
Pretrained Transformer Efficiently Learns Low-Dimensional Target Functions In-Context

Kazusato Oko^{1,3}, Yujin Song^{2,3}, Taiji Suzuki^{2,3}, Denny Wu^{4,5}

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

oko@berkeley.edu, song-yujin139@g.ecc.u-tokyo.ac.jp,
taiji@mist.i.u-tokyo.ac.jp, dennywu@nyu.edu

Abstract

Transformers can efficiently learn in-context from example demonstrations. Most existing theoretical analyses studied the in-context learning (ICL) ability of transformers for linear function classes, where it is typically shown that the minimizer of the pretraining loss implements one gradient descent step on the least squares objective. However, this simplified linear setting arguably does not demonstrate the statistical efficiency of ICL, since the pretrained transformer does not outperform directly solving linear regression on the test prompt. In this paper, we study ICL of a nonlinear function class via transformer with nonlinear MLP layer: given a class of *single-index* target functions $f_*(\mathbf{x}) = \sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)$, where the index features $\boldsymbol{\beta} \in \mathbb{R}^d$ are drawn from a r -dimensional subspace, we show that a nonlinear transformer optimized by gradient descent (with a pretraining sample complexity that depends on the *information exponent* of the link functions σ_*) learns f_* in-context with a prompt length that only depends on the dimension of the distribution of target functions r ; in contrast, any algorithm that directly learns f_* on test prompt yields a statistical complexity that scales with the ambient dimension d . Our result highlights the adaptivity of the pretrained transformer to low-dimensional structures of the function class, which enables sample-efficient ICL that outperforms estimators that only have access to the in-context data.

1 Introduction

Pretrained transformers [VSP⁺17] possess the remarkable ability of *in-context learning (ICL)* [BMR⁺20], whereby the model constructs a predictor from a prompt sequence consisting of pairs of labeled examples without updating any parameters. A common explanation is that the trained transformer can implement a learning algorithm, such as gradient descent on the in-context examples, in its forward pass [HvMM⁺19, DSD⁺22, VONR⁺23]. Such explanation has been empirically studied in synthetic tasks such as linear regression [GTLV22, ASA⁺22], which has motivated theoretical analyses of the statistical and computational complexity of ICL in learning simple function classes.

Many recent theoretical works on ICL focus on learning the function class of *linear models* using linear transformers trained via empirical risk minimization. In this setting, it can be shown that minima of the pretraining loss implements one (preconditioned) gradient descent step on the least squares objective computed on the test prompt [ZFB23, ACDS23, MHM23]. This implies that the forward pass of the pretrained transformer can learn linear targets with $n \gtrsim d$ in-context examples, hence matching the sample complexity of linear regression on the test prompt. Subsequent works also studied how the distribution and “diversity” of pretrained tasks affect the ICL solution in similar problem settings [WZC⁺23, RPCG23, ZWB24, LLZV⁺24].

The motivation of our work is the observation that the simple setting of learning linear models with linear transformers does not fully capture the statistical efficiency and adaptivity of ICL. Specifically,

- A linear transformer has limited expressivity and hence only implements simple operations in the forward pass, such as one gradient step for least squares regression. This limits the class of algorithms that can be executed in-context, and consequently, the pretrained transformer cannot outperform directly solving linear regression on the test prompt. We therefore ask the question:

With the aid of MLP layer, can a pretrained transformer learn a nonlinear function class in-context, and outperform baseline algorithms that only have access to the test prompt?

- A key feature of ICL is the *adaptivity* to structure of the learning problem. For example, [GTLV22] empirically showed that transformers can match the performance of either ridge regression or LASSO, depending on parameter sparsity of the target class; [RPCG23] observed that transformers transitions from a weighted-sum estimator to ridge regression as the number of pretraining tasks increases. Such adaptivity enables the pretrained model to outperform algorithms that only have access to the test prompt, which cannot take into account the “prior” distribution of target functions. Hence a natural question to ask is

Can a pretrained transformer adapt to certain structures of the target function class, and how does such adaptivity contribute to the statistical efficiency of ICL?

1.1 Our Contributions

1.1.1 Function Class: Gaussian Single-index Models

To address the above questions, we study the in-context learning of arguably the simplest *nonlinear* extension of linear regression where a nonlinearity is applied to the response. Specifically, we consider the following class of *single-index* target functions, where the t -th pretraining task is constructed as

$$\mathbf{x}_1^t, \mathbf{x}_2^t, \dots, \mathbf{x}_N^t, \mathbf{x}^t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d), \quad y_i^t = \sigma_*^t(\langle \mathbf{x}_i^t, \boldsymbol{\beta}^t \rangle) + \varsigma_i^t, \quad (1.1)$$

where $\sigma_*^t : \mathbb{R} \rightarrow \mathbb{R}$ is the unknown link function, and $\boldsymbol{\beta}^t \in \mathbb{R}^d$ is the index feature vector which is drawn from some fixed r -dimensional subspace for some $r \ll d$. The task is to learn this input-output relation by reading the context $(\mathbf{x}_1, y_1, \dots, \mathbf{x}_N, y_N)$ and predict the output corresponding to the query \mathbf{x} (See Section 2 for details). This problem setting is based on the following considerations.

- **Nonlinearity of function class.** Due to the nonlinear link function, single-index targets cannot be learned by linear transformers. For this class of functions, the statistical efficiency of simple algorithms has been extensively studied: given a link function with degree P and information exponent Q (defined as the index of the smallest non-zero coefficient in the Hermite expansion), we know that kernel methods require at least $n \gtrsim d^P$ samples [GMMM21, DWY21], whereas two-layer neural network trained by gradient descent can achieve better sample complexity $n \gtrsim d^{\Theta(Q)}$ [BAGJ21, BBSS22]. Hence these algorithms, if directly applied to the test prompt, require a context length *polynomial* in the ambient dimensionality d , which arguably deviates from practical settings of ICL where the pretrained transformer learns from a few in-context examples which may come from high dimensional data space. These algorithms serve as a baseline for comparing the statistical efficiency of ICL.

- **Two types of low-dimensional structures.** Prior works on gradient-based feature learning highlights the adaptivity of neural network to *low-dimensional functions*, i.e., single-index model $\sigma_*(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)$ that depends on one direction (index features) $\boldsymbol{\beta}$ in \mathbb{R}^d [AAM22, BES⁺22, BMZ23]. In such setting, the complexity of gradient descent is dominated by the search for direction $\boldsymbol{\beta} \in \mathbb{R}^d$, the difficulty of which can be characterized by various computational lower bounds [DLS22, AAM23, DPVLB24].

Importantly, our pretraining problem setup introduces *another notion of low dimensionality*: the “distribution” of target functions is low-dimensional, since the index features for each task $\boldsymbol{\beta}_t$ are drawn from a rank- r subspace. This low-dimensionality of *function class* cannot be exploited by any algorithms that directly estimate the target function from test prompt, but as we will see, transformers can adapt to this additional structure via gradient-based pretraining, which reduces the search problem (for the index features $\boldsymbol{\beta}$) to r -dimensional in the in-context phase. Therefore, when $r \ll d$, we expect pretrained transformers to outperform baseline algorithms on the in-context examples (kernel methods, neural network, etc.).

Empirical observations. We pretrain a GPT-2 model [RWC⁺19] (with the same configurations as the in-context linear regression setting in [GTLV22]) to learn the Gaussian single-index task (1.1) with degree-3 link function, and compare its in-context sample complexity against baseline algorithms (see Section 4 for details). In Figure 1 we observe that the pretrained transformer achieves low prediction risk using fewer in-context examples than two baseline algorithms: kernel ridge regression, and neural network trained by gradient descent. Moreover, we observe that unlike the baseline methods,

Transformer achieves an in-context sample complexity (almost) independent of the ambient dimensionality d .

The goal of this work is to rigorously establish a d -independent in-context sample complexity in an idealized theoretical setting for a shallow transformer optimized via gradient-based pretraining.

1.1.2 Main Result: Learning Single-index Models In-Context

We characterize the sample complexity of learning (1.1) in-context, using a nonlinear transformer optimized by gradient descent. Each single-index task is specified by an unknown index feature vector $\beta \in \mathbb{R}^d$ drawn from some r -dimensional subspace, and a link function σ_* with degree P and information exponent $Q \leq P$; we allow the degree and information exponent to vary across tasks, to model the scenario where the difficulty of pretraining tasks may differ. We show that pretraining of the nonlinear MLP layer can extract the low-dimensional structure of the function class, and the attention layer efficiently approximates the nonlinear link function. Our main theorem upper bounds the in-context generalization error of the pretrained transformer.

Theorem (Informal). *Let $f : (\mathbf{x}_1, y_1, \dots, \mathbf{x}_N, y_N, \mathbf{x}) \mapsto y$ be a transformer with nonlinear MLP layer pretrained with gradient descent (Algorithm 1) on the single-index regression task (1.1). With probability at least 0.99, the model f achieves in-context prediction risk $\mathbb{E}|f(\mathbf{x}) - f_*(\mathbf{x})| - \tau = o_d(1)$, where τ is the noise level, if the number of pretraining tasks T , the number of training examples N , the test prompt length N^* , and the network width m satisfy (we omit polylogarithmic factors)*

$$\underbrace{T \gtrsim d^{\Theta(Q)}, \quad NT \gg T \gtrsim d^{\Theta(Q)}}_{\text{pretraining}}, \quad \underbrace{N^* \gtrsim r^{\Theta(P)}}_{\text{inference}}, \quad \underbrace{m \gtrsim r^{\Theta(P)}}_{\text{approximation}},$$

where Q, P are the information exponent and the highest degree of link functions, respectively.

To the best of our knowledge, this is the first end-to-end optimization and statistical guarantee for in-context regression of this nonlinear function class. We make the following remarks.

- The required sample size for *pretraining* T, N scale with the ambient dimensionality d . In particular, the number of pretraining tasks T is parallel to the complexity of learning a single-index model with information exponent Q using a two-layer neural network. On the other hand, the sample complexity for the *in-context* phase only depends on the dimensionality of the function class $r \ll d$.
- Note that any estimator that only has access to the in-context examples requires $n \gtrsim d$ samples to learn the single-index model, as suggested by the information theoretic lower bound (e.g., see [MM18, BKM⁺19]). Therefore, when $r \ll d$, we see a separation between pretrained transformer and algorithms that directly learn from the test prompt, such as linear/kernel regression and neural network + gradient descent. This highlights the adaptivity (via pretraining) of transformers to low-dimensional structures of the target function class.
- Our analysis of pretraining reveals the following mechanism analogous to [GHM⁺23]: the nonlinear MLP layer extract useful features and adapt to the low-dimensionality of the function class, whereas the attention layer performs in-context function approximation on top of the learned features.

1.2 Related Works

Theory of in-context learning. Many recent works studied the in-context learning ability of transformers trained by gradient descent. [ZFB23, ACDS23, MHM23, WZC⁺23, ZWB24] studied

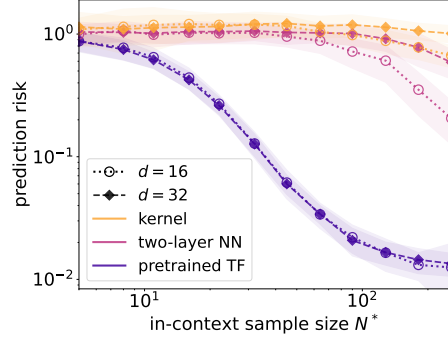


Figure 1: In-context generalization error (with standard deviation) of kernel ridge regression, neural network + gradient descent, and pretrained transformer. The target function is a polynomial single-index model. We fix $r = 8$ and vary $d = 16, 32$.

the training of linear transformer models to learn linear target functions in-context by implementing one gradient descent step in the forward pass. Similar optimization results have been established for looped linear transformers [GSR⁺24], transformers with SoftMax attention [HCL23, NDL24, CSWY24a, CSWY24b] or nonlinear MLP layer [KS24, LWL⁺24]. Our problem setting resembles [KS24], where a nonlinear MLP block is used to extract features, followed by a linear attention layer; the main difference is that we establish end-to-end guarantees on the optimization and sample complexity for learning a concrete nonlinear function class, whereas [KS24] focused on convergence of optimization. [CCS23] showed that transformers can learn nonlinear functions in-context via a functional gradient mechanism, but no statistical and optimization guarantees were given. If we do not take gradient-based optimization into account, the function class that can be implemented in-context by transformers has been characterized in many prior works [BCW⁺23, GHM⁺23, ZZYW23, JLLVR24, SGS⁺24, KNS24]. These results typically aim to encode specific algorithms (LASSO, gradient descent, etc.) in the forward pass or directly analyze the Bayes-optimal estimator.

Gradient-based learning of low-dimensional functions. The complexity of learning low-dimensional functions with neural network has been extensively studied in the feature learning theory literature. Typical target functions include single-index models [BAGJ21, BES⁺22, BBSS22, MHPG⁺23, DNGL23, BES⁺23] and multi-index models [DLS22, AAM22, AAM23, BBPV23]. While a shallow neural network can efficiently approximate such low-dimensional functions, the efficiency of gradient-based training is governed by properties of the nonlinearity σ^* . In the single-index setting, prior works established a sufficient sample size $n \gtrsim d^{\Theta(K)}$, where $K \in \mathbb{N}$ is the *information exponent* for algorithms utilizing correlational information [BAGJ21, BBSS22, DNGL23, MHWSE23], or the *generative exponent* for algorithms that employ suitable label transformations [DPVLB24, LOSW24, ADK⁺24, JMS24]. Moreover, $n \asymp d$ samples are information theoretically necessary without additional structural assumptions; this entails that estimators that only access the test prompt inevitably pay a sample size that scales with the ambient dimensionality. As we will see, pretrained transformers can avoid this “curse of dimensionality” by exploiting the low-dimensionality of the task distribution. Low-dimensional structure of the *function class* similar to our setting has been assumed to study the efficiency of transfer learning [DLS22] and multi-task learning [CHS⁺23].

2 Problem Setting

Notations. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and the $\ell_2 \rightarrow \ell_2$ operator norm for matrices. For a vector \mathbf{w} , we use $\mathbf{w}_{a:b}$ for $a \leq b$ to denote the vector $[w_a, w_{a+1}, \dots, w_b]^\top$. $\mathbf{1}_N$ denotes the all-one vector of size N . The indicator function of A is denoted by \mathbb{I}_A . Let N be a nonnegative integer; then $[N]$ denotes the set $\{n \in \mathbb{Z} \mid 1 \leq n \leq N\}$. For a nonnegative integer i , the i -th Hermite polynomial is defined as $\text{He}_i(z) = (-1)^i e^{\frac{z^2}{2}} \frac{d^i}{dz^i} e^{-\frac{z^2}{2}}$. For a set S , $\text{Unif}(S)$ denotes the uniform distribution over S . We denote the unit sphere $\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| = 1\}$ by \mathbb{S}^{d-1} . $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ represent $O(\cdot)$ and $\Omega(\cdot)$ notations where polylogarithmic terms are hidden. We write $a \lesssim b$ when there exists a constant c such that $a \leq cb$. If both $a \lesssim b$ and $b \lesssim a$ holds, we write $a \asymp b$.

2.1 Data Generating Process

2.1.1 In-context learning

We first introduce the basic setting of ICL [BMR⁺20] of simple function classes as investigated in [GTLV22, ASA⁺22]. In each *task*, the learner is given a sequence of inputs and outputs $(\mathbf{x}_1, y_1, \dots, \mathbf{x}_N, y_N, \mathbf{x})$ referred to as *prompt*, where $\mathbf{x}_i, \mathbf{x} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The labeled examples $\mathbf{X} = (\mathbf{x}_1 \ \dots \ \mathbf{x}_N) \in \mathbb{R}^{d \times N}$, $\mathbf{y} = (y_1 \ \dots \ y_N)^\top \in \mathbb{R}^N$ are called *context*, and \mathbf{x} is the *query*. Given input distribution $\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{x} \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_{\mathbf{x}}$, the output y_i is expressed as

$$y_i = f_*(\mathbf{x}_i) + \varsigma_i, \quad i \in [N],$$

where f_* is the true function describing the input-output relation and $\varsigma_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_{\varsigma}$ is i.i.d. label noise. Note that f_* also varies across tasks — we assume f_* is drawn i.i.d. from some true distribution \mathcal{D}_{f_*} . In the pretraining phase, we optimize the model parameters given training data from T distinct tasks $\{(\mathbf{x}_1^t, y_1^t, \dots, \mathbf{x}_M^t, y_M^t, \mathbf{x}^t, y^t)\}_{t=1}^T$, which is composed of prompts $\{(\mathbf{x}_1^t, y_1^t, \dots, \mathbf{x}_M^t, y_M^t, \mathbf{x}^t)\}_{t=1}^T$ and responses $\{y^t\}_{t=1}^T$ for queries $\{\mathbf{x}^t\}_{t=1}^T$, where $y^t = f_*^t(\mathbf{x}^t) + \varsigma^t$ and $\varsigma^t \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_{\varsigma}$.

We say a model learns these functional relations *in-context*, when the model can predict the output $f_*(\mathbf{x})$ corresponding to query \mathbf{x} by solely examining the context (\mathbf{X}, \mathbf{y}) , without updating model parameters for each task. Given the pretrained model $f(\mathbf{X}, \mathbf{y}, \mathbf{x}; \boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$ which predicts the label of query \mathbf{x} from context (\mathbf{X}, \mathbf{y}) , we define the *expected ICL risk* as

$$\mathcal{R}_{N^*}(f) := \mathbb{E}[|f(\mathbf{X}_{1:N^*}, \mathbf{y}_{1:N^*}, \mathbf{x}; \boldsymbol{\theta}) - y|], \quad (2.1)$$

where $y = f_*(\mathbf{x}) + \varsigma$ and the expectation is taken over the in-context data: $\mathbf{x}_1, \dots, \mathbf{x}_{N^*}, \mathbf{x} \sim \mathcal{D}_{\mathbf{x}}, f_* \sim \mathcal{D}_{f_*}, \varsigma_1, \dots, \varsigma_{N^*}, \varsigma \sim \mathcal{D}_{\varsigma}$. Note that we take the expectation with respect to contexts $(\mathbf{X}_{1:N^*}, \mathbf{y}_{1:N^*}) \in \mathbb{R}^{d \times N^*} \times \mathbb{R}^{N^*}$ of length N^* , in order to examine the behavior of ICL at a specific context length.

2.1.2 Gaussian single-index models

We consider the situation where the true input-output relation is expressed by *single-index models*, i.e., functions that only depend on the direction of the index vector $\boldsymbol{\beta}$ in the input space. An efficient learning algorithm should adapt to this feature and identify the relevant subspace from high-dimensional observations; hence this problem setting has been extensively studied in the deep learning theory literature [BL19, BES⁺22, BBSS22, MHPG⁺23, MZD⁺23] to demonstrate the adaptivity of gradient-based feature learning.

Assumption 1. Let $\tau \geq 0$ be the noise level. The prompt $(\mathbf{x}_1, y_1, \dots, \mathbf{x}_N, y_N, \mathbf{x})$ is generated as

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N, \mathbf{x} \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_{\mathbf{x}} = \mathcal{N}(0, \mathbf{I}_d), \quad y_i = f_*(\mathbf{x}_i) + \varsigma_i, \quad f_*(\mathbf{x}_i) = \sigma_*(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle),$$

where $\varsigma_1, \dots, \varsigma_N \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{-\tau, \tau\})$, and the distribution \mathcal{D}_{f_*} of the true function f_* is specified as:

1. **Index Features.** Let \mathcal{S} be an $r \leq d$ -dimensional linear subspace of \mathbb{R}^d . We draw $\boldsymbol{\beta}$ uniformly from the unit sphere $\mathbb{S}(\mathcal{S})$ in \mathcal{S} , i.e., from $\mathbb{S}(\mathcal{S}) := \{\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathcal{S}, \|\boldsymbol{\beta}\| = 1\}$.
2. **Link Function.** $\sigma_*(z) = \sum_{i=Q}^P \frac{c_i}{i!} \text{He}_i(z)$, where $2 \leq Q \leq P$. We draw the Hermite coefficients $\{c_i\}_{i=Q}^P$ from any distribution satisfying

$$\mathbb{E}[c_Q^2] = \Theta_{d,r}(1) \neq 0, \quad \sum_{i=Q}^P c_i^2 \leq R_c^2 \text{ (a.s.) and } (c_Q, \dots, c_P) \neq (0, \dots, 0) \text{ (a.s.)}. \quad (2.2)$$

Throughout the paper, we assume that $P \ll d, r$ and $r \ll d$; specifically, we take $P = \Theta_{d,r}(1)$ and $r \lesssim d^{1/2}$. Note that the condition $r \ll d$ entails that the class of target functions is *low-dimensional*, and as we will see, such structure can be adapted by the transformer via pretraining.

