mathcal{J}}\rangle^{k-1}\right| \overset{(a)}{\leq} C_{\mathcal{D}} \max_{l\leq r}|c_l| \left(\sum_{l=1}^{r} \langle \boldsymbol{v}_l, \boldsymbol{w}_{\mathcal{J}}\rangle^2\right)^{\frac{(k-1)}{2}}. \tag{H.1}$$

where (a) follows that $\|\boldsymbol{v}\|_p \geq \|\boldsymbol{v}\|_q$ for $1 \leq p \leq q \leq \infty$. On the other hand, by Corollaries 57 and 58, we have with probability at least $1-\delta$,

$$\sum_{l=1}^{r} \langle \boldsymbol{v}_l, \boldsymbol{w}_{\mathcal{J}}\rangle^2 = \sum_{l=1}^{r} \frac{\langle \boldsymbol{v}_l|_{\mathcal{J}}, \boldsymbol{w}\rangle^2}{\|\boldsymbol{w}|_{\mathcal{J}}\|_2^2} \leq \frac{3(r+\log(2/\delta))}{M/2} = \frac{6(r+\log(2/\delta))}{M}. \tag{H.2}$$

■

Lemma 60 *We have for $\delta \in (0,1]$ and $M \geq 16\log(2/\delta)$,*

$$\mathbb{P}_{\boldsymbol{w}}\left[\left(\sum_{l=1}^{r} c_l^2 \langle \boldsymbol{v}_l, \boldsymbol{w}_{\mathcal{J}}\rangle^{2(k-1)}\right)^{\frac{1}{2}} > 6^{\frac{k-1}{2}} \max_{l\leq r}|c_l| \left(\frac{r+\log(2/\delta)}{M}\right)^{\frac{k-1}{2}}\right] \leq \delta.$$

Proof We have $\left(\sum_{l=1}^{r} c_l^2 \langle \boldsymbol{v}_l, \boldsymbol{w}_{\mathcal{J}}\rangle^{2(k-1)}\right)^{\frac{1}{2}} \leq \max_{l\leq r}|c_l| \left(\sum_{l=1}^{r} \langle \boldsymbol{v}_l, \boldsymbol{w}_{\mathcal{J}}\rangle^{2(k-1)}\right)^{\frac{1}{2}}$. The statement follows the argument in (H.1) and (H.2). ■

Lemma 61 *Let $\mathcal{A} \subset \mathbb{R}^{d_1\times d}$ such that for any $\boldsymbol{A} \in \mathcal{A}$, $\|\boldsymbol{A}\|_2 \leq 1$ and $rank(\boldsymbol{A}) \leq r$. For $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \sim_{iid} \mathcal{N}(0, \boldsymbol{I}_d)$, we have with probability $1-\delta$,*

$$\sup_{\boldsymbol{A}\in\mathcal{A}}\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{A}^\top - \boldsymbol{A}\boldsymbol{A}^\top\right\|_2 \leq \sqrt{\frac{r}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}} + \sqrt{\frac{2\log|\mathcal{A}|}{n}}$$

Proof Let's fix a $\boldsymbol{A} \in \mathcal{A}$. By SVD, we can write $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{L}^\top$, where $\boldsymbol{U}, \boldsymbol{L} \in \mathbb{R}^{d\times r}$ are orthonormal and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ is diagonal. For $\tilde{\boldsymbol{x}}_i := \boldsymbol{L}^\top\boldsymbol{x}_i$, since $\|\boldsymbol{A}\|_2 = 1$, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{A}^\top - \boldsymbol{A}\boldsymbol{A}^\top\right\|_2 \leq \left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\top - \boldsymbol{I}_r\right\|_2.$$

By (Vershynin, 2010, Corollary 5.35), for a fixed $\mathcal{J} \in \mathcal{H}$, we have with probability at least $1-\delta$, $\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\top - \boldsymbol{I}_r\right\|_2 \leq \sqrt{\frac{r}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}$. By union bound and that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$, the statement follows. ■