Remark 1. We make the following remarks on the assumption of single-index function class.

- For each task, the target is a single-index model with different index features drawn from some rank- r subspace, and different link function with degree at most P and information exponent (defined as the index of the lowest degree non-zero coefficient in the Hermite expansion of the link function, i.e., $\min\{i \mid c_i \neq 0\}$ in this case; see [BAGJ21, DH18]) at least Q . This heterogeneity reflects the situation where the difficulty of learning the input-output relation varies across tasks. Note that we allow for different distributions of the Hermite coefficients $\{c_i\}$: for example, we may set $(c_Q, \dots, c_P) \sim \text{Unif}\left\{(c_Q, \dots, c_P) \mid \sum_{i=Q}^P \frac{c_i^2}{i!} = 1\right\}$ (manifold of coefficients satisfying $\mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})] = 1$), $\text{Unif}(\{0, 1\}^{P-Q+1} \setminus (0, \dots, 0))$, or $\text{Unif}(\{(1, \dots, 0), \dots, (0, \dots, 1)\})$.
- The condition $Q \geq 2$ (i.e., the Hermite expansion of σ_* does not contain constant and linear terms) ensures that the gradient update detects the entire r -dimensional subspace instead of the trivial rank-1 component. For generic polynomial σ_* , this assumption can be satisfied by a simple preprocessing step that subtracts the low-degree components, as done in [DLS22].

2.2 Student Model: Transformer with Nonlinear MLP Layer

We consider a transformer composed of a single-layer self-attention module preceded by an embedding module using a nonlinear multi-layer perceptron (MLP). Let $\mathbf{E} \in \mathbb{R}^{d_e \times d_N}$ be an embedding

matrix constructed from prompt $(\mathbf{x}_1, y_1, \dots, \mathbf{x}_N, y_N, \mathbf{x})$. A single-layer SoftMax self-attention module [VSP⁺17] is given as

$$f_{\text{Attn}}(\mathbf{E}; \mathbf{W}^P, \mathbf{W}^V, \mathbf{W}^K, \mathbf{W}^Q) = \mathbf{E} + \mathbf{W}^P \mathbf{W}^V \mathbf{E} \cdot \text{softmax}\left(\frac{(\mathbf{W}^K \mathbf{E})^\top \mathbf{W}^Q \mathbf{E}}{\rho}\right), \quad (2.3)$$

where ρ is the temperature, and $\mathbf{W}^K, \mathbf{W}^Q \in \mathbb{R}^{d_k \times d_e}$, $\mathbf{W}^V \in \mathbb{R}^{d_v \times d_e}$ and $\mathbf{W}^P \in \mathbb{R}^{d_e \times d_v}$ are the key, query, value, and projection matrix, respectively. Following prior theoretical works [ZFB23, ACDS23, MHM23, KNS24], we remove the SoftMax and instead analyze the *linear* attention with $\rho = N$; it has been argued that such simplification can reproduce phenomena in practical transformer training [ACS⁺23]. We further simplify the original self-attention module (2.3) by merging $\mathbf{W}^P \mathbf{W}^V$ as $\mathbf{W}^{PV} \in \mathbb{R}^{d_e \times d_e}$ and $(\mathbf{W}^K)^\top \mathbf{W}^Q$ as $\mathbf{W}^{KQ} \in \mathbb{R}^{d_e \times d_e}$, and consider the following parameterization also introduced in [ZFB23, ACDS23, WZC⁺23],

$$\mathbf{W}^{PV} = \begin{bmatrix} * & * \\ \mathbf{0}_{1 \times (d_e - 1)} & v \end{bmatrix}, \mathbf{W}^{KQ} = \begin{bmatrix} \mathbf{K} & * \\ \mathbf{0}_{1 \times (d_e - 1)} & * \end{bmatrix}, \quad (2.4)$$

where $v \in \mathbb{R}$ and $\mathbf{K} \in \mathbb{R}^{(d_e - 1) \times (d_e - 1)}$. Then, the simplified attention module is written as $\tilde{f}_{\text{Attn}}(\mathbf{E}; \mathbf{W}^{PV}, \mathbf{W}^{KQ}) = \mathbf{E} + \mathbf{W}^{PV} \mathbf{E} \left(\frac{\mathbf{E}^\top \mathbf{W}^{KQ} \mathbf{E}}{N} \right)$, and we take the right-bottom entry as the prediction of y corresponding to query \mathbf{x} .

Prior analyses of linear transformers [ZFB23, ACDS23, MHM23] defined the embedding matrix \mathbf{E} simply as the input-output pairs; however, the combination of linear embedding and linear attention is not sufficient to learn nonlinear single-index models. Instead, we set $d_e = m + 1$, $d_N = N + 1$ for $m \in \mathbb{N}$ and construct \mathbf{E} using an MLP layer:

$$\mathbf{E} = \begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x}_1 + b_1) & \cdots & \sigma(\mathbf{w}_1^\top \mathbf{x}_N + b_1) & \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots & \ddots & \vdots & \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x}_1 + b_m) & \cdots & \sigma(\mathbf{w}_m^\top \mathbf{x}_N + b_m) & \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \\ y_1 & \cdots & y_N & 0 \end{bmatrix} \in \mathbb{R}^{(m+1) \times (N+1)}, \quad (2.5)$$

where $\mathbf{w}_1, \dots, \mathbf{w}_m \in \mathbb{R}^d$ and $b_1, \dots, b_m \in \mathbb{R}$ are trainable parameters and $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is a nonlinear activation function; we use $\sigma(z) = \text{ReLU}(z) = \max\{z, 0\}$. This is to say, we define the embedding as the ‘‘hidden representation’’ $\sigma(\mathbf{w}^\top \mathbf{x} + b)$ of a width- m two-layer neural network. For concise notation we write $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_m)^\top$, $\mathbf{b} = (b_1, \dots, b_m)^\top$.

Remark 2. We make the following remarks on the considered architecture.

- The MLP embedding layer before attention has been adopted in recent works [GHM⁺23, KS24, KNS24]. This setting can be interpreted as an idealized version of the mechanism that lower layers of the transformer construct useful representation, on top of which upper attention layers implement the in-context learning algorithm.
- Our architecture is also inspired by recent theoretical analyses of gradient-based feature learning, where it is shown that gradient descent on the MLP layer yields adaptivity to features of the target function and hence improved statistical efficiency [AAM22, DLS22, BES⁺22].

Combining the attention \tilde{f}_{Attn} and MLP embedding (2.5), we can express the model prediction y as

$$f(\mathbf{X}, \mathbf{y}, \mathbf{x}; \mathbf{W}, \mathbf{\Gamma}, \mathbf{b}) = \left\langle \frac{\mathbf{\Gamma} \sigma(\mathbf{W} \mathbf{X} + \mathbf{b} \mathbf{1}_N^\top) \mathbf{y}}{N}, \begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \end{bmatrix} \right\rangle, \quad (2.6)$$

where $\mathbf{\Gamma} = v \mathbf{K}^\top$ and $\sigma(\mathbf{W} \mathbf{X} + \mathbf{b} \mathbf{1}_N^\top)$ denotes the $m \times N$ matrix whose (i, j) -th entry is $\sigma(\mathbf{w}_i^\top \mathbf{x}_j + b_i)$; see Appendix E for full derivation. We refer to $\mathbf{\Gamma}$ as the *attention matrix*.

2.3 Pretraining: Empirical Risk Minimization via Gradient Descent

We pretrain the transformer (2.6) by the gradient-based learning algorithm specified in Algorithm 1, which is inspired by the layer-wise training procedure studied in the neural network theory literature [AAM22, DLS22, BES⁺22]. In particular, we update the trainable parameters in a sequential manner.

Algorithm 1: Gradient-based training of transformer with MLP layer

Input : Learning rate η_1 , weight decay λ_1, λ_2 , prompt length N_1, N_2 , number of tasks T_1, T_2 , attention matrix initialization scale γ .

- 1 **Initialize** $w_j^{(0)} \sim \text{Unif}(\mathbb{S}^{d-1})$ ($j \in [m]$); $b_j^{(0)} \sim \text{Unif}([-1, 1])$ ($j \in [m]$);
 $\Gamma_{j,j}^{(0)} \sim \text{Unif}(\{\pm\gamma\})$ ($j \in [m]$) and $\Gamma_{i,j}^{(0)} = 0$ ($i \neq j \in [m]$).
- 2 **Stage I: Gradient descent for MLP layer**
- 3 $\left[\text{Draw data } \{(\mathbf{x}_1^t, y_1^t, \dots, \mathbf{x}_{N_1}^t, y_{N_1}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$ with prompt length N_1 .
- 4 $\left[w_j^{(1)} \leftarrow w_j^{(0)} - \eta_1 \left[\nabla_{w_j^{(0)}} \frac{1}{T_1} \sum_{t=1}^{T_1} (y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \Gamma^{(0)}, \mathbf{b}^{(0)}))^2 + \lambda_1 w_j^{(0)} \right]$.
- 5 **Initialize** $b_j \sim \text{Unif}([-\log d, \log d])$.
- 6 **Stage II: Empirical risk minimization for attention layer**
- 7 $\left[\text{Draw data } \{(\mathbf{x}_1^t, y_1^t, \dots, \mathbf{x}_{N_2}^t, y_{N_2}^t, \mathbf{x}^t, y^t)\}_{t=T_1+1}^{T_1+T_2}$ with prompt length N_2 .
- 8 $\left[\Gamma^* \leftarrow \text{argmin}_{\Gamma} \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} (y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \Gamma, \mathbf{b}))^2 + \frac{\lambda_2}{2} \|\Gamma\|_F^2$.

Output : Prediction function $\mathbf{x} \rightarrow f(\mathbf{X}, \mathbf{y}, \mathbf{x}; \mathbf{W}^{(1)}, \Gamma^*, \mathbf{b})$.

- In Stage I we optimize the parameters of the MLP (embedding) layer, which is a non-convex problem due to the nonlinear activation function. To circumvent the nonlinear training dynamics, we follow the recipe in recent theoretical analyses of gradient-based feature learning [DLS22, BES⁺22, BEG⁺22]; specifically, we zoom into the *early phase* of optimization by taking one gradient step on the regularized empirical risk. As we will see, the first gradient step already provides a reliable estimate of the target subspace \mathcal{S} , which enables the subsequent attention layer to implement a sample-efficient in-context learning algorithm.
- In Stage II we train the attention layer, which is a convex problem and the global minimizer can be efficiently found. We show that the optimized attention matrix Γ performs regression on the polynomial basis (defined by the MLP embedding), which can be seen as an in-context counterpart to the second-layer training (to learn the polynomial link) in [DLS22, AAM23, OSSW24].

3 Transformer Learns Single-index Models In-Context

3.1 Main Theorem

Our main theorem characterizes the pretraining and in-context sample complexity of the transformer with MLP layer (2.6) optimized by layer-wise gradient-based pretraining outlined in Algorithm 1.

Theorem 1. *Given Assumption 1 and $r \lesssim d^{\frac{1}{2}}$. We pretrain the transformer (2.6) using Algorithm 1 with $m = \tilde{\Omega}(r^P), T_1 = \tilde{\Omega}(d^{Q+1}r^Q), N_1 T_1 = \tilde{\Omega}(d^{2Q+1}r), \gamma \asymp \frac{1}{m^{\frac{3}{2}} r^{\frac{1}{2}} d^Q}, \eta_1 \asymp m^{\frac{3}{2}} r d^{2Q-\frac{1}{2}}$. $(\log d)^{-C_\eta}$ for constant C_η . Then, for appropriately chosen regularization parameters $\lambda_1, \lambda_2 > 0$, with probability at least 0.99 over the data distribution and random initialization, the ICL prediction risk (2.1) with test prompt length N^* for the output f of Algorithm 1 can be upper bounded as*

$$\mathcal{R}_{N^*}(f) - \tau = \tilde{O} \left(\sqrt{\frac{r^{3P}}{m} + r^{4P} \left(\frac{1}{T_2} + \frac{1}{N_2} + \frac{1}{N^*} \right)} \right).$$

Theorem 1 suggests that to achieve low in-context prediction risk, it is sufficient to set $T_1, N_1 = \tilde{\Omega}(d^{\Theta(Q)})$, and $m, T_2, N_2, N^* = \tilde{\Omega}(r^{\Theta(P)})$. Observe that the *pretraining* complexity T_1, N_1 scales with the ambient dimensionality d , but the *in-context* sample complexity N^* only scales with the dimensionality of the function class $r \ll d$. This illustrates the in-context efficiency of pretrained transformers and aligns with our observations in the GPT-2 experiment reported in Figure 1.

Remark 3. *We make the following remarks.*

- *The sample complexity highlights different roles of the two sources of low dimensionality in our setting. The low dimensionality of single-index f_* entails that the pretraining cost scales as $N \gtrsim d^{\Theta(Q)}$, which is consistent with prior analyses on gradient-based feature learning [BAGJ21, DNGL23]. On the other hand, the low dimensionality of function class (i.e., β lies in r -dimensional*

subspace) leads to an in-context sample complexity that scales as $N^* \gtrsim r^{\Theta(P)}$, which, roughly speaking, is the rate achieved by polynomial regression or kernel models on r -dimensional data.

- The multiplicative scaling between N_1 and T_1 in the sample complexity suggests that one can tradeoff between the two quantities, that is, pretraining on more diverse tasks (larger T_1) can reduce the required pretraining context length N_1 .
- Similar low-dimensional function class has been considered in the setting of transfer or multi-task learning with two-layer neural networks [DLS22, CHS⁺23], where the first-layer weights identify the span of all target functions, and the second-layer approximates the nonlinearity. However, the crucial difference is that we do not update the network parameters based on the in-context examples; instead, the single-index learning is implemented by the forward pass of the transformer.

Comparison against baseline methods. Below we summarize the statistical complexity of commonly-used estimators that only have access to N^* in-context examples, but not the pretraining data. Note that the in-context sample complexity all depends on the ambient dimensionality $d \gg r$.

- **Kernel models.** Recall that our target function is a degree- P polynomial, and hence kernel ridge regression requires $N^* \gtrsim d^P$ in-context examples [GMMM21, DWY21].
- **CSQ learners.** The correlational statistical query (CSQ) lower bound suggests that an algorithm making use of correlational information requires $N^* \gtrsim d^{\Theta(Q)}$ samples to learn a single-index model with information exponent Q [DLS22, AAM22]. This sample complexity can be achieved by online SGD training of shallow neural network [BAGJ21, DNGL23].
- **Information theoretic limit.** Since the single-index model (1.1) contains d unknown parameters, we can infer an information theoretic lower bound of $N^* \gtrsim d$ samples to estimate this function. This complexity can be achieved (up to polylog factors) by tailored SQ algorithms [CM20, DPVLB24] or modified gradient-base training of neural network [LOSW24, ADK⁺24, JMS24].

3.2 Proof Sketch of Main Theorem

We provide a sketch of derivation for Theorem 1. The essential mechanism is outlined as follows: after training the MLP embedding via one gradient descent step, the MLP parameters $\{w_j^{(1)}\}$ align with the common subspace of the target functions \mathcal{S} . Subsequently, the (linear) attention module estimates the input-output relation f_* (which varies across tasks) on this r -dimensional subspace. We explain these two ingredients in the ensuing sections.

3.2.1 Training the MLP Layer

We first show that the first gradient descent step on \mathbf{W} results in significant alignment with the target subspace \mathcal{S} , using a proof strategy pioneered in [DLS22]. Note that under sufficiently small initialization and appropriately chosen weight decay, we have

$$w_j^{(1)} \simeq \eta_1 \cdot \frac{2}{T_1} \sum_{t=1}^{T_1} y^t \nabla_{w_j^{(0)}} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \Gamma^{(0)}, \mathbf{b}^{(0)}).$$

The key observation is that this correlation between y and the gradient of model output contains information of \mathcal{S} . We establish the concentration of this empirical gradient around the expected (population) one, which determines the required rates of T_1 and N_1 . For the population gradient, we make use of the Hermite expansion of $\sigma_{b_j}(z) = \sigma(z + b_j)$ and show that its leading term is proportional to $\begin{bmatrix} w_{1:r}^{(0)} \\ \mathbf{0}_{d-r} \end{bmatrix}$, which is contained in the target subspace \mathcal{S} ; see Appendix B for details.

3.2.2 Attention Matrix with Good Approximation Property

Next we construct the attention matrix $\bar{\Gamma}$ which satisfies the following approximation property:

Proposition 2 (Informal). *There exists $\bar{\Gamma}$ such that with high probability,*

$$\left| f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \bar{\Gamma}, \mathbf{b}) - y^t \right| - \tau = \tilde{O}\left(\sqrt{r^{3P}/m + r^{2P}/N_2}\right)$$

for all $t \in \{T_1 + 1, \dots, T_2\}$. Moreover, we have $\|\bar{\Gamma}\|_F = \tilde{O}(r^{2P}/m)$.

We provide an intuitive explanation of this construction. Recall that the target function can be written in its Hermite expansion $f_*(\mathbf{x}) = \sum_{i=Q}^P \frac{c_i}{i!} \text{He}_i(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)$. We build an orthonormal basis for f_* as follows. Let $\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_r\}$ be an orthonormal basis of \mathcal{S} . Then, any f_* can be expressed as a linear combination of functions in $\mathcal{H} = \{\prod_{j=1}^r \text{He}_{p_j}(\langle \boldsymbol{\beta}_j, \cdot \rangle) \mid Q \leq p_1 + \dots + p_r \leq P, p_1 \geq 0, \dots, p_r \geq 0\}$, where $\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)}[h(\mathbf{x})h'(\mathbf{x})] = \mathbb{1}_{h=h'}$ holds for $h, h' \in \mathcal{H}$. We write $\mathcal{H} = \{h_1, \dots, h_{B_P}\}$ with $B_P = |\mathcal{H}| = \Theta(r^P)$, and observe that a two-layer neural network can be constructed to approximate each h_n . Specifically, there exist vectors $\mathbf{a}^1, \dots, \mathbf{a}^{B_P} \in \mathbb{R}^m$ such that $\sum_{j=1}^m a_j^n \sigma(\langle \mathbf{w}_j^{(1)}, \mathbf{x} \rangle + b_j) \simeq h_n(\mathbf{x})$ for each $n \in [B_P]$.

Consequently, we can build the desired attention matrix using coefficients $\mathbf{a}^1, \dots, \mathbf{a}^{B_P}$. Let $\mathbf{A} = (\mathbf{a}^1 \ \dots \ \mathbf{a}^{B_P}) \in \mathbb{R}^{m \times B_P}$ and $\bar{\boldsymbol{\Gamma}} = \mathbf{A}\mathbf{A}^\top$, the attention module can recover the true function as

$$\begin{aligned} & \left\langle \frac{1}{N_2} \bar{\boldsymbol{\Gamma}} \sigma(\mathbf{W}^{(1)} \mathbf{X}^t + \mathbf{b} \mathbf{1}_{N_2}^\top) \mathbf{y}^t, \sigma(\mathbf{W}^{(1)} \mathbf{x} + \mathbf{b}) \right\rangle \\ &= \sum_{n=1}^{B_P} \left(\frac{1}{N_2} \sum_{i=1}^{N_2} \left(\sum_{j=1}^m \mathbf{a}_j^n \sigma(\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle + b_j) \right) y_i \right) \left(\sum_{j=1}^m \mathbf{a}_j^n \sigma(\langle \mathbf{w}_j^{(1)}, \mathbf{x} \rangle + b_j) \right) \\ &\stackrel{(a)}{\simeq} \sum_{n=1}^{B_P} \left(\frac{1}{N_2} \sum_{i=1}^{N_2} h_n(\mathbf{x}_i) y_i \right) h_n(\mathbf{x}) \stackrel{(b)}{\simeq} \sum_{n=1}^{B_P} \mathbb{E}[h_n(\mathbf{x}) f_*(\mathbf{x})] h_n(\mathbf{x}) = f_*(\mathbf{x}). \end{aligned} \quad (3.1)$$

Roughly speaking, the self-attention architecture computes the correlation between the target function and basis element h_i to estimate the corresponding coefficient in this basis decomposition. We evaluate (a) the approximation error of two-layer neural network in Appendix C.1, and (b) the discrepancy between the empirical and true correlations in Appendix C.2.

Note that the approximation errors and the norm of $\bar{\boldsymbol{\Gamma}}$ at this stage scale only with r up to poly-logarithmic terms; this is because $\mathbf{W}^{(1)}$ already identifies the low-dimensional target subspace \mathcal{S} . Consequently, the in-context sample size N^* only scales with the target dimensionality $r \ll d$.

3.2.3 Generalization Error Analysis

Finally, we transfer the learning guarantee from the constructed $\bar{\boldsymbol{\Gamma}}$ to the (regularized) empirical risk minimization solution $\boldsymbol{\Gamma}^*$. By the equivalence between optimization with L_2 regularization and norm-constrained optimization, there exists λ_2 such that $\|\boldsymbol{\Gamma}^*\|_F \leq \|\bar{\boldsymbol{\Gamma}}\|_F$ and the empirical ICL loss by $\boldsymbol{\Gamma}^*$ is no larger than that of $\bar{\boldsymbol{\Gamma}}$. Hence, we can bound the generalization error by $\boldsymbol{\Gamma}^*$ using a standard Rademacher complexity bound for norm-constrained transformers provided in Appendix D.1. One caveat here is that the context length N^* at test time may differ from training time; hence we establish a context length-free generalization bound, which is discussed in Appendix D.2.

4 Synthetic Experiments

4.1 Experimental Setting

We pretrain a GPT-2 model [RWC⁺19] to learn the Gaussian single-index function class (1.1). Specifically, we consider the 12-layer architecture (with 22.3M parameters) used in [GTLV22] for in-context linear regression. The pretraining data is generated from random single-index models: for each task t , the context $\{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^N$ is generated as $\mathbf{x}_i^t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$ and $y_i^t = \sum_{i=Q}^P \frac{c_i^t}{i!} \text{He}_i(\langle \mathbf{x}_i^t, \boldsymbol{\beta}^t \rangle)$, where $(c_Q^t, \dots, c_P^t) \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{(c_Q, \dots, c_P) \mid \sum_{i=Q}^P \frac{c_i^2}{i!} = 1\}$ and $\boldsymbol{\beta}^t \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{\boldsymbol{\beta} \mid \boldsymbol{\beta} = [\beta_1, \dots, \beta_r, 0, \dots, 0]^\top, \|\boldsymbol{\beta}\| = 1\}$. See Appendix F for further details.

4.2 Empirical Findings

Ambient dimension-free sample complexity. In Figure 2 we examine how the in-context sample complexity of the GPT-2 model depends on the ambient dimensionality d and the function class dimensionality r . For each problem setting the model is pretrained for 100,000 steps using the data of degree $P = 4$ and information exponent $Q = 2$ (see Appendix F for details). In Figure 2(a) we observe that for fixed $r = 8$, varying the ambient dimensionality $d = 16, 32, 64$ leads to negligible change in the model performance for the in-context phase. In contrast, Figure 2(b) illustrates that for fixed d , the required sample size N^* scales with the dimensionality of the function class $r = 2, 4, 8$. This confirms our theoretical finding that transformers can adapt to low-dimensional structure of the distribution of target functions via gradient-based pretraining.

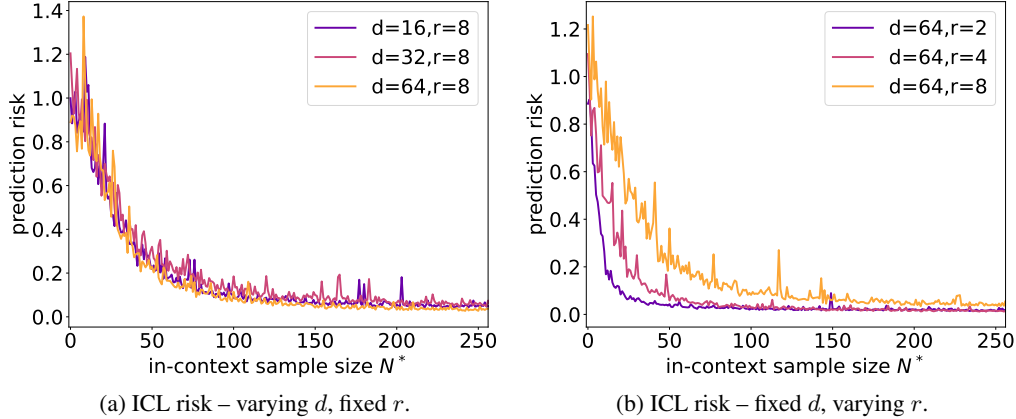


Figure 2: In-context sample complexity of GPT-2 model pretrained on Gaussian single-index function (see Section 4.1 for details) of degree-4 polynomial. Observe that (a) the ICL risk curve overlaps for different ambient dimensions d but the same target (subspace) dimensionality r , and (b) the required sample size N^* becomes larger as r increases.

Superiority over baseline algorithms. In Figure 1 we compare the in-context sample complexity of the GPT-2 model pretrained by data of $Q = 3$ and $P = 2$ against two baseline algorithms that directly learn f_* on the test prompt: (i) kernel ridge regression with the Gaussian RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\sigma^2)$, and (ii) two-layer neural network with ReLU activation $f_{\text{NN}}(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m a_i \sigma(\langle \mathbf{x}, \mathbf{w}_i \rangle)$ trained by the Adam optimizer [KB15]. We observe that for $r = 8, d = 16, 32$, the pretrained transformer outperforms both KRR and two-layer NN; moreover, the performance of these two baseline algorithms deteriorates significantly as the ambient dimensionality d becomes larger.

5 Conclusion and Future Direction

We study the complexity of in-context learning for the Gaussian single-index models using a pre-trained transformer with nonlinear MLP layer. We provide an end-to-end analysis of gradient-based pretraining and establish a generalization error bound that takes into account the number of pretraining tasks, the number of pretraining and in-context examples, and the network width. Our analysis suggests that when the distribution of target functions exhibits low-dimensional structure, transformers can identify and adapt to such structure during pretraining, whereas any algorithm that only has access to the test prompt necessarily requires a larger sample complexity.

We outline a few limitations and possible future directions.

- The in-context sample complexity we derived $r^{\Theta(P)}$ corresponds to that of polynomial regression or kernel methods in r -dimensional space. This rate is natural as discussed in Section 3.2.2, where the linear self-attention module extracts the coefficients with respect to fixed basis functions (of size $r^{\Theta(P)}$). An interesting question is whether transformers can implement a more efficient in-context algorithm that matches the complexity of gradient-based feature learning in r dimensions. This can be achieved if the pretrained model learns features in-context.
- Our pretraining complexity is based on one GD step analysis similar to [DLS22, BES⁺22] which makes use of the correlational information; hence the information exponent of the link functions plays an important role. We conjecture that the sample complexity can be improved if we modify the pretraining procedure or training objective, as done in [DTA⁺24, LOSW24, ADK⁺24, JMS24].

Acknowledgement

KO was partially supported by JST, ACT-X Grant Number JPMJAX23C4. TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2015). This research is unrelated to DW's work at xAI.

References

- [AAM22] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.
- [AAM23] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.
- [ACDS23] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2023.
- [ACS⁺23] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). *arXiv preprint arXiv:2310.01082*, 2023.
- [ADK⁺24] Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.
- [ASA⁺22] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.
- [BAGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.
- [BBPV23] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.
- [BBSS22] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 35, 2022.
- [BCW⁺23] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in Neural Information Processing Systems*, 36, 2023.
- [BEG⁺22] Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35, 2022.
- [BES⁺22] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35, 2022.
- [BES⁺23] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. *Advances in Neural Information Processing Systems*, 36, 2023.
- [BKM⁺19] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.
- [BL19] Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.
- [BMR⁺20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 2020.

- [BMZ23] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.
- [CCS23] Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. *arXiv preprint arXiv:2312.06528*, 2023.
- [CHS⁺23] Liam Collins, Hamed Hassani, Mahdi Soltanolkotabi, Aryan Mokhtari, and Sanjay Shakkottai. Provable multi-task representation learning by two-layer relu neural networks. *arXiv preprint arXiv:2307.06887*, 2023.
- [CM20] Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *Conference on Learning Theory*, pages 1161–1227. PMLR, 2020.
- [CSWY24a] Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.
- [CSWY24b] Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. *arXiv preprint arXiv:2409.10559*, 2024.
- [DH18] Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pages 1887–1930. PMLR, 2018.
- [DLS22] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.
- [DNGL23] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36, 2023.
- [DPVLB24] Alex Damian, Loucas Pillaud-Vivien, Jason D Lee, and Joan Bruna. The computational complexity of learning gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.
- [DSD⁺22] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022.
- [DTA⁺24] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024.
- [DWY21] Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.
- [GHM⁺23] Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do transformers learn in-context beyond simple functions? a case study on learning with representations. *arXiv preprint arXiv:2310.10616*, 2023.
- [GMMM21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.
- [GSR⁺24] Khashayar Gatmiry, Nikunj Saunshi, Sashank J Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *International Conference on Machine Learning*, 2024.
- [GSS21] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α -sub-exponential random variables. *Electronic Journal of Probability*, 26:1 – 22, 2021.

- [GTLV22] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35, 2022.
- [HCL23] Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.
- [HvMM⁺19] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.
- [JLLVR24] Hong Jun Jeon, Jason D Lee, Qi Lei, and Benjamin Van Roy. An information-theoretic analysis of in-context learning. *arXiv preprint arXiv:2401.15530*, 2024.
- [JMS24] Nimit Joshi, Theodor Misiakiewicz, and Nathan Srebro. On the complexity of learning sparse functions with statistical and gradient queries. *arXiv preprint arXiv:2407.05622*, 2024.
- [KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- [KNS24] Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. *arXiv preprint arXiv:2408.12186*, 2024.
- [KS24] Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. *arXiv preprint arXiv:2402.01258*, 2024.
- [LLZV⁺24] Yue M Lu, Mary I Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlivan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint arXiv:2405.11751*, 2024.
- [LOSW24] Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.
- [LWL⁺24] Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *International Conference on Machine Learning*, 2024.
- [Mau16] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.
- [MHM23] Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.
- [MHPG⁺23] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *International Conference on Learning Representations*, 2023.
- [MHWSE23] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36, 2023.
- [MM18] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450. PMLR, 2018.
- [MZD⁺23] Arvind Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. *Advances in Neural Information Processing Systems*, 36, 2023.

- [NDL24] Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.
- [OSSW24] Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. *arXiv preprint arXiv:2406.11828*, 2024.
- [RPCG23] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 36, 2023.
- [RWC⁺19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [SGS⁺24] Michael E Sander, Raja Giryes, Taiji Suzuki, Mathieu Blondel, and Gabriel Peyré. How do transformers perform in-context autoregressive learning? *arXiv preprint arXiv:2402.05787*, 2024.
- [Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.
- [VONR⁺23] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [Wai19] Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.
- [WZC⁺23] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter L Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? *arXiv preprint arXiv:2310.08391*, 2023.
- [ZFB23] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.
- [ZWB24] Ruiqi Zhang, Jingfeng Wu, and Peter L Bartlett. In-context learning of a linear transformer block: Benefits of the mlp component and one-step gd initialization. *arXiv preprint arXiv:2402.14951*, 2024.
- [ZZYW23] Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023.

Table of Contents

1	Introduction	1
1.1	Our Contributions	2
1.2	Related Works	3
2	Problem Setting	4
2.1	Data Generating Process	4
2.2	Student Model: Transformer with Nonlinear MLP Layer	5
2.3	Pretraining: Empirical Risk Minimization via Gradient Descent	6
3	Transformer Learns Single-index Models In-Context	7
3.1	Main Theorem	7
3.2	Proof Sketch of Main Theorem	8
4	Synthetic Experiments	9
4.1	Experimental Setting	9
4.2	Empirical Findings	9
5	Conclusion and Future Direction	10
A	Preliminaries	16
A.1	Definition of High Probability Event	16
A.2	Tensor Notations	16
A.3	Hermite Polynomials	16
A.4	Other Auxiliary Lemmas	18
B	Proofs for the MLP Layer	19
B.1	Calculation of Population Gradient	20
B.2	Concentration of Empirical Gradient	24
C	Construction of Attention Matrix	29
C.1	Approximation Error of Two-layer Neural Network	29
C.2	Concentration of Correlation	34
C.3	Proof of Proposition 2	36
D	Generalization Error Analysis and Proof of Theorem 1	37
D.1	Rademacher Complexity Bound	38
D.2	Prompt Length-free Generalization Bound	40
D.3	Proof of Theorem 1	41
E	Derivation of Simplified Self-attention Module	42
F	Details of Experiments	43
F.1	Detailed Experimental Settings	43
F.2	Experiment on Our Architecture and Algorithm 1	44

A Preliminaries

We consider the high-dimensional setting, i.e., our result holds for all $d \geq D$ where D is a constant which does not depend on d and r . Throughout the proofs, we take $\mathcal{S} = \{(x_1, \dots, x_r, 0, \dots, 0)^\top \mid x_1, \dots, x_r \in \mathbb{R}\}$ and $\beta \sim \text{Unif}(\mathbb{S}(\mathcal{S})) = \text{Unif}(\{(\beta_1, \dots, \beta_r, 0, \dots, 0)^\top \mid \beta_1^2 + \dots + \beta_r^2 = 1\})$. Note that this assumption is without loss of generality by the rotational invariance of the Gaussian distribution.

A.1 Definition of High Probability Event

Definition 3. We say that an event A occurs with high probability when there exists a sufficiently large constant C^* which does not depend on d, r and

$$1 - \Pr[A] \leq O(d^{-C^*})$$

holds.

We implicitly assume that we can redefine the constant C^* to be sufficiently large as needed. The lemma below is a basic example and will be used in the proofs:

Lemma 4. Let $x \sim \mathcal{N}(0, 1)$. Then, $|x| \lesssim \sqrt{\log d}$ holds with high probability.

Proof. From the tail bound of Gaussian, $\mathbb{P}(|x| \geq t) \leq 2\exp(-t^2/2)$ holds. Hence we get $\mathbb{P}(|x| \geq \sqrt{2C^* \log d}) \leq O(d^{-C^*})$. \square

Note that C^* can be redefined by changing the hidden constant in $|x| \lesssim \sqrt{\log d}$.

If A_1, \dots, A_M occurs with high probability where $M = O(\text{poly}(d))$, then $A_1 \cap \dots \cap A_M$ also occurs with high probability (by redefining C^*). In particular, throughout this paper we assume that $m, N_1, N_2, T_1, T_2 = O(\text{poly}(d))$, which allows us to take such unions.

A.2 Tensor Notations

In this paper, a k -tensor is a multidimensional array which has k indices: for example, matrices are 2-tensors. Let \mathbf{A} be a k -tensor. A_{i_1, \dots, i_k} denotes (i_1, \dots, i_k) -th entry of \mathbf{A} . Let \mathbf{A} be a k -tensor and \mathbf{B} be a l -tensor where $k \geq l$. $\mathbf{A}(\mathbf{B})$ denotes a $k-l$ tensor whose (i_1, \dots, i_{k-l}) -th entry is

$$\mathbf{A}(\mathbf{B})_{i_1, \dots, i_{k-l}} = \sum_{j_1, \dots, j_l} A_{i_1, \dots, i_{k-l}, j_1, \dots, j_l} B_{j_1, \dots, j_l},$$

and is defined only when sizes are compatible. If $k = l$, we sometimes write $\mathbf{A}(\mathbf{B})$ as $\mathbf{A} \circ \mathbf{B}$ or $\langle \mathbf{A}, \mathbf{B} \rangle$. Let $\mathbf{v} \in \mathbb{R}^d$ be a vector and k be a positive integer. Then, $\mathbf{v}^{\otimes k} \in \mathbb{R}^{d \times \dots \times d}$ denotes a k -tensor whose (i_1, \dots, i_k) -th entry is $v_{i_1} \dots v_{i_k}$.

Let $f(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ be a d -variable differentiable function. A k -tensor $\nabla^k f(\mathbf{x})$ is defined as

$$(\nabla^k f(\mathbf{x}))_{i_1, \dots, i_k} = \frac{\partial}{\partial x_{i_1}} \dots \frac{\partial}{\partial x_{i_k}} f(\mathbf{x}).$$

The following properties can be verified easily.

Lemma 5. For tensors \mathbf{A} and \mathbf{B} , $\|\mathbf{A}(\mathbf{B})\|_F^2 \leq \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2$ holds.

Lemma 6. For a vector \mathbf{v} , $\|\mathbf{v}^{\otimes k}\|_F^2 = (\|\mathbf{v}\|_2^2)^k$ holds.

A.3 Hermite Polynomials

We frequently use (probablists') Hermite polynomials, which is defined as

$$\text{He}_i(z) = (-1)^i e^{\frac{z^2}{2}} \frac{d^i}{dz^i} e^{-\frac{z^2}{2}},$$

where i is a nonnegative integer. We often make use of the orthogonality property: $\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\text{He}_i(z)\text{He}_j(z)] = i! \delta_{i,j}$. The Hermite expansion for $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is defined as $\sigma(z) = \sum_{i \geq 0} \frac{a_i}{i!} \text{He}_i(z)$ where $a_i = \mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma(z)\text{He}_i(z)]$. Similarly, the multivariate Hermite

expansion for $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined as $f(\mathbf{z}) = \sum_{i_1 \geq 0, \dots, i_d \geq 0} \frac{a_{i_1, \dots, i_d}}{(i_1)! \dots (i_d)!} \text{He}_{i_1}(z_1) \cdots \text{He}_{i_d}(z_d)$, where $a_{i_1, \dots, i_d} = \mathbb{E}_{z_1, \dots, z_d \sim \mathcal{N}(0,1)} [f(\mathbf{z}) \text{He}_{i_1}(z_1) \cdots \text{He}_{i_d}(z_d)]$. The coefficient a_{i_1, \dots, i_d} can also be obtained by $a_{i_1, \dots, i_d} = \mathbb{E}_{z_1, \dots, z_d \sim \mathcal{N}(0,1)} \left[\frac{\partial^{i_1}}{\partial z_1^{i_1}} \cdots \frac{\partial^{i_d}}{\partial z_d^{i_d}} f(\mathbf{z}) \right]$.

Consider f_* in our problem setting (Assumption 1). We can bound $\mathbb{E}[f_*(\mathbf{x})^2]$ and $\mathbb{E}[f_*(\mathbf{x})^4]$ as follows.