H.2. Lemmas for Bounding Polynomials of Gaussian Random Vectors

Lemma 62 (Moments of Gaussian Vector) *For $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, we have $\mathbb{E}[\|\boldsymbol{x}\|_2^{2k}] = d(d+2)\cdots(d+2k-2)$. For $d \geq 2k$, we have $\mathbb{E}[\|\boldsymbol{x}\|_2^{2k}]^{-1} \geq 2^{-k}d^{-k}$.*

Lemma 63 (Hypercontractivity) *Let $P_k : \mathbb{R}^d \to \mathbb{R}$ be a polynomial of degree-k. For $q \geq 2$, we have $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[P_k(\boldsymbol{x})^q\right]^{1/q} \leq (q-1)^{k/2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[P_k(\boldsymbol{x})^2\right]^{1/2}$.*

In the following, we will state some consequences of Lemmas 62 and 63.

Corollary 64 *For $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_r)$ and $p \geq 2$, $\mathbb{E}[(1+\|\boldsymbol{z}\|_2^2)^p]^{\frac{1}{p}} \leq (p-1)(r+2)$.*

Proof By Lemma 62 and 63, $\mathbb{E}[(1+\|\boldsymbol{z}\|_2^2)^p]^{\frac{1}{p}} \leq (p-1)\mathbb{E}[(1+\|\boldsymbol{z}\|^2)^2]^{\frac{1}{2}} \leq (p-1)(r+2)$. ■

Proposition 65 *For $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_r)$ and $C > 0$, $\mathbb{P}\left[(1+\|\boldsymbol{z}\|_2^2)^C \geq u^C(r+2)^C\right] \leq \exp\left(\frac{-u}{e}\right)$, for $u \geq 2e$.*

Proof By Corollary 64, we have for $p \geq 2$ that $\mathbb{P}\left[(1+\|\boldsymbol{z}\|_2^2)^C \geq u^C(r+2)^C\right] \leq p^p u^{-p}$. By using $p^* = \frac{u}{e}$ and $u \geq 2e$, we have the statement. ■

Corollary 66 *By Proposition 65, $\mathbb{P}_{\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I}_r)}\left[|\sigma^*(\boldsymbol{z})| \geq C_1 u^{C_2}(r+2)^{C_2}\right] \leq \exp\left(\frac{-u}{e}\right)$, for $u \geq 2e$.*

Proposition 67 *We have for $u \geq 2e$, $\mathbb{P}\left[|y| \geq C_1(r+2)^{C_2}u^{C_2} + \sqrt{\Delta/e}u^{\frac{1}{2}}\right] \leq 3\exp\left(\frac{-u}{e}\right)$.*

Proof By $|y| \leq |\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})| + \sqrt{\Delta}|\epsilon|$, Corollary 66, $\mathbb{P}\left[|\epsilon| > t\right] \leq 2e^{-t^2}$, the statement follows. ■

Proposition 68 *For $R = C_1(r+2)^{C_2}u^{C_2} + \sqrt{\Delta/e}u^{\frac{1}{2}}$ and $u \geq 2e$, we have*

$$\sup_{\substack{\boldsymbol{w},\boldsymbol{v}\in S^{d-1}\\ b\in\mathbb{R}}} \left|\mathbb{E}\left[y\mathbb{1}_{|y|>R}\left\langle \boldsymbol{v}, \boldsymbol{x}\right\rangle \phi'(\left\langle \boldsymbol{w}, \boldsymbol{x}\right\rangle + b)\right]\right| \leq 6^{\frac{3}{4}}\exp\left(\frac{-u}{2e}\right)\left(C_1^4(4C_2)^{4C_2}(r+2)^{4C_2} + 2\Delta^2\right)^{\frac{1}{4}}.$$

Proof Choose arbitrary $\boldsymbol{w}, \boldsymbol{v} \in S^{d-1}$ and $b \in \mathbb{R}$. By using Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left|\mathbb{E}\left[y\mathbb{1}_{|y|>R}\left\langle u, \boldsymbol{x}\right\rangle \phi'(\left\langle \boldsymbol{w}, \boldsymbol{x}\right\rangle + b)\right]\right| &\leq \mathbb{P}\left[|y| \geq R\right]^{\frac{1}{2}}\mathbb{E}[y^4]^{\frac{1}{4}}\mathbb{E}[|\left\langle u, \boldsymbol{x}\right\rangle \phi'(\left\langle \boldsymbol{w}, \boldsymbol{x}\right\rangle + b)|^4]^{\frac{1}{4}} \\ &\leq 3^{\frac{3}{4}}\exp\left(\frac{-u}{2e}\right)\mathbb{E}[y^4]^{\frac{1}{4}}, \end{aligned} \tag{H.3}$$