Lemma 7. *Under Assumption 1, $\mathbb{E}_{\mathbf{x}, f_*} [f_*(\mathbf{x})^2] = \Theta_{d,r}(1)$ and $\mathbb{E}_{\mathbf{x}, f_*} [f_*(\mathbf{x})^4] = O_{d,r}(1)$ holds.*

Proof. From (2.2) and $P = \Theta_{d,r}(1)$,

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d), f_* \sim \mathcal{D}_{f_*}} [f_*(\mathbf{x})^2] &= \mathbb{E}_{z \sim \mathcal{N}(0,1), \{c_i\}} \left[\left(\sum_{i=Q}^P \frac{c_i}{i!} \text{He}_i(z) \right)^2 \right] \\ &= \mathbb{E} \left[\sum_{i=Q}^P \frac{c_i^2}{i!} \right] \\ &= \Theta_{d,r}(1). \end{aligned}$$

We can bound $\mathbb{E}_{\mathbf{x}} [f_*(\mathbf{x})^4] = \mathbb{E}_{z \sim \mathcal{N}(0,1), \{c_i\}} \left[\left(\sum_{i=Q}^P \frac{c_i}{i!} \text{He}_i(z) \right)^4 \right]$ naively as follows: let $\zeta_P^4 = \max_{Q \leq i \leq P} \mathbb{E}_{z \sim \mathcal{N}(0,1)} [\text{He}_i(z)^4]$. ζ_P is an $O(1)$ quantity depending only on P and Q . Then,

$$\begin{aligned} &\mathbb{E}_{z \sim \mathcal{N}(0,1), \{c_i\}} \left[\left(\sum_{i=Q}^P \frac{c_i}{i!} \text{He}_i(z) \right)^4 \right] \\ &\leq \sum_{Q \leq i,j,k,l \leq P} \frac{\mathbb{E}[c_i c_j c_k c_l]}{i! j! k! l!} \mathbb{E}[|\text{He}_i(z) \text{He}_j(z) \text{He}_k(z) \text{He}_l(z)|] \\ &\leq \sum_{Q \leq i,j,k,l \leq P} R_c^4 \mathbb{E}_z [\text{He}_i(z)^4]^{1/4} \mathbb{E}_z [\text{He}_j(z)^4]^{1/4} \mathbb{E}_z [\text{He}_k(z)^4]^{1/4} \mathbb{E}_z [\text{He}_l(z)^4]^{1/4} \\ &\leq P^4 R_c^4 \zeta_P^4. \end{aligned}$$

□

The lemma below is useful to find a basis of the set of single-index functions.

Lemma 8. *Suppose $\beta \in \mathbb{S}(\mathcal{S})$. Then,*

$$\text{He}_p(\langle \mathbf{x}, \beta \rangle) = \sum_{p_1 \geq 0, \dots, p_r \geq 0} \frac{p!}{p_1! \cdots p_r!} \cdot \beta_1^{p_1} \cdots \beta_r^{p_r} \cdot \text{He}_{p_1}(x_1) \cdots \text{He}_{p_r}(x_r)$$

holds.

Proof. Note that $\mathbb{E}_{x_1, \dots, x_r \sim \mathcal{N}(0,1)} \left[\frac{\partial^{i_1}}{\partial x_1^{i_1}} \cdots \frac{\partial^{i_r}}{\partial x_r^{i_r}} \text{He}_p(\langle \mathbf{x}, \beta \rangle) \right]$ is nonzero only when $i_1 + \cdots + i_r = p$, because $\frac{\partial^{i_1}}{\partial x_1^{i_1}} \cdots \frac{\partial^{i_r}}{\partial x_r^{i_r}} \text{He}_p(\langle \mathbf{x}, \beta \rangle) = p(p-1) \cdots (p - (i_1 + \cdots + i_r) + 1) \beta_1^{i_1} \cdots \beta_r^{i_r} \text{He}_{p-(i_1+\dots+i_r)}(\langle \mathbf{x}, \beta \rangle)$ and $\mathbb{E}_{x_1, \dots, x_r \sim \mathcal{N}(0,1)} [\text{He}_q(\langle \mathbf{x}, \beta \rangle)] = \mathbb{E}_{z \sim \mathcal{N}(0,1)} [\text{He}_q(z)] = 0$ if $q > 0$. When $i_1 + \cdots + i_r = p$, then $\mathbb{E}_{x_1, \dots, x_r \sim \mathcal{N}(0,1)} \left[\frac{\partial^{i_1}}{\partial x_1^{i_1}} \cdots \frac{\partial^{i_r}}{\partial x_r^{i_r}} \text{He}_p(\langle \mathbf{x}, \beta \rangle) \right] = p! \beta_1^{i_1} \cdots \beta_r^{i_r}$ holds. Thus, from the multivariate Hermite expansion we obtain the claim. □

Corollary 9. *Let $f_*(\mathbf{x}) = \sum_{i=Q}^P \frac{c_i}{i!} \text{He}_i(\langle \mathbf{x}, \beta \rangle)$. Then, $\mathbb{E}_{\mathbf{x}} [\nabla^k f_*(\mathbf{x})] = c_k \beta^{\otimes k}$ if $Q \leq k \leq P$ and otherwise it is the zero tensor.*

A.4 Other Auxiliary Lemmas

Lemmas on random vectors. Let $\chi(d)$ be the distribution of $\sqrt{z_1^2 + \dots + z_d^2}$ where z_i i.i.d. $\mathcal{N}(0, 1)$. Note that the Gaussian vector $\mathbf{v} \sim \mathcal{N}(0, I_d)$ can be decomposed as $\mathbf{v} = z\mathbf{w}$ where $z \sim \chi(d)$ and $\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})$, because the norm and the direction of the Gaussian vector are independent of each other.

Lemma 10 (Example 2.11 in [Wai19]). For $z \sim \chi(d)$ and $0 < t < 1$,

$$\mathbb{P}[|z^2 - d| \geq dt] \leq 2 \exp(-dt^2/8)$$

holds.

Lemma 11. Let $C > 0$. For $\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})$, $\|\mathbf{w}_{1:r}\|^2 \leq \frac{4Cr}{d} \log d$ holds with probability at least $1 - 2 \exp(-d/32) - 2rd^{-C}$ (then with high probability).

Proof. Let $z^2 \sim \chi^2(d)$ be independent from \mathbf{w} . We observe that the distribution of $z^2 \|\mathbf{w}_{1:r}\|^2$ is the $\chi^2(r)$ -distribution. First, from Lemma 10, with probability at least $1 - 2 \exp(-d/32)$, $z^2 \geq d/2$ holds. Moreover, we can say that $(z')^2 \sim \chi^2(r)$ satisfies $(z')^2 \leq 2Cr \log d$ with high probability for sufficiently large constant: let us decompose $(z')^2 = v_1^2 + \dots + v_r^2$ where $v_i \sim \mathcal{N}(0, 1)$; from Lemma 4, $v_i^2 \leq 2C \log d$ holds with probability at least $1 - 2d^{-C}$, and thus $(z')^2 \leq 2Cr \log d$ holds with probability at least $1 - 2rd^{-C}$. Taking the union bound yields the result. \square

Lemma 12. For $k \geq 1$ and $r \leq d$,

$$\mathbb{E}_{\mathbf{w} \sim \mathbb{S}^{d-1}} [\|\mathbf{w}_{1:r}\|^{2k}] = O_{d,r} \left(\left(\frac{r}{d} \right)^k \right)$$

holds.

Proof. Since $\|\mathbf{w}_{1:r}\|^2$ is a polynomial of \mathbf{w} , Lemma 24 in [DLS22] yields $\mathbb{E}_{\mathbf{w} \sim \mathbb{S}^{d-1}} [\|\mathbf{w}_{1:r}\|^{2k}] \lesssim \mathbb{E}_{\mathbf{w} \sim \mathbb{S}^{d-1}} [\|\mathbf{w}_{1:r}\|^4]^{k/2}$. Here, $\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\|\mathbf{w}_{1:r}\|^4] = \frac{\mathbb{E}_{v_i \sim \mathcal{N}(0,1)} [(v_1^2 + \dots + v_r^2)^2]}{\mathbb{E}_{z \sim \chi^4(d)} [z]} = \frac{3r+r(r-1)}{d(d+2)} = \Theta(r^2/d^2)$ yields the result. \square

Lemma 13 (Sub Gaussian vector). Let \mathbf{x} be a d -dimensional random vector and suppose that $\langle \mathbf{u}, \mathbf{x} \rangle$ is zero-mean σ^2 -sub Gaussian for any $\mathbf{u} \in \mathbb{R}^d$ such that $\|\mathbf{u}\| = 1$. Then, $\|\mathbf{x}\| = \tilde{O}(\sigma\sqrt{d})$ holds with high probability.

Lemma 14. Let X be a σ^2 -sub Gaussian random variable and Y be a bounded random variable satisfying $|Y| \leq R$ almost surely. Then, $XY - \mathbb{E}[XY]$ is $O(\sigma^2 R^2)$ -sub Gaussian.

Proof. XY is $\sigma^2 R^2$ -Gaussian by examining the tail probability. Hence, $XY - \mathbb{E}[XY]$ is $O(\sigma^2 R^2)$ -sub Gaussian from Lemma 2.6.8 in [Ver18]. \square

Polynomial concentration. We frequently use the following bounds on an output data y (which is a polynomial of \mathbf{x}).

Lemma 15 (Theorem 1.5 in [GSS21]). Let $\mathbf{x} \sim \mathcal{N}(0, I_n)$ and f be a degree- D n -variable polynomial function. Then, for all $t \geq 0$ it holds that

$$\mathbb{P}(|f(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]| \geq t) \leq 2 \exp \left(-\frac{1}{C_D M^2} \min_{1 \leq k \leq D} \left(\frac{t}{\|\mathbb{E}_{\mathbf{x}}[\nabla^k f(\mathbf{x})]\|_F} \right)^{2/k} \right),$$

where M is an absolute constant and C_D is a constant depending only on D (if $\|\mathbb{E}_{\mathbf{x}}[\nabla^k f(\mathbf{x})]\|_F = 0$ then we ignore the term with k).

Lemma 16. Let $\mathbf{x} \sim \mathcal{N}(0, I_d)$ and $y = f_*(\mathbf{x}) + \varsigma = \sum_{i=Q}^P \frac{c_i}{i!} \text{He}_i(\langle \mathbf{x}, \boldsymbol{\beta} \rangle) + \varsigma$ where c_i and $\boldsymbol{\beta}$ are drawn according to Assumption 1, and $\varsigma \sim \text{Unif}\{\pm\tau\}$. Then, for any fixed $\{c_i\}$ and $\boldsymbol{\beta}$ on the support of their distributions, it holds that

$$\mathbb{P}(|y| \geq t + \tau) \leq 2 \exp \left(-\frac{1}{C_P M^2} \min_{2 \leq k \leq P} \left(\frac{t}{R_c} \right)^{2/k} \right),$$

where M is an absolute constant and C_P is a constant depending only on P .

Proof. Note that from Corollary 9, $\mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})] = 0$, $\mathbb{E}_{\mathbf{x}}[\nabla f(\mathbf{x})] = \mathbf{0}$ and $\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})] = c_i(\boldsymbol{\beta})^{\otimes i}$; by (2.2), Lemma 6 and $\|\boldsymbol{\beta}\| = 1$, $\|\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})]\|_F \leq R_c$ holds. Lemma 15 yields the result. \square

Corollary 17. Fix any $\{c_i\}$ and $\boldsymbol{\beta}$ on the support of their distributions. Then, $|f_*(\mathbf{x})|, |y| \lesssim (\log d)^{P/2}$ holds with high probability over the distribution of \mathbf{x} and ς .

B Proofs for the MLP Layer

This section analyzes the behavior of MLP weight \mathbf{w} after one gradient descent step (line 4 in Algorithm 1), as sketched in Section 3.2.1.

We set $\lambda_1 = \eta_1^{-1}$. From line 4 of Algorithm 1, for each $j \in [m]$,

$$\begin{aligned} \mathbf{w}_j^{(1)} &= 2\eta_1 \frac{1}{T_1} \sum_{t=1}^{T_1} y^t \nabla_{\mathbf{w}_j^{(0)}} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}) \\ &\quad - 2\eta_1 \frac{1}{T_1} \sum_{t=1}^{T_1} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}) \nabla_{\mathbf{w}_j^{(0)}} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}) \\ &= 2\eta_1 \Gamma_{j,j}^{(0)} \left(\sum_{t=1}^{T_1} \frac{1}{T_1} y^t \sigma_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}^t) \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma'_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}_i^t) \mathbf{x}_i^t \right. \\ &\quad \left. + \sum_{t=1}^{T_1} \frac{1}{T_1} y^t \sigma'_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}^t) \mathbf{x}^t \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}_i^t) \right) \\ &\quad - 2\eta_1 \frac{1}{T_1} \sum_{t=1}^{T_1} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}) \nabla_{\mathbf{w}_j^{(0)}} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}). \end{aligned} \tag{B.1}$$

holds where $\sigma_{b_j}(z) := \sigma(z + b_j)$. Here, \mathbf{x}_i^t and y_i^t is the input and output of i -th context at t -th task, respectively. First we show that the final term is sufficiently small, if we set $\gamma = |\Gamma_{j,j}^{(0)}|$ sufficiently small.

Lemma 18. With high probability over the distribution of $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$ and $\{\mathbf{w}_j^{(0)}\}_{j=1}^m$,

$$\begin{aligned} &\left\| 2\eta_1 \frac{1}{T_1} \sum_{t=1}^{T_1} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}) \nabla_{\mathbf{w}_j^{(0)}} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}) \right\| \\ &= \tilde{O}\left(\eta_1 \gamma^2 m \sqrt{d}\right). \end{aligned}$$

Proof. Recall that

$$\begin{aligned} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}) &= \sum_{j=1}^m \Gamma_{j,j}^{(0)} \left(\frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}_i^t) \right) \sigma_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}^t), \\ \nabla_{\mathbf{w}_j^{(0)}} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{b}^{(0)}) &= \Gamma_{j,j}^{(0)} \left(\frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma'_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}_i^t) \mathbf{x}_i^t \right) \sigma_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}^t) \\ &\quad + \Gamma_{j,j}^{(0)} \left(\frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}_i^t) \right) \sigma'_{b_j}(\mathbf{w}_j^{(0)\top} \mathbf{x}^t) \mathbf{x}^t. \end{aligned}$$

Fix i, t, j and consider the inner product between $\mathbf{w}_j^{(0)} \sim \text{Unif}(\mathbb{S}^{d-1})$ and $\mathbf{x}_i^t \sim \mathcal{N}(0, I_d)$. From the rotational invariance, without loss of generality, we can assume that $\mathbf{w}_j^{(0)} = [1, 0, \dots, 0]^\top$ and $\mathbf{x}_i^t \sim \mathcal{N}(0, I_d)$. Therefore, $\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i^t \rangle \sim \mathcal{N}(0, 1)$. From Lemma 4, $\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i^t \rangle^2 \lesssim \log d$ holds with high probability. Moreover, from Lemma 10 and Corollary 17, we know that $|y_i^t|, |y^t| \lesssim (\log d)^{P/2}$

and $\|\mathbf{x}_i^t\|, \|\mathbf{x}^t\| \lesssim \sqrt{d}$ holds. We can take the union bound with respect to all the i, t, j , retaining the high probability bound (see Appendix A). Therefore, noting that $|b_j| \leq 1$, we obtain

$$|f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \mathbf{\Gamma}^{(0)}, \mathbf{b}^{(0)})| \lesssim m\gamma(\log d)^{P/2+1}$$

and

$$\|\nabla_{\mathbf{w}_j^{(0)}} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \mathbf{\Gamma}^{(0)}, \mathbf{b}^{(0)})\| \lesssim \gamma(\log d)^{P/2+1/2} \sqrt{d}.$$

Thus we obtain the assertion. \square

From now on we focus on analyzing (B.1), which expresses the correlation between the output label y and the gradient of model output $\nabla_{\mathbf{w}} f$. Let

$$\begin{aligned} & \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) \\ &:= \sum_{i=1}^{T_1} \frac{1}{T_1} y^t \sigma_b(\mathbf{w}^\top \mathbf{x}^t) \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t \\ & \quad + \sum_{i=1}^{T_1} \frac{1}{T_1} y^t \sigma'_b(\mathbf{w}^\top \mathbf{x}^t) \mathbf{x}^t \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_b(\mathbf{w}^\top \mathbf{x}_i^t), \end{aligned} \quad (\text{B.2})$$

and $\mathbf{g}(\mathbf{w}, b) = \mathbb{E}[\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})]$, where the expectation is taken with respect to the data $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$. Note that

$$\mathbf{w}_j^{(1)} = 2\eta_1 \mathbf{\Gamma}_{j,j}^{(0)} \mathbf{g}_{T_1, N_1}(\mathbf{w}_j^{(0)}, b_j^{(0)}, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) + \tilde{O}(\eta_1 \gamma^2 m \sqrt{d})$$

from Lemma 18.

In Appendix B.1, we analyze the explicit form of $\mathbf{g}(\mathbf{w}, b)$ via the Hermite expansion: we can check that its main term aligns with the low-dimensional subspace \mathcal{S} . In Appendix B.2, we show the concentration of \mathbf{g}_{T_1, N_1} around \mathbf{g} .

B.1 Calculation of Population Gradient

B.1.1 Hermite Expansion

First, note that

$$\begin{aligned} \mathbf{g}(\mathbf{w}, b) &= \mathbb{E}[y \sigma_b(\mathbf{w}^\top \mathbf{x}) y_i \sigma'_b(\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i] + \mathbb{E}[y \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x} y_i \sigma_b(\mathbf{w}^\top \mathbf{x}_i)] \\ &= \mathbb{E}_{f_*} [\mathbb{E}_{\{\mathbf{x}_i\}, \mathbf{x}, \{\varsigma_i\}, \varsigma} [(f_*(\mathbf{x}) + \varsigma) \sigma_b(\mathbf{w}^\top \mathbf{x}) (f_*(\mathbf{x}_i) + \varsigma_i) \sigma'_b(\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i]] \\ & \quad + \mathbb{E}_{f_*} [\mathbb{E}_{\{\mathbf{x}_i\}, \mathbf{x}, \{\varsigma_i\}, \varsigma} [(f_*(\mathbf{x}) + \varsigma) \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x} (f_*(\mathbf{x}_i) + \varsigma_i) \sigma_b(\mathbf{w}^\top \mathbf{x}_i)]] \\ &= \mathbb{E}_{f_*} [\mathbb{E}_{\mathbf{x}, \varsigma} [(f_*(\mathbf{x}) + \varsigma) \sigma_b(\mathbf{w}^\top \mathbf{x})] \mathbb{E}_{\{\mathbf{x}_i\}, \{\varsigma_i\}} [(f_*(\mathbf{x}_i) + \varsigma_i) \sigma'_b(\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i]] \\ & \quad + \mathbb{E}_{f_*} [\mathbb{E}_{\mathbf{x}, \varsigma} [(f_*(\mathbf{x}) + \varsigma) \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \mathbb{E}_{\{\mathbf{x}_i\}, \{\varsigma_i\}} [(f_*(\mathbf{x}_i) + \varsigma_i) \sigma_b(\mathbf{w}^\top \mathbf{x}_i)]] \\ &= 2\mathbb{E}_{f_*} [\mathbb{E}_{\mathbf{x}} [f_*(\mathbf{x}) \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \mathbb{E}_{\mathbf{x}} [f_*(\mathbf{x}) \sigma_b(\mathbf{w}^\top \mathbf{x})]]. \end{aligned}$$

Here \mathbb{E}_{f_*} denotes the expectation with respect to the distribution of the true function f_* (i.e., the distribution of $\{c_i\}_{i=Q}^P$ and β) specified in Assumption 1. Now let $\sigma_b(z) = \sum_{i \geq 0} \frac{a_i(b)}{i!} \text{He}_i(z)$ be the Hermite expansion of student activation $\sigma_b(z) = \text{ReLU}(z + b)$.