where we use $|\phi'| \leq 1$ and Proposition 67 in (H.3). We observe that

$$\begin{aligned}\mathbb{E}[y^4] &\leq 2^3(\mathbb{E}\left[(\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})^4\right] + \Delta^2\mathbb{E}\left[\epsilon^4\right]) \\ &\overset{(a)}{\leq} 2^3(\mathbb{E}\left[(\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})^4\right] + 2\Delta^2) \overset{(b)}{\leq} 2^3\left(C_1^4(4C_2)^{4C_2}(r+2)^{4C_2} + 2\Delta^2\right). \end{aligned} \tag{H.4}$$

where (a) follows from the tail inequality for ϵ, and (b) follows from Corollary 64 since $C_2 \geq 1/2$. By using (H.4) in (H.3) , we have the statement. ■

H.3. Magnitude Pruning

Lemma 69 *For $\boldsymbol{u} \in \mathbb{R}^d$, let $\mathcal{I}_u$ denotes the index set that includes the largest M entries of u and let $\boldsymbol{u}|_{top(M)}$ denote the vector $\boldsymbol{u}$ with everything except M largest coefficients set 0. For any $\boldsymbol{v} \in \mathbb{R}^d$ and $q \in (0, 2]$, we have*

$$(4^{(q-1)\vee 0} + 1) \sum_{i \in \mathcal{I}_u \cup \mathcal{I}_v} |\boldsymbol{u}_i - \boldsymbol{v}_i|^q \geq \|\boldsymbol{u}|_{top(M)} - \boldsymbol{v}\|_q^q - 4^{(q-1)\vee 0}\|\boldsymbol{v} - \boldsymbol{v}|_{top(M)}\|_q^q.$$

Proof Without loss of generality, we can assume $|\boldsymbol{v}_1| \geq |\boldsymbol{v}_2| \geq |\boldsymbol{v}_3| \cdots \geq |\boldsymbol{v}_d|$. We have

$$\|\boldsymbol{u}|_{\text{top}(M)} - \boldsymbol{v}\|_q^q = \sum_{i \in \mathcal{I}_u \cap [M]} |\boldsymbol{u}_i - \boldsymbol{v}_i|^q + \sum_{i \in \mathcal{I}_u - [M]} |\boldsymbol{u}_i - \boldsymbol{v}_i|^q + \sum_{i \in [M] - \mathcal{I}_u} |\boldsymbol{v}_i|^q + \sum_{i \in [d] - (\mathcal{I}_u \cup [M])} |\boldsymbol{v}_i|^q. \quad \text{(H.5)}$$

If $\mathcal{I}_u = [M]$, the statement follows by Proposition 72. Therefore, suppose $\mathcal{I}_u \neq [M]$. Let $[M] - \mathcal{I}_u := \{j_1, \cdots, j_\kappa\}$ and $\mathcal{I}_u - [M] := \{l_1, \cdots, l_\kappa\}$. For some $\iota = 1, \cdots, \kappa$, we get

$$\begin{aligned} |\boldsymbol{v}_{j_\iota}|^q = |\boldsymbol{v}_{j_\iota} \pm \boldsymbol{u}_{j_\iota}|^q &\overset{(a)}{\leq} 2^{(q-1)\vee 0}|\boldsymbol{v}_{j_\iota} - \boldsymbol{u}_{j_\iota}|^q + 2^{(q-1)\vee 0}|\boldsymbol{u}_{l_\iota}|^q \\ &\overset{(b)}{\leq} 2^{(q-1)\vee 0}|\boldsymbol{v}_{j_\iota} - \boldsymbol{u}_{j_\iota}|^q + 4^{(q-1)\vee 0}|\boldsymbol{v}_{l_\iota} - \boldsymbol{u}_{l_\iota}|^q + 4^{(q-1)\vee 0}|\boldsymbol{v}_{l_\iota}|^q, \quad \text{(H.6)} \end{aligned}$$