Lemma 19. *It holds that*

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}} [f_*(\mathbf{x}) \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \\ &= \sum_{i=Q-1}^{P-1} \frac{a_{i+1}(b) \mathbb{E}_{\mathbf{x}} [\nabla^{i+1} f_*(\mathbf{x})] (\mathbf{w}^{\otimes i})}{i!} + \mathbf{w} \sum_{i=Q}^P \frac{a_{i+2}(b) \mathbb{E}_{\mathbf{x}} [\nabla^i f_*(\mathbf{x})] (\mathbf{w}^{\otimes i})}{i!}, \quad (\text{B.3}) \\ & \mathbb{E}_{\mathbf{x}} [f_*(\mathbf{x}) \sigma_b(\mathbf{w}^\top \mathbf{x})] = \sum_{i=Q}^P \frac{a_i(b) \mathbb{E}_{\mathbf{x}} [\nabla^i f_*(\mathbf{x})] (\mathbf{w}^{\otimes i})}{i!}. \end{aligned}$$

Proof. Consider a function $g(\mathbf{w}^\top \mathbf{x}) = \sum_{i \geq 0} \frac{\tilde{a}_i}{i!} \text{He}_i(\mathbf{w}^\top \mathbf{x})$. Here $\mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})g(\mathbf{w}^\top \mathbf{x})]$ can be calculated as

$$\begin{aligned} \mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})g(\mathbf{w}^\top \mathbf{x})] &= \sum_{i \geq 0} \frac{\tilde{a}_i}{i!} \mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})\text{He}_i(\mathbf{w}^\top \mathbf{x})] \\ &= \sum_{i \geq 0} \sum_{i_1 \geq 0, \dots, i_d \geq 0}^{i_1 + \dots + i_d = i} \tilde{a}_i \frac{w_1^{i_1} \dots w_d^{i_d}}{i_1! \dots i_d!} \mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})\text{He}_{i_1}(x_1) \dots \text{He}_{i_d}(x_d)] \quad (\because \text{Lemma 8}) \\ &= \sum_{i \geq 0} \sum_{i_1 \geq 0, \dots, i_d \geq 0}^{i_1 + \dots + i_d = i} \tilde{a}_i \frac{w_1^{i_1} \dots w_d^{i_d}}{i_1! \dots i_d!} \mathbb{E}_{\mathbf{x}} \left[\frac{\partial^{i_1}}{\partial x_1^{i_1}} \dots \frac{\partial^{i_d}}{\partial x_d^{i_d}} f_*(\mathbf{x}) \right] \\ &= \sum_{i \geq 0} \frac{\tilde{a}_i \mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!}. \end{aligned}$$

Where the final line is obtained noting that the tensor $\nabla^i f_*(\mathbf{x})$ has $\frac{i!}{i_1! \dots i_d!}$ entries whose value is $\frac{\partial^{i_1}}{\partial x_1^{i_1}} \dots \frac{\partial^{i_d}}{\partial x_d^{i_d}} f_*(\mathbf{x})$.

Note that $\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})] = \mathbb{E}_{\mathbf{x}} \left[\nabla^i \sum_{j=Q}^P \frac{c_j}{j!} \text{He}_j(\langle \mathbf{x}, \boldsymbol{\beta} \rangle) \right]$ is a nonzero tensor only when $Q \leq i \leq P$, from Lemma 8. We can also show (B.3) as

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})\sigma'_b(\mathbf{w}^\top \mathbf{x})] \\ &= \mathbb{E}_{\mathbf{x}}[\nabla f_*(\mathbf{x})\sigma'_b(\mathbf{w}^\top \mathbf{x})] + \mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})\sigma''_b(\mathbf{w}^\top \mathbf{x})]\mathbf{w} \quad (\because \text{Stein's lemma}) \\ &= \sum_{i \geq 0} \frac{a_{i+1}(b)\mathbb{E}_{\mathbf{x}}[\nabla^{i+1} f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} + \mathbf{w} \sum_{i \geq 0} \frac{a_{i+2}(b)\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} \\ &= \sum_{i=Q-1}^{P-1} \frac{a_{i+1}(b)\mathbb{E}_{\mathbf{x}}[\nabla^{i+1} f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} + \mathbf{w} \sum_{i=Q}^P \frac{a_{i+2}(b)\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} \end{aligned}$$

□

Using Lemma 19, $g(\mathbf{w}, b)$ can be expanded as

$$\begin{aligned} g(\mathbf{w}, b) &= \mathbb{E}_{f_*} \left[2 \left(\sum_{i=Q-1}^{P-1} \frac{a_{i+1}(b)\mathbb{E}_{\mathbf{x}}[\nabla^{i+1} f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} \right. \right. \\ &\quad \left. \left. + \mathbf{w} \sum_{i=Q}^P \frac{a_{i+2}(b)\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} \right) \left(\sum_{i=Q}^P \frac{a_i(b)\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} \right) \right] \\ &=: \mathbb{E}_{f_*} \left[\frac{2a_Q(b)}{(Q-1)!} \mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes(Q-1)}) \frac{a_Q(b)\mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes Q})}{Q!} \right] + \mathbf{s}(\mathbf{w}, b) \\ &= \frac{2a_Q(b)^2}{Q!(Q-1)!} \mathbb{E}_{f_*} [\mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes(Q-1)}) \cdot \mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes Q})] + \mathbf{s}(\mathbf{w}, b), \quad (\text{B.4}) \end{aligned}$$

where

$$\begin{aligned} &\mathbf{s}(\mathbf{w}, b) \\ &:= 2\mathbb{E}_{f_*} \left[\left(\sum_{i=Q-1}^{P-1} \frac{a_{i+1}(b)\mathbb{E}_{\mathbf{x}}[\nabla^{i+1} f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} + \mathbf{w} \sum_{i=Q}^P \frac{a_{i+2}(b)\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} \right) \right. \\ &\quad \left. \cdot \left(\sum_{i=Q}^P \frac{a_i(b)\mathbb{E}_{\mathbf{x}}[\nabla^i f_*(\mathbf{x})](\mathbf{w}^{\otimes i})}{i!} \right) - \frac{a_Q(b)^2}{(Q-1)!} \mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes(Q-1)}) \frac{\mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes Q})}{Q!} \right] \end{aligned}$$

expresses terms in the asymptotic expansion which are not the leading term.

Note that, in this explicit formulation, $a_i(b) = \mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_b(z)\text{He}_i(z)]$ appears. It will be helpful to show that this coefficient is well-controlled.

Lemma 20. *There are constants C_H and C_L depending only on Q and P satisfying the following condition. Let $\sigma_b(z) = \sigma(z + b)$ and $b \sim \text{Unif}([-1, 1])$. Then,*

$$|\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_b(z)\text{He}_i(z)]| \leq C_H \quad (2 \leq i \leq P+2) \text{ and } |\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_b(z)\text{He}_Q(z)]| \geq C_L \quad (\text{B.5})$$

holds with probability 1/2. Especially, $|\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_b(z)\text{He}_i(z)]| \leq C_H \quad (2 \leq i \leq P+2)$ holds with probability 1.

Proof. From Lemma 15 in [BES⁺23], we know that for $i \geq 2$, $|\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_b(z)\text{He}_i(z)]| = \frac{e^{-b^2/2}}{\sqrt{2\pi}} |\text{He}_{i-2}(b)|$ holds. First, as $b \sim \text{Unif}([-1, 1])$, by setting sufficiently large C_H , $|\text{He}_i(b)| \leq C_H$ holds for all $0 \leq i \leq P$. Therefore, as $|\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_b(z)\text{He}_i(z)]| \leq |\text{He}_{i-2}(b)|$, $|\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_b(z)\text{He}_i(z)]| \leq C_H$ holds for all $2 \leq i \leq P+2$ with probability one.

From continuity we know that there exists C_L such that $|\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma_b(z)\text{He}_Q(z)]| \geq C_L$ with probability 1/2. \square

In the following, we first calculate the main term and then upper bound the residual terms.

B.1.2 Calculation of the Main Term

Let us calculate the main term in (B.4) explicitly.

Lemma 21.

$$\begin{aligned} & \frac{2a_Q(b)^2}{Q!(Q-1)!} \mathbb{E}_{f_*} [\mathbb{E}_{\mathbf{x}} [\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes(Q-1)}) \cdot \mathbb{E}_{\mathbf{x}} [\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes Q})] \\ &= \frac{2a_Q(b)^2 \mathbb{E}_{c_Q} [c_Q^2] (2Q-1)!!}{Q!(Q-1)! \mathbb{E}_{z \sim \chi_r} [z^{2Q}]} \|\mathbf{w}_{1:r}\|^{2Q-2} \begin{bmatrix} \mathbf{w}_{1:r} \\ \mathbf{0}_{d-r} \end{bmatrix} \end{aligned}$$

holds.

Proof. From Corollary 9, it holds that $\mathbb{E}_{\mathbf{x}} [\nabla^Q f_*(\mathbf{x})] = c_Q(\boldsymbol{\beta})^{\otimes Q}$. Then, we obtain

$$\begin{aligned} & \mathbb{E}_{f_*} [\mathbb{E}_{\mathbf{x}} [\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes(Q-1)}) \cdot \mathbb{E}_{\mathbf{x}} [\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes Q})] \\ &= \mathbb{E}_{c_Q} [c_Q^2] \mathbb{E}_{\boldsymbol{\beta}} [(\boldsymbol{\beta}^{\otimes Q} \mathbf{w}^{\otimes(Q-1)}) \cdot (\boldsymbol{\beta}^{\otimes Q} \circ \mathbf{w}^{\otimes Q})] \\ &= \mathbb{E}_{c_Q} [c_Q^2] \mathbb{E}_{\boldsymbol{\beta}} [\boldsymbol{\beta}^{\otimes 2Q}] (\mathbf{w}^{\otimes(2Q-1)}). \end{aligned}$$

Next, let us show

$$\mathbb{E}_{\boldsymbol{\beta}} [\boldsymbol{\beta}^{\otimes 2Q}] (\mathbf{w}^{\otimes(2Q-1)}) = \frac{(2Q-1)!!}{\mathbb{E}_{z \sim \chi_r} [z^{2Q}]} \|\mathbf{w}_{1:r}\|^{2Q-2} \begin{bmatrix} \mathbf{w}_{1:r} \\ \mathbf{0}_{d-r} \end{bmatrix}.$$

First, note that by letting $\boldsymbol{\beta}' \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0}_r \\ \mathbf{0}_r^\top & \mathbf{O}_{d-r} \end{bmatrix},$$

and $z \sim \chi_r$ independent from $\boldsymbol{\beta}$, $\boldsymbol{\beta}' \sim \boldsymbol{\beta}z$ holds (recall that we assumed that $\boldsymbol{\beta} \sim \text{Unif}(\{(\beta_1, \dots, \beta_r, 0, \dots, 0)^\top \mid \beta_1^2 + \dots + \beta_r^2 = 1\})$ – see Appendix A). Then,

$$\mathbb{E}_{\boldsymbol{\beta}} [\boldsymbol{\beta}^{\otimes 2Q}] = \frac{1}{\mathbb{E}_{z \sim \chi_r} [z^{2Q}]} \mathbb{E}_{\mathbf{v}'_{1:r} \sim \mathcal{N}(0, \mathbf{I}_r), \mathbf{v}'_{r+1:d} = \mathbf{0}} [\mathbf{v}'^{\otimes 2Q}].$$

It suffices to show that $\mathbb{E}_{\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_r)} [\mathbf{v}^{\otimes 2Q}] (\mathbf{z}^{\otimes(2Q-1)}) = (2Q-1)!! \|\mathbf{z}\|^{2Q-2}$ for $\mathbf{z} \in \mathbb{R}^r$. As an example, we show that the first entry of $\mathbb{E}_{\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_r)} [\mathbf{v}^{\otimes 2Q}] (\mathbf{z}^{\otimes(2Q-1)})$ equals to $(2Q-1)!! z_1 \|\mathbf{z}\|^{2Q-2}$; the other components can be calculated similarly. First, we expand $(\mathbb{E}_{\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_r)} [\mathbf{v}^{\otimes 2Q}] (\mathbf{z}^{\otimes(2Q-1)}))_1$ as

$$(\mathbb{E}_{\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_r)} [\mathbf{v}^{\otimes 2Q}] (\mathbf{z}^{\otimes(2Q-1)}))_1 = \sum_{i_1, \dots, i_{2Q-1} \in [r]} \mathbb{E}[v_1 v_{i_1} \cdots v_{i_{2Q-1}}] z_{i_1} \cdots z_{i_{2Q-1}}.$$

Note that $\mathbb{E}[v_1 v_{i_1} \cdots v_{i_{2Q-1}}] \neq 0$ if and only if the degree of $v_1 v_{i_1} \cdots v_{i_{2Q-1}}$ with respect to v_j is even, for all $j \in [r]$. Then, we obtain

$$(\mathbb{E}_{\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_r)} [\mathbf{v}^{\otimes 2Q}] (\mathbf{z}^{\otimes(2Q-1)}))_1$$

$$\begin{aligned}
&= \sum_{\substack{q_1+\dots+q_r=Q \\ q_1 \geq 1, q_2, \dots, q_r \geq 0}} c_{q_1, \dots, q_r} \mathbb{E}[v_1^{2q_1} v_2^{2q_2} \dots v_r^{2q_r}] z_1^{2q_1-1} z_2^{2q_2} \dots z_r^{2q_r} \\
&= z_1 \sum_{\substack{q_1+\dots+q_r=Q \\ q_1 \geq 1, q_2, \dots, q_r \geq 0}} c_{q_1, \dots, q_r} \mathbb{E}[v_1^{2q_1} v_2^{2q_2} \dots v_r^{2q_r}] z_1^{2q_1-2} z_2^{2q_2} \dots z_r^{2q_r},
\end{aligned}$$

where $c_{q_1, \dots, q_r} = \frac{(2Q-1)!}{(2q_1-1)!(2q_2)! \dots (2q_r)!}$ is the number of ways to assign $(2q_1 - 1)$ “1”s, $(2q_2)$ “2”s, ..., among i_1 to i_{2Q-1} .

Note that $\mathbb{E}_{v \sim \mathcal{N}(0,1)}[v^{2q}] = (2q - 1)!!$. Therefore, we obtain

$$\begin{aligned}
&(\mathbb{E}_{\mathbf{v} \sim \mathcal{N}(0, I_r)}[\mathbf{v}^{\otimes 2Q}](\mathbf{z}^{\otimes (2Q-1)}))_1 \\
&= z_1 \sum_{\substack{q_1+\dots+q_r=Q \\ q_1 \geq 1, q_2, \dots, q_r \geq 0}} \frac{(2Q-1)!}{(2q_1-1)!(2q_2)! \dots (2q_r)!} (2q_1-1)!! \dots (2q_r-1)!! z_1^{2q_1-2} z_2^{2q_2} \dots z_r^{2q_r} \\
&= z_1 \sum_{\substack{q_1+\dots+q_r=Q \\ q_1 \geq 1, q_2, \dots, q_r \geq 0}} \frac{(Q-1)!}{(q_1-1)!(q_2)! \dots (q_r)!} (2Q-1)!! (z_1^2)^{q_1-1} (z_2^2)^{q_2} \dots (z_r^2)^{q_r} \\
&= (2Q-1)!! z_1 (z_1^2 + \dots + z_r^2)^{Q-1} \quad (\because \text{multinomial theorem})
\end{aligned}$$

which concludes the assertion. \square

By Lemma 21, we observe that the main term of the expected one-step gradient aligns to the r -dimensional subspace \mathcal{S} . In other words, the MLP layer captures the information of \mathcal{S} via pretraining, which will be useful in the later stage.

B.1.3 Bounding Residual Terms

First, similarly to the Lemma 21, we can obtain the explicit formulation of the entire $\mathbf{g}(\mathbf{w}, b)$.

Lemma 22.

$$\begin{aligned}
&\mathbf{g}(\mathbf{w}, b) \\
&= \sum_{\substack{Q-1 \leq i \leq P-1, \\ Q \leq j \leq P, \\ i+j \text{ is odd}}} \frac{2a_{i+1}(b)a_j(b)\mathbb{E}[c_{i+1}c_j]}{i!j!} \frac{(i+j)!!}{\mathbb{E}_{z \sim \chi_r}[z^{i+j+1}]} \|\mathbf{w}_{1:r}\|^{i+j-1} \begin{bmatrix} \mathbf{w}_{1:r} \\ \mathbf{0}_{d-r} \end{bmatrix} \\
&+ \mathbf{w} \sum_{\substack{Q \leq i \leq P, \\ Q \leq j \leq P, \\ i+j \text{ is even}}} \frac{2a_{i+2}(b)a_j(b)\mathbb{E}[c_i c_j]}{i!j!} \frac{(i+j-1)!!}{\mathbb{E}_{z \sim \chi_r}[z^{i+j}]} \|\mathbf{w}_{1:r}\|^{i+j}
\end{aligned} \tag{B.6}$$

holds.

Proof. Note that

$$\mathbb{E}_{\beta}[\beta^{\otimes n}](\mathbf{w}^{\otimes (n-1)}) = \begin{cases} \frac{(n-1)!!}{\mathbb{E}_{z \sim \chi_r}[z^n]} \|\mathbf{w}_{1:r}\|^{n-2} \begin{bmatrix} \mathbf{w}_{1:r} \\ \mathbf{0}_{d-r} \end{bmatrix} & (n \text{ is even}) \\ 0 & (n \text{ is odd}) \end{cases}$$

and

$$\mathbb{E}_{\beta}[\beta^{\otimes n}](\mathbf{w}^{\otimes n}) = \begin{cases} \frac{(n-1)!!}{\mathbb{E}_{z \sim \chi_r}[z^n]} \|\mathbf{w}_{1:r}\|^n & (n \text{ is even}) \\ 0 & (n \text{ is odd}) \end{cases},$$

which can be obtained similarly to the proof of Lemma 21. \square

Let us upper bound the non-leading terms; we need to show that the moment of first r components in residual term $\mathbf{s}(\mathbf{w}, b)$ are sufficiently small.

Lemma 23. Let $j = 4k$ where $k \in [P]$. then,

$$\mathbb{E}_{\mathbf{w} \sim \mathbb{S}^{d-1}} [\|\mathbf{s}(\mathbf{w}, b)_{1:r}\|^j]^{1/j} = \tilde{O}_{d,r} \left(\sqrt{\frac{r}{d^{2Q+1}}} \right) \quad (\text{B.7})$$

holds uniformly over all $b \in [-1, 1]$.

Proof. Note that $\|\mathbf{s}(\mathbf{w}, b)_{1:r}\|^2$ is a polynomial in \mathbf{w} ; from Lemma 24 in [DLS22], it suffices to show the case $k = 1$.

$$\begin{aligned} & \mathbf{s}(\mathbf{w}, b)_{1:r} \\ = & \sum_{\substack{Q-1 \leq i \leq P-1, \\ Q \leq j \leq P, \\ i+j \text{ is odd,} \\ i, j \neq (Q-1, Q)}} \frac{2a_{i+1}(b)a_j(b)\mathbb{E}[c_{i+1}c_j]}{i!j!} \frac{(i+j)!!}{\mathbb{E}_{z \sim \chi_r}[z^{i+j+1}]} \|\mathbf{w}_{1:r}\|^{i+j-1} \mathbf{w}_{1:r} \\ + & \sum_{\substack{Q \leq i \leq P, \\ Q \leq j \leq P, \\ i+j \text{ is even}}} \frac{2a_{i+2}(b)a_j(b)\mathbb{E}[c_i c_j]}{i!j!} \frac{(i+j-1)!!}{\mathbb{E}_{z \sim \chi_r}[z^{i+j}]} \|\mathbf{w}_{1:r}\|^{i+j} \mathbf{w}_{1:r}. \end{aligned}$$

By Minkowski's inequality $\mathbb{E}[\|\mathbf{x} + \mathbf{y}\|^4]^{1/4} \leq \mathbb{E}[\|\mathbf{x}\|^4]^{1/4} + \mathbb{E}[\|\mathbf{y}\|^4]^{1/4}$, it suffices to show that (B.7) holds for each term. Note that $\mathbb{E}[c_{i+1}c_j] \leq R_c^2$ and $\mathbb{E}[c_i c_j] \leq R_c^2$ from Assumption 1, $|a_i(b)| \leq C_H$ from Lemma 20, and $\mathbb{E}_{\mathbf{w}}[\|\mathbf{w}_{1:r}\|^{4k}] = O((r/d)^{2k})$ from Lemma 12. Moreover, it is known that $\mathbb{E}_{z \sim \chi_r}[z^{2l}] = \Theta(r^l)$. Putting these things together yields the assertion. \square

It is also needed in later stages to give a high probability upper bound on the $\|\mathbf{g}(\mathbf{w}, b)\|$;

Lemma 24.

$$\sup_{b \in [-1, 1]} \|\mathbf{g}(\mathbf{w}, b)\| = \tilde{O}_{d,r} \left(\sqrt{\frac{1}{r d^{2Q-1}}} \right)$$

holds with high probability over $\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})$.

Proof. This can be shown similarly to Lemma 23 – we use Lemma 11 instead of Lemma 12 to obtain a high probability bound (note that this bound includes not only $s(\mathbf{w}, b)$ but also the main term, and it is for $\mathbf{g}(\mathbf{w}, b)$, instead of $\mathbf{g}(\mathbf{w}, b)_{1:r}$). \square

B.2 Concentration of Empirical Gradient

Now we control the deviation of $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})$ from $\mathbf{g}(\mathbf{w}, b)$.

Lemma 25. Under Assumption 1,

$$\sup_{f_*, b \in [-1, 1]} \|\mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})\sigma'_b(\mathbf{w}^\top \mathbf{x})]\| \lesssim \sqrt{\left(\frac{r}{d} \log d\right)^{Q-1}}, \quad (\text{B.8})$$

$$\sup_{f_*, b \in [-1, 1]} \|\mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})\sigma'_b(\mathbf{w}^\top \mathbf{x})]_{1:r}\| \lesssim \sqrt{\left(\frac{r}{d} \log d\right)^{Q-1}}, \quad (\text{B.9})$$

$$\sup_{f_*, b \in [-1, 1]} |\mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})\sigma_b(\mathbf{w}^\top \mathbf{x})]| \lesssim \sqrt{\left(\frac{r}{d} \log d\right)^Q} \quad (\text{B.10})$$

holds with high probability over $\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})$, respectively. Here \sup_{f_*} denotes the supremum over the support of f_* whose distribution is specified in Assumption 1.

Proof. We bound the leading term $\frac{a_Q(b)\mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes(Q-1)})}{(Q-1)!}$ in (B.3) of Lemma 19 to show (B.8). From Lemmas 5 and 6, we have

$$\left\| \frac{a_Q(b)\mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes(Q-1)})}{(Q-1)!} \right\|^2 = \left\| \frac{a_Q(b)c_Q}{(Q-1)!} (\boldsymbol{\beta})^{\otimes Q} (\mathbf{w}^{\otimes(Q-1)}) \right\|^2$$

$$\begin{aligned}
&= \left\| \frac{a_Q(b)c_Q}{(Q-1)!} (\boldsymbol{\beta})^{\otimes Q} (\mathbf{w}_{1:r}^{\otimes(Q-1)}) \right\|^2 \\
&\leq \left| \frac{a_Q(b)c_Q}{(Q-1)!} \right|^2 \|\boldsymbol{\beta}\|^{2Q} \|\mathbf{w}_{1:r}\|^{2(Q-1)} \\
&= \left| \frac{a_Q(b)c_Q}{(Q-1)!} \right|^2 \|\mathbf{w}_{1:r}\|^{2(Q-1)}.
\end{aligned}$$

Moreover, Lemma 20 tells us that $|a_Q(b)| \leq C_H$ always holds when $b \in [-1, 1]$, and Assumption 1 ensures that $|c_Q| \leq R_c$. Therefore, from Lemma 11, we can show that $\left\| \frac{a_Q(b)\mathbb{E}_{\mathbf{x}}[\nabla^Q f_*(\mathbf{x})](\mathbf{w}^{\otimes(Q-1)})}{(Q-1)!} \right\|^2 \leq \left(\frac{C_H R_c}{(Q-1)!} \right)^2 \left(\frac{4Cr}{d} \log d \right)^{Q-1}$ with probability at least $1 - 2\exp(-d/32) - 2rd^{-C}$. (B.9) and (B.10) can be obtained similarly. \square

The following lemma is parallel to Lemma 19 in [DLS22].

Lemma 26. Fix any $t \in [T_1]$, $b \in [-1, 1]$ and $\mathbf{w} \in \mathbb{S}^{d-1}$ satisfying the conditions (B.8) to (B.10). Then, with high probability over the distribution of $(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)$, it holds that

$$\begin{aligned}
&\left\| \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t - \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x}) \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \right\| \leq \tilde{O}\left(\sqrt{\frac{d}{N_1}}\right), \\
&\left\| \left(\frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t - \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x}) \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \right)_{1:r} \right\| \leq \tilde{O}\left(\sqrt{\frac{r}{N_1}}\right), \\
&\left| \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_b(\mathbf{w}^\top \mathbf{x}_i^t) - \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x}) \sigma_b(\mathbf{w}^\top \mathbf{x})] \right| \leq \tilde{O}\left(\sqrt{\frac{1}{N_1}}\right).
\end{aligned}$$

Here the right-hand sides do not depend on t .

Proof. We only present the proof of the first inequality, as other bounds can be attained similarly. From Corollary 17, $|f_*^t(\mathbf{x})| \leq R$ where $R \lesssim (\log d)^{P/2}$ holds with high probability for all $t \in [T_1]$. Conditioned on this,

$$\begin{aligned}
&\left\| \frac{1}{N_1} \sum_{i=1}^{N_1} f_*^t(\mathbf{x}_i^t) \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t - \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x}) \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \right\| \\
&\leq \left\| \frac{1}{N_1} \sum_{i=1}^{N_1} f_*^t(\mathbf{x}_i^t) \mathbb{I}_{f_*^t(\mathbf{x}_i^t) \leq R} \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t - \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x}) \mathbb{I}_{f_*^t(\mathbf{x}) \leq R} \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \right\| \\
&\quad + \left\| \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x}) \mathbb{I}_{f_*^t(\mathbf{x}) > R} \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \right\|
\end{aligned}$$

Note that $f_*^t(\mathbf{x}_i^t) \mathbb{I}_{f_*^t(\mathbf{x}_i^t) \leq R} \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \langle \mathbf{x}_i^t, \mathbf{u} \rangle - \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x}) \mathbb{I}_{f_*^t(\mathbf{x}) \leq R} \sigma'_b(\mathbf{w}^\top \mathbf{x}) \langle \mathbf{x}, \mathbf{u} \rangle]$ is R^2 -sub Gaussian for any $\|\mathbf{u}\| = 1$ from Lemma 14. Hence the first term can be upper bounded by $\tilde{O}\left(R\sqrt{\frac{d}{N_1}}\right)$ with high probability from Lemma 13. For the second term, we observe that

$$\begin{aligned}
\left\| \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x}) \mathbb{I}_{f_*^t(\mathbf{x}) > R} \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}] \right\| &\leq \mathbb{E}_{\mathbf{x}}[\|f_*^t(\mathbf{x}) \mathbb{I}_{f_*^t(\mathbf{x}) > R} \sigma'_b(\mathbf{w}^\top \mathbf{x}) \mathbf{x}\|] \\
&\leq \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x})^2]^{1/2} \mathbb{E}_{\mathbf{x}}[(\sigma'_b(\mathbf{w}^\top \mathbf{x}) \|\mathbf{x}\|)^2]^{1/2} \mathbb{E}_{\mathbf{x}}[(\mathbb{I}_{f_*^t(\mathbf{x}) > R})^4]^{1/4} \\
&\leq \mathbb{E}_{\mathbf{x}}[f_*^t(\mathbf{x})^2]^{1/2} \mathbb{E}_{\mathbf{x}}[\|\mathbf{x}\|^2]^{1/2} \mathbb{E}_{\mathbf{x}}[(\mathbb{I}_{f_*^t(\mathbf{x}) > R})]^{1/4} \\
&\leq O(\sqrt{d} \cdot d^{-C/4})
\end{aligned}$$

for sufficiently large C ; hence this term can be ignored. Finally, $\varsigma_i \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \langle \mathbf{x}_i^t, \mathbf{u} \rangle$ is τ^2 -sub Gaussian for any $\|\mathbf{u}\| = 1$ and then $\left\| \frac{1}{N_1} \sum_{i=1}^{N_1} \varsigma_i \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t \right\| = \tilde{O}\left(\sqrt{d/N_1}\right)$ with high probability. \square

By similar procedure, we can obtain a bound of $\|\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t\}_{t=1}^{T_1}) - \mathbf{g}(\mathbf{w}, b)\|$.

Lemma 27. Fix any $b \in [-1, 1]$ and $\mathbf{w} \in \mathbb{S}^{d-1}$ satisfying the conditions (B.8)-(B.10). Then, with high probability over the distribution of $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$,

$$\|\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) - \mathbf{g}(\mathbf{w}, b)\| = \tilde{O}\left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}}\sqrt{\frac{d}{T_1}} + \sqrt{\frac{d^2}{N_1 T_1}}\right),$$

$$\|(\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) - \mathbf{g}(\mathbf{w}, b))_{1:r}\| = \tilde{O}\left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}}\sqrt{\frac{r}{T_1}} + \sqrt{\frac{r^2}{N_1 T_1}}\right),$$

holds.

Proof. In this proof, the term ‘‘with high probability’’ indicates the probability with respect to the distribution of $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$. Let

$$\mathbf{z}_1(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) = \sum_{t=1}^{T_1} \frac{1}{T_1} y^t \sigma_b(\mathbf{w}^\top \mathbf{x}^t) \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_b'(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t$$

and

$$\mathbf{z}_2(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) = \sum_{t=1}^{T_1} \frac{1}{T_1} y^t \sigma_b'(\mathbf{w}^\top \mathbf{x}^t) \mathbf{x}^t \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_b(\mathbf{w}^\top \mathbf{x}_i^t).$$

Note that $\mathbf{g}_{T_1, N_1} = \mathbf{z}_1 + \mathbf{z}_2$ holds.

Let us consider $\mathbf{z}_1(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})$. Since we assume (B.8)-(B.10) hold, by taking union bound for Lemma 26 with respect to $t \in [T_1]$, with high probability $\|\frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_b'(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t\| \leq$

$R = \tilde{O}\left(\sqrt{\frac{d}{N_1}} + \sqrt{\frac{r^{Q-1}}{d^{Q-1}}}\right)$ for each t . Moreover, from Corollary 17, $|y^t| \leq R' = \tilde{O}(1)$ for each t .

Here, we notice that the procedure in the proof of Lemma 26 can be applied, yielding

$$\begin{aligned} & \|\mathbf{z}_1(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) - \mathbb{E}[\mathbf{z}_1(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})]\| \\ &= \tilde{O}\left(RR' \sqrt{d/T_1}\right) \\ &= \tilde{O}\left(\left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} + \sqrt{\frac{d}{N_1}}\right) \sqrt{\frac{d}{T_1}}\right) \end{aligned}$$

with high probability. Similarly we can obtain

- $\|\mathbf{z}_2 - \mathbb{E}[\mathbf{z}_2]\| = \tilde{O}\left(\left(\sqrt{\frac{r^Q}{d^Q}} + \sqrt{\frac{1}{N_1}}\right) \sqrt{\frac{d}{T_1}}\right),$
- $\|(\mathbf{z}_1 - \mathbb{E}[\mathbf{z}_1])_{1:r}\| = \tilde{O}\left(\left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} + \sqrt{\frac{r}{N_1}}\right) \sqrt{\frac{r}{T_1}}\right),$ and
- $\|(\mathbf{z}_2 - \mathbb{E}[\mathbf{z}_2])_{1:r}\| = \tilde{O}\left(\left(\sqrt{\frac{r^Q}{d^Q}} + \sqrt{\frac{1}{N_1}}\right) \sqrt{\frac{r}{T_1}}\right),$

which concludes the proof. \square

We need to obtain a bound uniformly over $b \in [-1, 1]$. We first introduce the following definition.

Definition 28. Fix any $\mathbf{w} \in \mathbb{S}^{d-1}$ and $X = \bigcup_{t=1}^{T_1} \{\mathbf{x}_1^t, \dots, \mathbf{x}_{N_1}^t, \mathbf{x}^t\} \subset \mathbb{R}^d$. Then, define a finite disjoint partition $\mathcal{B}(\mathbf{w}, X) = \{B_1, \dots, B_{N(\mathbf{w}, X)}\}$ of $[-1, 1]$, i.e., $\cup_i B_i = [-1, 1]$ and $B_i \cap B_j = \emptyset$ ($i \neq j$) as follows: b and b' belong to the same B_i if and only if $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b')$ for all $\mathbf{x} \in X$.

It is clear that $|\mathcal{B}(\mathbf{w}, X)| \lesssim |X| \leq (N_1 + 1)T_1$ holds because it suffices to divide $[-1, 1]$ at the point satisfying $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ for each $\mathbf{x} \in X$.

Lemma 29. Fix any $\mathbf{w} \in \mathbb{S}^{d-1}$. Then, with high probability over the distribution of $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$, the following holds:

- $g(\mathbf{w}, b)$ is L_1 -Lipschitz continuous with respect to b in the interval $[-1, 1]$.
- Let $X = \bigcup_{t=1}^{T_1} \{\mathbf{x}_1^t, \dots, \mathbf{x}_{N_1}^t, \mathbf{x}^t\}$. Then, $g_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})$ is L_2 -Lipschitz continuous with respect to b in each interval $B_i \in \mathcal{B}(\mathbf{w}, X)$.

Here $L_1 = O(1)$ and $L_2 = \tilde{O}(\sqrt{d})$ do not depend on \mathbf{w} .

Proof. By the explicit form of $g(\mathbf{w}, b)$ (B.6), we notice that each term is made by taking the product between the term as $a_j(b)a_k(b)$ and a vector independent of b whose norm is $O(1)$. Thus, it suffices to show that $a_j(b)a_k(b)$ is $O(1)$ -Lipschitz. This follows from the fact $a_j = \pm \frac{e^{-b^2/2}}{\sqrt{2\pi}} \text{He}_{j-2}(b)$ (see the proof of Lemma 20) and $\text{He}_{j-2}(b)e^{-b^2/2}$ is $O(1)$ -Lipschitz continuous on the bounded set $[-1, 1]$.

Next, let us see $g_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})$. Recall that

$$\begin{aligned} & g_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) \\ &:= \sum_{i=1}^{T_1} \frac{1}{T_1} y^t \sigma_b(\mathbf{w}^\top \mathbf{x}^t) \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma'_b(\mathbf{w}^\top \mathbf{x}_i^t) \mathbf{x}_i^t + \sum_{i=1}^{T_1} \frac{1}{T_1} y^t \sigma'_b(\mathbf{w}^\top \mathbf{x}^t) \mathbf{x}^t \frac{1}{N_1} \sum_{i=1}^{N_1} y_i^t \sigma_b(\mathbf{w}^\top \mathbf{x}_i^t). \end{aligned}$$

From Lemma 10 and Corollary 17, we know that $|y_i^t|, |y^t| \lesssim (\log d)^{P/2}$ and $\|\mathbf{x}_i^t\|, \|\mathbf{x}^t\| \lesssim \sqrt{d}$ holds for all i, t with high probability. Assuming this and noting that σ is 1-Lipschitz, in each interval in $\mathcal{B}(\mathbf{w}, X)$, $g_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})$ is $L_2 = \tilde{O}(\sqrt{d})$ -Lipschitz. \square

Corollary 30. With high probability over the distribution of \mathbf{w} and $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$,

$$\begin{aligned} \sup_{b \in [-1, 1]} \|g_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) - g(\mathbf{w}, b)\| &= \tilde{O} \left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{d}{T_1}} + \sqrt{\frac{d^2}{N_1 T_1}} \right), \\ \sup_{b \in [-1, 1]} \|(g_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) - g(\mathbf{w}, b))_{1:r}\| &= \tilde{O} \left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{r}{T_1}} + \sqrt{\frac{r^2}{N_1 T_1}} \right) \end{aligned} \quad (\text{B.11})$$

holds.

Proof. Let $X = \bigcup_{t=1}^{T_1} \{\mathbf{x}_1^t, \dots, \mathbf{x}_{N_1}^t, \mathbf{x}^t\}$ and consider $\mathcal{B}(\mathbf{w}, X) = \{B_1, \dots, B_{N(\mathbf{w}, X)}\}$. Then, let B be the union of $\tilde{O} \left(\left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{r}{T_1}} + \sqrt{\frac{r^2}{N_1 T_1}} \right) / \sqrt{d} \right)$ -coverings of $B_i \in \mathcal{B}(\mathbf{w}, X)$. Since $|B|$ is polynomial in d , Lemma 27 holds with high probability uniformly over $b \in B$.

Moreover, from the construction and Lemma 29, for $b \in [-1, 1] \setminus B$, there exists $\pi(b) \in B$ such that

$$\begin{aligned} & \|g_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) - g(\mathbf{w}, b)\| \\ & \quad - \|g_{T_1, N_1}(\mathbf{w}, \pi(b), \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) - g(\mathbf{w}, \pi(b))\| \\ & \leq \tilde{O} \left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{r}{T_1}} + \sqrt{\frac{r^2}{N_1 T_1}} \right) \end{aligned}$$

holds. \square

Lemma 31. With high probability over the distribution of $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$,

- It holds that

$$\begin{aligned} & \sup_{b \in [-1, 1]} \|g(\mathbf{w}, b) - g_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\| \\ &= \tilde{O} \left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{d}{T_1}} + \sqrt{\frac{d^2}{N_1 T_1}} \right) \end{aligned} \quad (\text{B.12})$$

with high probability over $\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})$.

• It holds that

$$\begin{aligned} & \sup_{b \in [-1, 1]} \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\|\mathbf{g}(\mathbf{w}, b) - \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\|^j]^{1/j} \\ &= \tilde{O} \left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{r}{T_1}} + \sqrt{\frac{r^2}{N_1 T_1}} \right) \end{aligned} \quad (\text{B.13})$$

for $j \in [4P]$.

Proof. If an event A occurs with probability at least $1 - d^{-C}$ over the simultaneous distribution of $(\mathbf{w}, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})$, then we can say that

- with probability at least $1 - d^{C/2}$ over $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$, $\mathbb{P}_{\mathbf{w}}\{A \mid \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}\} \geq 1 - d^{-C/2}$ holds.

(B.12) follows from this remark and Corollary 30.

For (B.13), with high probability over the distribution of $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$, $\mathbb{P}_{\mathbf{w}}\{(\text{B.11}) \mid \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}\} \geq 1 - d^{-\tilde{C}}$ is satisfied for some \tilde{C} . Also, from Lemma 10 and Corollary 17, with high probability over the distribution of $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$, we have $\|\mathbf{x}_i^t\|, \|\mathbf{x}^t\| = O(\sqrt{d})$ and $|y^t| = \tilde{O}(1)$. From the definition of $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})$ (B.2) and expansion of $\mathbf{g}(\mathbf{w}, b)$ (B.6), $\sup_{\mathbf{w}} \|\mathbf{g}(\mathbf{w}, b) - \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\| \leq \sup_{\mathbf{w}} \|\mathbf{g}(\mathbf{w}, b)\| + \sup_{\mathbf{w}} \|\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\|$ is upper bounded by $\tilde{O}(d)$. Therefore, with high probability

$$\begin{aligned} & \sup_{b \in [-1, 1]} \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\|\mathbf{g}(\mathbf{w}, b) - \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\|^j]^{1/j} \\ & \leq \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} \left[\left(\sup_{b \in [-1, 1]} \|\mathbf{g}(\mathbf{w}, b) - \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\| \right)^j \right]^{1/j} \\ & = \tilde{O} \left(\left((1 - d^{-\tilde{C}}) \left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{r}{T_1}} + \sqrt{\frac{r^2}{N_1 T_1}} \right)^j + (d^{-\tilde{C}}) d^j \right)^{1/j} \right). \end{aligned}$$

By setting sufficiently large \tilde{C} , we arrive at the claim. \square

Summary. We combine the obtained results and show that $\|\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\|$ and the moment of the residual terms are bounded with high probability.