where in (a), we use Proposition 72 and $|\boldsymbol{u}_{j_\iota}| \leq |\boldsymbol{u}_{l_\iota}|$, $j_\iota \in \mathcal{I}_u$, and Proposition 72 for (b). By using (H.6) for $\iota = 1, \cdots, \kappa$, we get

$$\begin{aligned} \text{(H.5)} &\overset{(a)}{\leq} \sum_{i \in \mathcal{I}_u \cap [M]} |\boldsymbol{u}_i - \boldsymbol{v}_i|^q + (4^{(q-1)\vee 0} + 1) \sum_{i \in \mathcal{I}_u - [M]} |\boldsymbol{u}_i - \boldsymbol{v}_i|^q \\ &\qquad + 2^{(q-1)\vee 0} \sum_{i \in [M] - \mathcal{I}_u} |\boldsymbol{u}_i - \boldsymbol{v}_i|^q + 4^{(q-1)\vee 0} \sum_{i \in [d] - [M]} |\boldsymbol{v}_i|^q \\ &\leq (4^{(q-1)\vee 0} + 1) \sum_{i \in \mathcal{I}_u \cup [M]} |\boldsymbol{u}_i - \boldsymbol{v}_i|^q + 4^{(q-1)\vee 0} \sum_{i \in [d] - [M]} |\boldsymbol{v}_i|^q, \end{aligned} \quad \text{(H.7)}$$

where (a) follows $(\mathcal{I}_u - [M]) \cup ([d] - (\mathcal{I}_u \cup [M])) = [d] - [M]$. By (H.7), the statement follows. ■

Lemma 70 *Let $q \in (0, 2)$ and $v \in \mathbb{R}^d$. We have $\left\|\boldsymbol{v} - \boldsymbol{v}|_{top(M)}\right\|_2 \leq \left(\left(1 - \frac{q}{2}\right)^{\frac{2-q}{q}} \frac{q}{2} \right)^{1/2} \|v\|_q M^{\frac{-1}{q} + \frac{1}{2}}$, for $M = 1, 2, \cdots, d$.*

Proof Without loss of generality, we assume $|\boldsymbol{v}_1| \geq |\boldsymbol{v}_2| \geq \cdots \geq |\boldsymbol{v}_d|$. Then, we have

$$\left\|\boldsymbol{v} - \boldsymbol{v}|_{\text{top}(M)}\right\|_2^2 = \sum_{i=M+1}^{d} \boldsymbol{v}_i^2 \leq |\boldsymbol{v}_{M+1}|^{2-q} \sum_{i=M+1}^{d} |\boldsymbol{v}_i|^q. \quad \text{(H.8)}$$

Let $\sum_{i=M+1}^d |\boldsymbol{v}_i|^q = r$ and $\sum_{i=1}^d |\boldsymbol{v}_i|^q = R$. Then, we have

$$\begin{aligned} R - r = \sum_{i=1}^{M} |\boldsymbol{v}_i|^q \geq M|\boldsymbol{v}_{M+1}|^q \ \Rightarrow\ &|\boldsymbol{v}_{M+1}|^{2-q} \leq (R - r)^{\frac{2-q}{q}} M^{-\frac{2-q}{q}} \\ \Rightarrow\ &\text{(H.8)} \leq (R - r)^{\frac{2-q}{q}} r M^{-\frac{2-q}{q}}. \end{aligned}$$

The statement follows from $\max_{r \in [0,R]} (R - r)^{\frac{2-q}{q}} r \leq \left(1 - \frac{q}{2}\right)^{\frac{2-q}{q}} \frac{q}{2} R^{\frac{2}{q}}$. ■

H.4. Elementary Results

Corollary 71 *For any $M \in [d]$ and $\epsilon > 0$, let $\mathcal{N}_M^\epsilon \subseteq S_M^{d-1}$ be the minimal ϵ-cover. We have $|\mathcal{N}_M^\epsilon| \leq \binom{d}{M}\left(1 + \frac{2}{\epsilon}\right)^M$.*