Corollary 32. With high probability over the distribution of $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}$,

- (I) $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) = \mathbf{m}(\mathbf{w}, b) + \mathbf{r}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})$ holds where

$$\mathbf{m}(\mathbf{w}, b) := \frac{2a_Q(b)^2 \mathbb{E}_{c_Q}[c_Q^2]}{Q!(Q-1)!} \frac{(2Q-1)!!}{\mathbb{E}_{z \sim \chi_r}[z^{2Q}]} \|\mathbf{w}_{1:r}\|^{2Q-2} \begin{bmatrix} \mathbf{w}_{1:r} \\ \mathbf{0}_{d-r} \end{bmatrix},$$

and

$$\begin{aligned} & \sup_{b \in [-1, 1]} \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} \left[\left\| \left(\mathbf{r}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}) \right)_{1:r} \right\|^j \right]^{1/j} \\ & = \tilde{O}_{d,r} \left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{r}{T_1}} + \sqrt{\frac{r^2}{N_1 T_1}} + \sqrt{\frac{r}{d^{2Q+1}}} \right). \end{aligned}$$

(II) If $\eta_1 \gamma \leq \sqrt{rd^{2Q-1}}/C_1(\log d)^{C_2}$ for sufficiently large C_1 and C_2 , $T_1 \gtrsim r^Q d^{Q+1}$, and $N_1 T_1 \gtrsim rd^{2Q+1}$, then

$$\sup_{b \in [-1, 1]} \|2\eta_1 \gamma \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\| \leq 1$$

holds with high probability over $\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})$. Moreover,

$$\sup_{b \in [-1, 1]} \left| \left\langle 2\eta_1 \gamma \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}, \mathbf{z}) \right\rangle \right| \leq \log d$$

holds with high probability over $\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})$ and the training data for the second stage $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=T_1+1}^{T_2}$, where $\mathbf{z} \in \bigcup_{t=T_1+1}^{T_2} \{\mathbf{x}_1^t, \dots, \mathbf{x}_{N_2}^t, \mathbf{x}^t\}$.

Proof. Combining Lemmas 23 and 31 yields (I). For (II), by the condition for T_1 and N_1 , we obtain

$$\begin{aligned} \sup_{b \in [-1, 1]} \|\mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\| &\leq \tilde{O}_{d,r} \left(\sqrt{\frac{r^{Q-1}}{d^{Q-1}}} \sqrt{\frac{d}{T_1}} + \sqrt{\frac{d^2}{N_1 T_1}} + \sqrt{\frac{1}{rd^{2Q-1}}} \right) \\ &= \tilde{O}_{d,r} \left(\sqrt{\frac{1}{rd^{2Q-1}}} \right) \end{aligned}$$

from Lemmas 24 and 31. Then, as $\eta_1 \gamma \lesssim \sqrt{rd^{2Q-1}}/C_1(\log d)^{C_2}$, $\sup_{b \in [-1, 1]} \|2\eta_1 \gamma \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1})\| \leq 1$ is achieved.

By an argument similar to the one in the proof of Lemma 18, we obtain that with high probability, $\left| \left\langle 2\eta_1 \gamma \mathbf{g}_{T_1, N_1}(\mathbf{w}, b, \{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)\}_{t=1}^{T_1}, \mathbf{z}) \right\rangle \right| \lesssim \log d$ for $\mathbf{z} \in \bigcup_{t=T_1+1}^{T_2} \{\mathbf{x}_1^t, \dots, \mathbf{x}_{N_2}^t, \mathbf{x}^t\}$. To obtain the supremum bound over $b \in [-1, 1]$, it suffices to utilize Lemma 29 again. \square

In the following sections, we assume that these two events (I) and (II) occur.

C Construction of Attention Matrix

We construct an attention matrix $\bar{\Gamma}$ with a good approximation property: see Section 3.2.2 for the proof outline.

In this section, a set of basis functions \mathcal{H} is defined as

$$\mathcal{H} = \left\{ \mathbf{x} \mapsto \prod_{j=1}^r \frac{1}{\sqrt{j!}} \text{He}_{p_j}(x_j) \mid Q \leq p_1 + \dots + p_r \leq P, p_1 \geq 0, \dots, p_r \geq 0 \right\}.$$

We let $B_P := |\mathcal{H}|$ and introduce a numbering $\mathcal{H} = \{h_1, \dots, h_{B_P}\}$. Note that

$$B_P = \sum_{i=Q}^P \binom{r+i-1}{i} = \Theta(r^P).$$

C.1 Approximation Error of Two-layer Neural Network

As sketched in Section 3.2.2, we need to show that there is a set of two-layer neural networks which can approximate basis functions well. The goal of this section is to show the following proposition.

Proposition 33. Assume $T_1 = \tilde{\Omega}(d^{Q+1}r^Q)$ and $N_1 T_1 = \tilde{\Omega}(d^{2Q+1}r)$. Let $\{\mathbf{w}_j^{(1)}\}_j$ be neurons obtained by line 4 of Algorithm 1 with $\gamma \asymp \frac{1}{m^{\frac{3}{2}} r^{\frac{1}{2}} d^Q}$, $\eta_1 \asymp m^{\frac{3}{2}} r d^{2Q-\frac{1}{2}} \cdot (\log d)^{-C_n}$ for constant C_n . Assume \mathbf{b} is re-initialized in Line 5 of Algorithm 1. Then, with high probability over the distribution of training data and the initialization of model parameters, there exist vectors $\mathbf{a}^1, \dots, \mathbf{a}^{B_P} \in \mathbb{R}^m$ such that for each $n \in [B_P]$,

$$\sup_{i \in [N_2], T_1+1 \leq t \leq T_2} \left(\sum_{j=1}^m a_j^n \sigma(\mathbf{w}_j^{(1)\top} \mathbf{x}_i^t + b_j) - h_n(\mathbf{x}_i^t) \right)^2 = \tilde{O} \left(\frac{r^P}{m} + \frac{1}{N_2} \right),$$

$$\sup_{T_1+1 \leq t \leq T_2} \left(\sum_{j=1}^m a_j^n \sigma(\mathbf{w}_j^{(1)\top} \mathbf{x}^t + b_j) - h_n(\mathbf{x}^t) \right)^2 = \tilde{O}\left(\frac{r^P}{m} + \frac{1}{N_2}\right)$$

holds. Moreover, $\sup_{n \in [B_P]} \|\mathbf{a}^n\|^2 = \tilde{O}\left(\frac{r^P}{m}\right)$.

This proposition states that there exists a two-layer neural network $\sum_{j=1}^m a_j^n \sigma(\langle \mathbf{w}_j^{(1)}, \cdot \rangle + b_j)$ which can approximate the basis function h_n , and its second-layer magnitude $\|\mathbf{a}^n\|^2$ only scales with r . Note that the condition of T_1 and $N_1 T_1$ comes from Corollary 32, which implies that the noise terms of the one-step gradient is well controlled.

C.1.1 Alignment to \mathcal{S} and Efficient Approximation

From now on, we suppose that $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)$ satisfies the conditions in Corollary 32 (to simplify the notation, we drop the dependency on $\{(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^{T_1}$). An important previous result [DLS22] is that, if $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)$ aligns with the r -dimensional subspace \mathcal{S} , then it can be used to approximate polynomials on \mathcal{S} efficiently.

Lemma 34 (Lemma 21 in [DLS22]). *There exists a set of symmetric r -dimensional tensors $\{\mathbf{T}_k^n\}_{0 \leq k \leq P, n \in [B_P]}$ such that*

$$h_n(\mathbf{x}) = \sum_{0 \leq k \leq P} \langle \mathbf{T}_k^n, \mathbf{x}_{1:r}^{\otimes k} \rangle$$

and $\|\mathbf{T}_k^n\|_F \lesssim r^{\frac{P-k}{4}}$.

Definition 35 ($f(r, d)$ -alignment). *We call that $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)$ has $f(r, d)$ -alignment with \mathcal{S} , if the following holds:*

- For any $0 \leq k \leq P$ -symmetric r -dimensional tensor \mathbf{T} ,

$$\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\langle \mathbf{T}, \mathbf{g}_{T_1, N_1}(\mathbf{w}, b)_{1:r}^{\otimes k} \rangle^2] \gtrsim f(r, d)^k \|\mathbf{T}\|_F^2,$$

and the hidden constant in “ \gtrsim ” needs to be uniformly lower bounded.

Lemma 36 (Corollary 4 in [DLS22], adapted). *Suppose that $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)$ has $f(r, d)$ -alignment with \mathcal{S} . Let \mathbf{T} be a $0 \leq k \leq P$ -symmetric r -dimensional tensor. Then, there exists $\psi_{\mathbf{T}}(\mathbf{w}, b)$ such that*

$$\begin{aligned} \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\psi_{\mathbf{T}}(\mathbf{w}, b) \langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b), \mathbf{x} \rangle^k] &= \langle \mathbf{T}, \mathbf{x}_{1:r}^{\otimes k} \rangle, \\ \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\psi_{\mathbf{T}}(\mathbf{w}, b)^2] &\lesssim f(r, d)^{-k} \|\mathbf{T}\|_F^2, \\ |\psi_{\mathbf{T}}(\mathbf{w}, b)| &\lesssim f(r, d)^{-k} \|\mathbf{T}\|_F \|\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)\|^k. \end{aligned}$$

Intuitively speaking, $f(r, d)$ -alignment ensures that the MLP gradient $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)$ has sufficiently large component on \mathcal{S} . If so, the lemma states that, an infinite-width neural network $\mathbb{E}_{\mathbf{w}} [\psi_{\mathbf{T}}(\mathbf{w}, b) \langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b), \cdot \rangle^k]$, whose hidden layer weight is $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)$ and the activation is $z \mapsto z^k$, can express the functions $\langle \mathbf{T}, \mathbf{x}_{1:r}^{\otimes k} \rangle$ with small output layer $\psi_{\mathbf{T}}(\mathbf{w}, b)$.

Let us show that our $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)$ has sufficient alignment if b satisfies (B.5).

Lemma 37 (Tensor expectation lower bound). *Let \mathbf{T} be a $k \leq P$ -symmetric r -dimensional tensor and $\bar{\mathbf{w}} = \mathbf{w}_{1:r} \|\mathbf{w}_{1:r}\|^{2Q-2}$. Then*

$$\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\mathbf{T}(\bar{\mathbf{w}}^{\otimes k})^2] \gtrsim \frac{r^{Qi}}{d^{(Q+1)i}} \mathbb{E}[\|\mathbf{T}(\bar{\mathbf{w}}^{\otimes k-i})\|_F^2]$$

holds for all $0 \leq i \leq k$. Here $\mathbf{T}(\bar{\mathbf{w}}^{\otimes 0}) := \mathbf{T}$.

Proof. Let $\mathbf{u} \sim \mathcal{N}(0, I_d)$, $z \sim \chi(d)$ and $\bar{\mathbf{u}} = \mathbf{u}_{1:r} \|\mathbf{u}_{1:r}\|^{2Q-2}$. We can decompose $\bar{\mathbf{u}}$ as $\bar{\mathbf{u}} = z^{2Q-1} \bar{\mathbf{w}}$. Therefore

$$\mathbb{E}[\mathbf{T}(\bar{\mathbf{u}}^{\otimes k})^2] = \mathbb{E}[z^{(4Q-2)k}] \mathbb{E}[\mathbf{T}(\bar{\mathbf{w}}^{\otimes k})^2]$$

holds. On the other hand, let $\mathbf{x} \sim \mathbb{S}^{r-1}$ and $z' \sim \chi(r)$. Then we can decompose $\bar{\mathbf{u}}$ as $\bar{\mathbf{u}} = (z')^{2Q-1}\mathbf{x}$. Thus

$$\mathbb{E}[\mathbf{T}(\bar{\mathbf{u}}^{\otimes k})^2] = \mathbb{E}[(z')^{(4Q-2)k}] \mathbb{E}[\mathbf{T}(\mathbf{x}^{\otimes k})^2]$$

holds. It implies that

$$\mathbb{E}[\mathbf{T}(\bar{\mathbf{w}}^{\otimes k})^2] = \frac{\mathbb{E}[(z')^{(4Q-2)k}]}{\mathbb{E}[z^{(4Q-2)k}]} \mathbb{E}[\mathbf{T}(\mathbf{x}^{\otimes k})^2].$$

Similarly,

$$\mathbb{E}[\|\mathbf{T}(\bar{\mathbf{w}}^{\otimes k-i})\|_F^2] = \frac{\mathbb{E}[(z')^{(4Q-2)(k-i)}]}{\mathbb{E}[z^{(4Q-2)(k-i)}]} \mathbb{E}[\|\mathbf{T}(\mathbf{x}^{\otimes(k-i)})\|_F^2]$$

is satisfied. Now from Corollary 13 in [DLS22],

$$\mathbb{E}[\|\mathbf{T}(\mathbf{x}^{\otimes(k-i)})\|_F^2] \lesssim r^i \mathbb{E}[\mathbf{T}(\mathbf{x}^{\otimes k})^2]$$

holds. Therefore we obtain

$$\begin{aligned} \mathbb{E}[\mathbf{T}(\bar{\mathbf{w}}^{\otimes k})^2] &= \frac{\mathbb{E}[(z')^{(4Q-2)k}]}{\mathbb{E}[z^{(4Q-2)k}]} \mathbb{E}[\mathbf{T}(\mathbf{x}^{\otimes k})^2] \\ &\gtrsim r^{-i} \frac{\mathbb{E}[(z')^{(4Q-2)k}]}{\mathbb{E}[z^{(4Q-2)k}]} \mathbb{E}[\|\mathbf{T}(\mathbf{x}^{\otimes(k-i)})\|_F^2] \\ &= r^{-i} \frac{\mathbb{E}[(z')^{(4Q-2)k}]}{\mathbb{E}[z^{(4Q-2)k}]} \frac{\mathbb{E}[z^{(4Q-2)(k-i)}]}{\mathbb{E}[(z')^{(4Q-2)(k-i)}]} \mathbb{E}[\|\mathbf{T}(\bar{\mathbf{w}}^{\otimes k-i})\|_F^2] \\ &= \frac{r^{(2Q-2)i}}{d^{(2Q-1)i}} \mathbb{E}[\|\mathbf{T}(\bar{\mathbf{w}}^{\otimes k-i})\|_F^2]. \end{aligned}$$

□

Corollary 38. Let \mathbf{T} be a $k \leq P$ -symmetric r -dimensional tensor and let $\mathbf{m}(\mathbf{w}, b)_{1:r} = \frac{2a_Q(b)^2}{Q!(Q-1)!} \frac{\mathbb{E}_{c_Q}[c_Q^2]^{(2Q-1)!!}}{\mathbb{E}_{z \sim \chi_r}[z^{2Q}]}$ $\|\mathbf{w}_{1:r}\|^{2Q-2} \mathbf{w}_{1:r}$. Moreover, assume that (B.5) holds. Then,

$$\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})}[\langle \mathbf{T}, \mathbf{m}(\mathbf{w}, b)_{1:r}^{\otimes k} \rangle^2] \gtrsim \frac{1}{d^{(2Q-1)i} r^{2i}} \mathbb{E}[\|\mathbf{T}(\mathbf{m}(\mathbf{w}, b)_{1:r}^{\otimes k-i})\|_F^2]$$

holds for all $0 \leq i \leq k$.

Proof. This is obtained by calibrating the coefficients from the previous lemma. □

Lemma 39. Let \mathbf{T} be a $0 \leq k \leq P$ -symmetric r -dimensional tensor. Assume that (B.5) holds. Furthermore, suppose $r \lesssim d^{\frac{1}{2}}$, $T_1 = \tilde{\Omega}(d^{Q+1} r^Q)$ and $N_1 T_1 = \tilde{\Omega}(d^{2Q+1} r)$. Then,

$$\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})}[\langle \mathbf{T}, \mathbf{g}_{T_1, N_1}(\mathbf{w}, b)_{1:r}^{\otimes k} \rangle^2] \gtrsim \frac{1}{d^{(2Q-1)k} r^{2k}} \|\mathbf{T}\|_F^2$$

holds for all $0 \leq k \leq P$. Therefore, $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b)$ has $d^{-2Q+1} r^{-2}$ -alignment with \mathcal{S} .

Proof. Decompose $\mathbf{g}_{T_1, N_1}(\mathbf{w}, b) = \mathbf{m}(\mathbf{w}, b) + \mathbf{r}(\mathbf{w}, b)$ as defined in Corollary 32. From the proof of Lemma 11 in [DLS22], we can show that

$$\begin{aligned} &\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})}[\langle \mathbf{T}, \mathbf{g}_{T_1, N_1}(\mathbf{w}, b)_{1:r}^{\otimes k} \rangle^2] \\ &\geq \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} \left[\frac{1}{2} \langle \mathbf{T}, \mathbf{m}(\mathbf{w}, b)_{1:r}^{\otimes k} \rangle^2 \right] - \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})}[\delta(\mathbf{w}, b)^2] \end{aligned} \quad (\text{C.1})$$

for some $\delta(\mathbf{w}, b)$ satisfying

$$\begin{aligned} &\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})}[\delta(\mathbf{w}, b)^2] \\ &\lesssim \sum_{i=1}^k \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})}[\|\mathbf{T}(\mathbf{m}(\mathbf{w}, b)_{1:r}^{\otimes k-i})\|_F^2] \sqrt{\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})}[\|\mathbf{r}(\mathbf{w}, b)_{1:r}\|^{4i}]}. \end{aligned}$$

Here, from Corollaries 32 and 38,

$$\begin{aligned} & \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\delta(\mathbf{w}, b)^2] \\ & \lesssim \sum_{i=1}^k d^{(2Q-1)i} r^{2i} \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\langle \mathbf{T}, \mathbf{m}(\mathbf{w}, b)_{1:r}^{\otimes k} \rangle^2] \tilde{O}_{d,r} \left(\left(\frac{r^{Q-1}}{d^{Q-1}} \frac{r}{T_1} + \frac{r^2}{N_1 T_1} + \frac{r}{d^{2Q+1}} \right)^i \right) \end{aligned}$$

holds. Now, from the condition of N_1 and T_1 , and $r \lesssim \sqrt{d}$, we know

$$d^{(2Q-1)} r^2 \left(\frac{r^{Q-1}}{d^{Q-1}} \frac{r}{T_1} + \frac{r^2}{N_1 T_1} + \frac{r}{d^{2Q+1}} \right) \lesssim 1.$$

Therefore we can achieve $\mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\delta(\mathbf{w}, b)^2] \leq \frac{1}{4} \mathbb{E}_{\mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1})} [\langle \mathbf{T}, \mathbf{m}(\mathbf{w}, b)_{1:r}^{\otimes k} \rangle^2]$. Applying this and Corollary 38 ($i = k$) to (C.1) yields the result. \square

C.1.2 Approximation Power of Infinite-width Neural Network

Now we can show that some ‘‘infinite-width’’ neural network can approximate the functions $h \in \mathcal{H}$.

Lemma 40 (Lemma 9 in [DLS22], adapted). *Let $a \sim \text{Unif}(\{-1, 1\})$ and $b \sim \text{Unif}([-\log d, \log d])$. Then, for any $k \geq 0$ there exists $v_k(a, b)$ such that for $|x| \leq \log d$,*

$$\mathbb{E}_{a,b} [v_k(a, b) \sigma(ax + b)] = x^k, \sup_{a,b} |v_k(a, b)| = \tilde{O}_d(1).$$

Lemma 41 (Lemma 12 and Corollary 5 in [DLS22], adapted). *There exists $\phi_n(a, \mathbf{w}, b_1, b_2)$ such that if*

$$f_{\phi_n}(\mathbf{x}) := \mathbb{E}_{a, \mathbf{w}, b_1, b_2} [\phi_n(a, \mathbf{w}, b_1, b_2) \sigma(2\eta_1 a \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1) + b_2)]$$

where

$$a \sim \text{Unif}\{\pm\gamma\}, \mathbf{w} \sim \text{Unif}(\mathbb{S}^{d-1}), b_1 \sim \text{Unif}([-1, 1]), b_2 \sim \text{Unif}([-\log d, \log d]),$$

it holds that

$$\begin{aligned} \sup_{i \in [N_2], T_1+1 \leq t \leq T_2} (f_{\phi_n}(\mathbf{x}_i^t) - h_n(\mathbf{x}_i^t))^2 & \lesssim \frac{1}{N_2}, \\ \sup_{T_1+1 \leq t \leq T_2} (f_{\phi_n}(\mathbf{x}^t) - h_n(\mathbf{x}^t))^2 & \lesssim \frac{1}{N_2}, \\ \mathbb{E}_{a, \mathbf{w}, b_1, b_2} [\phi_n(a, \mathbf{w}, b_1, b_2)^2] & = \tilde{O}(r^P), \\ \sup_{a, \mathbf{w}, b_1, b_2} |\phi_n(a, \mathbf{w}, b_1, b_2)| & = \tilde{O}(r^P) \end{aligned}$$

for each $n \in [B_P]$.