Proof By (Vershynin, 2018, Corollary 4.2.13), we know that the minimal ϵ-cover of the unit sphere, i.e., $\mathcal{N}^\epsilon \subseteq S^{d-1}$, satisfies $|\mathcal{N}^\epsilon| \leq (1 + 2/\epsilon)^d$. Then, by choosing M subsets of S^{d-1} and taking the union of ϵ-covers restricted on the chosen indices, we can construct an ϵ-cover for S_M^{d-1}. Therefore, the statement follows. ■

Proposition 72 *For any $q \in (0, \infty]$, we have $|a + b|^q \leq 2^{(q-1)\vee 0}(|a|^q + |b|^q)$.*

Proof Without loss of generality, let's assume $|b| \geq |a|$. For $q \in (0, 1]$, we have $|a + b|^q \leq (|a| + |b|)^q \leq |a|^q + q|a|^{q-1}|b| \leq |a|^q + |b|^q$, where we use that $x \to x^q$ is concave in the second inequality. For $q > 1$, we have $|a + b|^q \leq (|a| + |b|)^q \leq 2^{q-1}(|a|^q + |b|^q)$ where we use Jensen's inequality in the last step. ■

Lemma 73 *Let $\cosh(t) := \frac{e^t + e^{-t}}{2}$. For $Z \sim \mathcal{N}(0, 1)$, we have*

(i) $\mathbb{E}[\cosh(\lambda Z^2)] \leq \exp\left(4\lambda^2\right),\ |\lambda| \leq \frac{1}{2\sqrt{2}}$ *and* (ii) $\mathbb{E}[\exp(\lambda^2 Z^2)] \leq \exp\left(2\lambda^2\right),\ |\lambda| \leq \frac{1}{2}$.

Proof Since $|\lambda| \leq \frac{1}{2\sqrt{2}}$, we have $\mathbb{E}\left[\exp(\lambda Z^2)\right] = \frac{1}{\sqrt{1-2\lambda}}$ and $\mathbb{E}\left[\exp(-\lambda Z^2)\right] = \frac{1}{\sqrt{1+2\lambda}}$. Therefore,

$$\mathbb{E}[\cosh(\lambda Z^2)] = \frac{1}{2}\left(\frac{\sqrt{1-2\lambda} + \sqrt{1+2\lambda}}{\sqrt{1-4\lambda^2}}\right) \leq \frac{1}{\sqrt{1-4\lambda^2}} \overset{(a)}{\leq} \exp(4\lambda^2)$$

where (a) follows $\frac{1}{1-t} \leq \exp(2t)$ for $|t| \leq 1/2$. The second statement also follows the same argument. ■

H.5. Lemmas for Feature Learning

Proposition 74 *For $m \in \mathbb{N}$, $M \in [d]$ and $(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{u}) \in \mathbb{R}^m \times \mathbb{R}^{d\times m} \times \mathbb{R}^m \times \mathbb{R}^d$, let*

$$\Theta := \Big\{(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{u}) \mid \|\boldsymbol{a}\|_2 \leq \frac{r_a}{\sqrt{m}},\ \|\boldsymbol{b}\|_\infty \leq r_b,\ \|\boldsymbol{u}\|_2 \leq r_u,\ \|\boldsymbol{W}_{j*}\|_2 \leq r_W,\\ \|\boldsymbol{u}\|_0 \leq M,\ \|\boldsymbol{W}_{j*}\|_0 \leq M,\ j \in [m]\Big\}.$$

and for some $\tau > 0$, let $\mathcal{G} := \Big\{(\boldsymbol{x}, y) \to \big(y - \langle \boldsymbol{u}, \boldsymbol{x}\rangle - \big\langle \boldsymbol{a}, \phi(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b})\big\rangle\big)^2 \wedge \tau^2 \mid (\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{u}) \in \Theta\Big\}$ and let $\mathcal{R}(\mathcal{G})$ denote the Rademacher complexity of $\mathcal{G}$. Then, with $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, we have

$$\mathcal{R}(\mathcal{G}) \leq 4\tau C\left((r_a r_W + r_u)\sqrt{\frac{M\log\left(\frac{6d}{M}\right)}{n}} + \frac{r_a r_b}{\sqrt{n}}\right)$$

where n is number of samples and $C > 0$ is a universal constant.

Proof Let $\mathcal{F} := \left\{(\boldsymbol{x}, y) \to \langle \boldsymbol{u}, \boldsymbol{x}\rangle + \left\langle \boldsymbol{a}, \phi(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b})\right\rangle \mid (\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{u}) \in \Theta\right\}$. By Talagrand's contraction principle, we have $\mathcal{R}(\mathcal{G}) \leq 2\tau \mathcal{R}(\mathcal{F})$. Hence, in the following, we will bound $\mathcal{R}(\mathcal{F})$. Indeed, let $(\varepsilon_i)_{i\in[n]}$ be a sequence of i.i.d Radamacher random variables. Then, we have

$$\begin{aligned}
\mathcal{R}(\mathcal{F}) &= \mathbb{E}\left[\sup_{(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b},\boldsymbol{u})} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle + \left\langle \boldsymbol{a}, \phi(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{b})\right\rangle\right)\right] \\
&\leq \mathbb{E}\left[\sup_{(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left\langle \boldsymbol{a}, \phi(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{b})\right\rangle\right] + \mathbb{E}\left[\sup_{\boldsymbol{u}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \boldsymbol{u}, \boldsymbol{x}_i\rangle\right] \\
&\leq \mathbb{E}\left[\sup_{(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left\langle \boldsymbol{a}, \phi(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{b})\right\rangle\right] + C r_u \sqrt{\frac{M \log\left(\frac{6d}{M}\right)}{n}} \qquad \text{(H.9)}
\end{aligned}$$

where we use (Vershynin, 2018, Exercise 10.3.8) in the last line. To bound the first term, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left\langle \boldsymbol{a}, \phi(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{b})\right\rangle\right] &\leq \frac{r_a}{\sqrt{m}} \mathbb{E}\left[\sup_{(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})} \left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \phi(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{b})\right\|_2\right] \\
&\leq r_a \mathbb{E}\left[\sup_{(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})} \left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \phi(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{b})\right\|_\infty\right] \\
&\leq 2 r_a \mathbb{E}\Big[\sup_{\substack{\|\boldsymbol{w}\|_2 \leq r_W \\ \|\boldsymbol{w}\|_0 \leq M \\ |b| \leq r_b}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i (\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b)\right|\Big] \qquad \text{(H.10)}
\end{aligned}$$

where we use Cauchy Schwartz inequality in the first line, and the contraction lemma in the last line (note that $\phi(0) = 0$ and it is 1-Lipschitz). Then, since the set we take supremum over is symmetric, we have

$$\begin{aligned}
\text{(H.10)} &= 2 r_a \mathbb{E}\Big[\sup_{\substack{\|\boldsymbol{w}\|_2 \leq r_W \\ \|\boldsymbol{w}\|_0 \leq M \\ |b| \leq r_b}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i (\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b)\Big] \\
&\leq 2 r_a r_W \mathbb{E}\Big[\sup_{\substack{\|\boldsymbol{w}\|_2 \leq 1 \\ \|\boldsymbol{w}\|_0 \leq M}} \left\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{x}_i \right\rangle\Big] + 2 r_a r_b \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] \\
&\leq 2 C r_a r_W \sqrt{\frac{M \log\left(\frac{6d}{M}\right)}{n}} + 2 r_a r_b \frac{1}{\sqrt{n}} \qquad \text{(H.11)}
\end{aligned}$$

where we use (Vershynin, 2018, Exercise 10.3.8) in the last line. By (H.9) and (H.11), the statement follows. ■

Lemma 75 *For fixed* $(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b}) \in \mathbb{R}^m \times \mathbb{R}^{d\times m} \times \mathbb{R}^m$, *let* $\hat{y}(\boldsymbol{x}; (\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b})) := \boldsymbol{a}^\top \phi(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b})$. *For* $x \sim \mathcal{N}(0, \boldsymbol{I}_d)$, *we have the following:*

1. $\mathbb{E}_{\boldsymbol{x}}[\hat{y}(\boldsymbol{x}; (\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{b}))^2] \leq \|\boldsymbol{a}\|_2^2 \left(\|\boldsymbol{b}\|_2^2 + \|\boldsymbol{W}\|_F^2\right)$

2. $\mathbb{E}_{\boldsymbol{x}}[\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b}))^4] \leq \|\boldsymbol{a}\|_2^4 m \sum_{j=1}^m \left(3\|\boldsymbol{W}_{j*}\|_2^4 + 6\|\boldsymbol{W}_{j*}\|_2^2\boldsymbol{b}_j^2 + \boldsymbol{b}_j^4\right)$

Proof For the first item, by using Cauchy Schwartz inequality and that $\phi(t) \leq |t|$, we have

$$\begin{aligned}\mathbb{E}[\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b}))^2] = \mathbb{E}\left[\left\langle \boldsymbol{a}, \phi(\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{b})\right\rangle^2\right] &\leq \|\boldsymbol{a}\|_2^2 \mathbb{E}\left[\|\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{b}\|_2^2\right] \\ &= \|\boldsymbol{a}\|_2^2\left(\|\boldsymbol{b}\|_2^2 + \|\boldsymbol{W}\|_F^2\right).\end{aligned}$$

For the second item, by using the same arguments,

$$\begin{aligned}\mathbb{E}[\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b}))^2] = \|\boldsymbol{a}\|_2^4\mathbb{E}\left[\|\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{b}\|_2^4\right] &\overset{(a)}{\leq} \|\boldsymbol{a}\|_2^4 m \sum_{j=1}^m \mathbb{E}\left[(\langle \boldsymbol{W}_{j*},\boldsymbol{x}\rangle + \boldsymbol{b}_j)^4\right] \\ &= \|\boldsymbol{a}\|_2^4 m \sum_{j=1}^m \left(3\|\boldsymbol{W}_{j*}\|_2^4 + 6\|\boldsymbol{W}_{j*}\|_2^2\boldsymbol{b}_j^2 + \boldsymbol{b}_j^4\right)\end{aligned}$$

where we use $\|\boldsymbol{v}\|_4 \leq m^{1/4}\|\boldsymbol{v}\|_2$ for $\boldsymbol{v} \in \mathbb{R}^m$ for (a). ■

Lemma 76 *For fixed* $(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b}) \in \mathbb{R}^m \times \mathbb{R}^{d\times m} \times \mathbb{R}^m$, *and* $\boldsymbol{u} \in \mathbb{R}^d$, *let* $\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})) := \boldsymbol{a}^\top\phi(\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{b}) + \boldsymbol{u}^\top\boldsymbol{x}$. *For* $\boldsymbol{x} \sim \mathcal{N}(0,\boldsymbol{I}_d)$, *we have with probability at least* $1-\delta$,

$$\begin{aligned}|\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})) - \sigma^*(\boldsymbol{V}^\top\boldsymbol{x})| \leq \|\boldsymbol{a}\|_2\sqrt{\|\boldsymbol{b}\|_2^2 + \|\boldsymbol{W}\|_F^2} + (\|\boldsymbol{a}\|_2\|\boldsymbol{W}\|_F + \|\boldsymbol{u}\|_2)\sqrt{2\log(4/\delta)} \\ + C_1(r+2)(2e)^{C_2}\log^{C_2}(6/\delta).\end{aligned}$$

Proof We first observe that

$$\begin{aligned}|\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})) - \sigma^*(\boldsymbol{V}^\top\boldsymbol{x})| &= |\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})) - \mathbb{E}[\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b}))]| \\ &\quad + |\mathbb{E}[\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b}))]| + |\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})| \\ &\leq |\hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b})) - \mathbb{E}[\hat{y}(x;(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b}))]| + |\sigma^*(\boldsymbol{V}^\top\boldsymbol{x})| \\ &\quad + \|\boldsymbol{a}\|_2\left(\|\boldsymbol{b}\|_2^2 + \|\boldsymbol{W}\|_F^2\right)^{1/2}.\end{aligned}$$

Moreover, since ϕ is 1-Lipschitz that $\boldsymbol{x} \to \hat{y}(\boldsymbol{x};(\boldsymbol{a},\boldsymbol{W},\boldsymbol{b}))$ is $\|\boldsymbol{a}\|_2\|\boldsymbol{W}\|_F + \|\boldsymbol{u}\|_2$ - Lipschitz. Then, by using Gaussian Lipschitz concentration inequality (see (Vershynin, 2018, Theorem 5.2.2)) and Corollary 66, we obtain the statement. ■