Proof. We first construct $\phi_{\mathbf{T}_k^n}(a, \mathbf{w}, b_1, b_2)$ for basis tensors defined in Lemma 34, such that if

$$f_{\phi_{\mathbf{T}_k^n}}(\mathbf{x}) := \mathbb{E}_{a, \mathbf{w}, b_1, b_2} [\phi_{\mathbf{T}_k^n}(a, \mathbf{w}, b_1, b_2) \sigma(\langle 2\eta_1 a \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle + b_2)],$$

it holds that

$$\sup_{i \in [N_2], T_1+1 \leq t \leq T_2} \left(f_{\phi_{\mathbf{T}_k^n}}(\mathbf{x}_i^t) - \langle \mathbf{T}_k^n, (\mathbf{x}_i^t)_{1:r}^{\otimes k} \rangle \right)^2 \lesssim \frac{1}{N_2}, \quad (\text{C.2})$$

$$\sup_{T_1+1 \leq t \leq T_2} \left(f_{\phi_{\mathbf{T}_k^n}}(\mathbf{x}^t) - \langle \mathbf{T}_k^n, (\mathbf{x}^t)_{1:r}^{\otimes k} \rangle \right)^2 \lesssim \frac{1}{N_2}. \quad (\text{C.3})$$

Let $\phi_{\mathbf{T}_k^n}(a, \mathbf{w}, b_1, b_2) = \frac{2v_k(a, b_2)\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1)}{(2\eta_1)^k \gamma^k} \mathbb{I}_{A(\mathbf{w})} \mathbb{I}_{B(b_1)}$ where v_k and $\psi_{\mathbf{T}_k^n}$ are constructed in Lemma 36 and the events $A(\mathbf{w})$ and $B(b_1)$ are defined as

$$A(\mathbf{w}) := \left\{ \sup_{b_1 \in [-1, 1]} 2\eta_1 \gamma |\langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x}_i^t \rangle|, \sup_{b_1 \in [-1, 1]} 2\eta_1 \gamma |\langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x}^t \rangle| \leq \log d \right.$$

$$\begin{aligned}
& \left. \text{for } i \in [N_2], T_1 + 1 \leq t \leq T_2 \right\} \\
& \cap \left\{ \sup_{b_1 \in [-1, 1]} \|2\eta_1 \gamma \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1)\| \leq 1 \right\} \\
B(b_1) & := \left\{ |\mathbb{E}_{z \sim \mathcal{N}(0, 1)}[\sigma_{b_1}(z) \text{He}_i(z)]| \leq C_H \ (2 \leq i \leq P + 2) \right\} \\
& \cap \left\{ |\mathbb{E}_{z \sim \mathcal{N}(0, 1)}[\sigma_{b_1}(z) \text{He}_Q(z)]| \geq C_L \right\}.
\end{aligned}$$

Then,

$$\begin{aligned}
& f_{\phi_{\mathbf{T}_k^n}}(\mathbf{x}) \\
& = \mathbb{E}_{a, \mathbf{w}, b_1, b_2} \left[\frac{2v_k(a, b_2) \psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1)}{(2\eta_1)^k \gamma^k} \mathbb{I}_{A(\mathbf{w})} \mathbb{I}_{B(b_1)} \sigma(\langle 2\eta_1 a \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle + b_2) \right] \\
& = \mathbb{E}_{b_1} \left[2 \cdot \mathbb{I}_{B(b_1)} \mathbb{E}_{\mathbf{w}} \left[\frac{\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1)}{(2\eta_1)^k \gamma^k} \mathbb{I}_{A(\mathbf{w})} \mathbb{E}_{a', b_2} [v_k(a, b_2) \sigma(\langle 2\eta_1 a' \gamma \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle + b_2)] \right] \right],
\end{aligned}$$

where $a' \sim \text{Unif}(\{\pm 1\})$. Here, if $\mathbf{x} \in \bigcup_{t=T_1+1}^{T_2} \{\mathbf{x}_1^t, \dots, \mathbf{x}_{N_2}^t, \mathbf{x}^t\}$,

$$\begin{aligned}
& \mathbb{E}_{\mathbf{w}} \left[\frac{\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1)}{(2\eta_1)^k \gamma^k} \mathbb{I}_{A(\mathbf{w})} \mathbb{E}_{a' \sim \text{Unif}(\{\pm 1\}), b_2} [v_k(a, b_2) \sigma(\langle 2\eta_1 a' \gamma \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle + b_2)] \right] \\
& = \mathbb{E}_{\mathbf{w}} \left[\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1) \langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle^k \right] + \mathbb{E}_{\mathbf{w}} \left[\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1) (1 - \mathbb{I}_{A(\mathbf{w})}) \langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle^k \right].
\end{aligned}$$

Note that the absolute value of second term above can be upper bounded by $\mathbb{E}_{\mathbf{w}} [(1 - \mathbb{I}_{A(\mathbf{w})})]^{1/2} \mathbb{E}_{\mathbf{w}} \left[(\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1) \langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle^k)^2 \right]^{1/2}$. Recall that $A(\mathbf{w})$ occurs

with high probability from Corollary 32, and $\mathbb{E}_{\mathbf{w}} \left[(\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1) \langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle^k)^2 \right]^{1/2}$ is polynomial in d . Hence we can set the second term to be $O(1/\sqrt{N_2})$.

Moreover, as $\mathbb{P}[B(b_1)] = 1/2$ from Lemma 20, we have

$$\begin{aligned}
& \mathbb{E}_{b_1} \left[2 \cdot \mathbb{I}_{B(b_1)} \mathbb{E}_{\mathbf{w}} \left[\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1) \langle \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1), \mathbf{x} \rangle^k \right] \right] \\
& = \mathbb{E}_{b_1} \left[2 \cdot \langle \mathbf{T}_k^n, \mathbf{x}^{\otimes k} \rangle \right] + \mathbb{E}_{b_1} \left[2(1 - \mathbb{I}_{B(b_1)}) \langle \mathbf{T}_k^n, \mathbf{x}^{\otimes k} \rangle \right] = \langle \mathbf{T}_k^n, \mathbf{x}^{\otimes k} \rangle.
\end{aligned}$$

Therefore, (C.2) and (C.3) holds. Noting that $\eta_1 \gamma \asymp \sqrt{rd^{2Q-1}}/C_1 (\log d)^{C_2}$, We obtain

$$\begin{aligned}
& \mathbb{E}_{a, \mathbf{w}, b_1, b_2} [\phi_{\mathbf{T}_k^n}(a, \mathbf{w}, b_1, b_2)^2] \\
& \leq \frac{4}{(2\eta_1)^{2k} \gamma^{2k}} \mathbb{E}_{a, b_2} [v_k(a, b_2)^2] \mathbb{E}_{\mathbf{w}, b_1} [\mathbb{I}_{B(b_1)} \psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1)^2] \\
& \lesssim \frac{4}{(2\eta_1)^{2k} \gamma^{2k}} \mathbb{E}_{a, b_2} [v_k(a, b_2)^2] (d^{(2Q-1)} r^2)^k \|\mathbf{T}_k^n\|_F^2 \\
& = \tilde{O} \left(r^k r^{\frac{P-k}{2}} \right)
\end{aligned}$$

and

$$\begin{aligned}
& \sup_{a, \mathbf{w}, b_1, b_2} |\phi_{\mathbf{T}_k^n}(a, \mathbf{w}, b_1, b_2)| \\
& \leq \frac{2}{(2\eta_1)^k \gamma^k} |v_k(a, b_2)| \mathbb{I}_{A(\mathbf{w})} \mathbb{I}_{B(b_1)} |\psi_{\mathbf{T}_k^n}(\mathbf{w}, b_1)| \\
& \lesssim \frac{2}{(2\eta_1)^{2k} \gamma^{2k}} (d^{(2Q-1)} r^2)^k \|\mathbf{T}_k^n\|_F \|(2\eta_1 \gamma) \mathbf{g}_{T_1, N_1}(\mathbf{w}, b_1)\|^k \mathbb{I}_{A(\mathbf{w})} \mathbb{I}_{B(b_1)} \\
& \lesssim \tilde{O} \left(r^k r^{\frac{P-k}{4}} \right)
\end{aligned}$$

from Lemmas 36 and 39. To show the original claim, it suffices to set $\phi_n(a, \mathbf{w}, b_1, b_2) = \sum_{k=0}^P \phi_{\mathbf{T}_k^n}(a, \mathbf{w}, b_1, b_2)$, which yields the upper bounds on the magnitude of ϕ_n because $\sum_{k=0}^P r^k r^{\frac{P-k}{2}} \lesssim r^P$ and $\sum_{k=0}^P r^k r^{\frac{P-k}{4}} \lesssim r^P$. \square

C.1.3 Proof of Proposition 33

We discretize Lemma 41 to establish Proposition 33.

Proof. [Proof of Proposition 33] First, in the same discretization manner as the proof of Lemma 13 in [DLS22], we can show that there exists $\mathbf{a}^1, \dots, \mathbf{a}^{B_P}$ such that for each $n \in [B_P]$,

$$\begin{aligned} & \sup_{i \in [N_2], T_1+1 \leq t \leq T_2} \left(\sum_{j=1}^m a_j^n \sigma \left(\left\langle 2\eta_1 \Gamma_{j,j}^{(0)} \mathbf{g}_{T_1, N_1}(\mathbf{w}_j^{(0)}, b_j^{(0)}), \mathbf{x}_i^t \right\rangle + b_j \right) - h_n(\mathbf{x}_i^t) \right)^2 \\ &= \tilde{O} \left(\frac{r^P}{m} + \frac{1}{N_2} \right), \\ & \sup_{T_1+1 \leq t \leq T_2} \left(\sum_{j=1}^m a_j^n \sigma \left(\left\langle 2\eta_1 \Gamma_{j,j}^{(0)} \mathbf{g}_{T_1, N_1}(\mathbf{w}_j^{(0)}, b_j^{(0)}), \mathbf{x}^t \right\rangle + b_j \right) - h_n(\mathbf{x}^t) \right)^2 \\ &= \tilde{O} \left(\frac{r^P}{m} + \frac{1}{N_2} \right) \end{aligned}$$

and $\sup_{n \in [B_P]} \|\mathbf{a}^n\|^2 = \tilde{O} \left(\frac{r^P}{m} \right)$ with high probability ($b_j^{(0)}$, b_j denotes the biases at the first and second initialization, respectively). Recall that $\mathbf{w}_j^{(1)} = 2\eta_1 \Gamma_{j,j}^{(0)} \mathbf{g}_{T_1, N_1}(\mathbf{w}_j^{(0)}, b_j^{(0)}) - 2\eta_1 \frac{1}{T_1} \sum_{t=1}^{T_1} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \Gamma^{(0)}, \mathbf{b}^{(0)}) \nabla_{\mathbf{w}_j^{(0)}} f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(0)}, \Gamma^{(0)}, \mathbf{b}^{(0)})$. It remains to bound the effect of the second term. By Lemma 18, we know that the norm of second term is $\tilde{O}(\eta_1 \gamma^2 m \sqrt{d})$. Similar to the argument in the proof of Lemma 18, the inner product between the second term and $\mathbf{x}^t, \mathbf{x}_i^t$ ($i \in [N_2], T_1 + 1 \leq t \leq T_2$) is also $\tilde{O}(\eta_1 \gamma^2 m \sqrt{d})$, with high probability. From the Cauchy-Schwarz inequality and the Lipschitz continuity of σ ,

$$\begin{aligned} & \left| \sum_{j=1}^m a_j^n \sigma \left(\left\langle 2\eta_1 \Gamma_{j,j}^{(0)} \mathbf{g}_{T_1, N_1}(\mathbf{w}_j^{(0)}, b_j^{(0)}), \mathbf{x} \right\rangle + b_j \right) - \sum_{j=1}^m a_j^n \sigma \left(\left\langle \mathbf{w}_j^{(1)}, \mathbf{x} \right\rangle + b_j \right) \right| \\ &= \tilde{O} \left(\sqrt{\frac{r^P}{m}} \eta_1 \gamma^2 m \sqrt{d} \sqrt{m} \right) = \tilde{O} \left(\sqrt{\frac{r^P}{m}} \right). \end{aligned}$$

We obtain the assertion. \square

C.2 Concentration of Correlation

Next, we evaluate the discrepancy between the empirical correlation $\frac{1}{N_2} \sum_{j=1}^{N_2} y_j h(\mathbf{x}_j)$ and its expectation for each $h \in \mathcal{H}$, which decides the approximation error in (b) of (3.1).

Recall that $h \in \mathcal{H}$ can be written in the form as

$$h(\mathbf{x}) = \prod_{i=1}^r \frac{1}{\sqrt{p_i!}} \text{He}_{p_i}(x_i) \quad (Q \leq p_1 + \dots + p_r \leq P, p_1 \geq 0, \dots, p_r \geq 0).$$

Also note that

$$\begin{aligned} & \frac{1}{N_2} \sum_{j=1}^{N_2} y_j h(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}, \varsigma} [y h(\mathbf{x})] \\ &= \frac{1}{N_2} \sum_{j=1}^{N_2} \varsigma_j h(\mathbf{x}_j) + \frac{1}{N_2} \sum_{j=1}^{N_2} \sigma_*(\langle \mathbf{x}_j, \boldsymbol{\beta} \rangle) h(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}} [\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle) h(\mathbf{x})], \end{aligned}$$

where $\sigma_*(z) = \sum_{i=Q}^P \frac{c_i}{i!} \text{He}_i(z)$. First we give an upper bound for $|\frac{1}{N_2} \sum_{j=1}^{N_2} \sigma_*(\langle \mathbf{x}_j, \boldsymbol{\beta} \rangle) h(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}} [\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle) h(\mathbf{x})]|$.

Lemma 42. Let $D \leq 2P$ be the degree of $\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x})$, when we regard it as an r -variable polynomial with respect to (x_1, \dots, x_r) . For all $t > 0$ there exist absolute constants M and C_D depending only on D such that

$$\begin{aligned} & \mathbb{P} \left(\left| \frac{1}{N_2} \sum_{j=1}^{N_2} \sigma_*(\langle \mathbf{x}_j, \boldsymbol{\beta} \rangle)h(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}}[\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x})] \right| \geq t \right) \\ & \leq 2 \exp \left(-\frac{1}{C_D M^2} \min_{1 \leq k \leq D} \left(\frac{t \sqrt{N_2}}{\|\mathbb{E}[\nabla_{\mathbf{x}}^k(\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))]\|_F} \right)^{2/k} \right) \end{aligned}$$

holds, where C_D and M are constants.

Proof. Let $F(\mathbf{x}_1, \dots, \mathbf{x}_N) = \frac{1}{N_2} \sum_{j=1}^{N_2} \sigma_*(\langle \mathbf{x}_j, \boldsymbol{\beta} \rangle)h(\mathbf{x}_j)$. It is a degree- D polynomial for $x_{1,1}, \dots, x_{1,r}, \dots, x_{N_2,r}$, which are standard Gaussian variables. Also,

$$\|\mathbb{E}[\nabla^k F(\mathbf{x}_1, \dots, \mathbf{x}_N)]\|_F = \sqrt{N_2}^{-1} \|\mathbb{E}[\nabla^k(\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))]\|_F$$

is satisfied. Then, applying Lemma 15 to F yields the desired result. \square

Then, our aim is to bound $\|\mathbb{E}[\nabla^k(\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))]\|_F$.

Lemma 43. Let $h(\mathbf{x}) = \prod_{i=1}^r \frac{1}{\sqrt{p_i}} \text{He}_{p_i}(x_i)$ and $\tilde{P} = p_1 + \dots + p_r$. For $Q + \tilde{P} \leq k \leq P + \tilde{P}$,

$$\|\mathbb{E}[\nabla^k(\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))]\|_F^2 \leq R_c^2 \tilde{C}_k$$

holds where \tilde{C}_k is a constant depending only on k . For other k , $\|\mathbb{E}[\nabla^k(\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))]\|_F = 0$.

Proof. First,

$$\|\mathbb{E}[\nabla^k(\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))]\|_F^2 = \left\| \sum_{i=Q}^P \frac{c_i}{i!} \mathbb{E}[\nabla^k(\text{He}_i(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))] \right\|_F^2$$

holds. To calculate $\mathbb{E}[\nabla^k(\text{He}_i(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))]$, note that from Lemma 8,

$$\begin{aligned} & \text{He}_i(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}) \\ & = \left(\sum_{\substack{q_1 + \dots + q_r = i \\ q_j \geq 0}} \frac{i!}{q_1! \dots q_r!} \beta_1^{q_1} \dots \beta_r^{q_r} \text{He}_{q_1}(x_1) \dots \text{He}_{q_r}(x_r) \right) \frac{\text{He}_{p_1}(x_1)}{\sqrt{p_1!}} \dots \frac{\text{He}_{p_r}(x_r)}{\sqrt{p_r!}}. \quad (\text{C.4}) \end{aligned}$$

Also, from the discussion in Appendix A.3, $\mathbb{E}_{\mathbf{x}} \left[\frac{\partial^{s_1}}{\partial x_1^{s_1}} \dots \frac{\partial^{s_r}}{\partial x_r^{s_r}} \text{He}_i(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}) \right]$ is equal to the coefficient of (C.4) with respect to $\frac{1}{s_1! \dots s_r!} \text{He}_{s_1}(x_1) \dots \text{He}_{s_r}(x_r)$. Therefore, $\mathbb{E}[\nabla^k(\text{He}_i(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))] \neq \mathbf{0}$ if and only if $i + \tilde{P} = k$. Then, $\left\| \sum_{i=Q}^P \frac{c_i}{i!} \mathbb{E}[\nabla^k(\text{He}_i(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))] \right\|_F^2 = \left\| \frac{c_{k-\tilde{P}}}{(k-\tilde{P})!} \mathbb{E}[\nabla^k(\text{He}_{k-\tilde{P}}(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))] \right\|_F^2$.

Here, we can observe that the (j_1, \dots, j_k) -th entry of the tensor $\frac{c_{k-\tilde{P}}}{(k-\tilde{P})!} \mathbb{E}[\nabla^k(\text{He}_{k-\tilde{P}}(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))]$ is equal to

$$c_{k-\tilde{P}} \frac{s_1! \dots s_r!}{(s_1 - p_1)! \dots (s_r - p_r)!} \cdot \beta_1^{s_1 - p_1} \dots \beta_r^{s_r - p_r} \frac{1}{\sqrt{p_1!} \dots \sqrt{p_r!}} \mathbb{1}_{s_1 - p_1 \geq 0, \dots, s_r - p_r \geq 0}$$

where $s_i = \sum_{k'=1}^k \mathbb{1}_{j_{k'}=i}$ ($i \in [r]$). There are $\frac{k!}{s_1! \dots s_r!}$ index sets (j_1, \dots, j_k) having the same entry, and thus

$$\begin{aligned} & \left\| \frac{c_{k-\tilde{P}}}{(k-\tilde{P})!} \mathbb{E}[\nabla^k(\text{He}_{k-\tilde{P}}(\langle \mathbf{x}, \boldsymbol{\beta} \rangle)h(\mathbf{x}))] \right\|_F^2 \\ & = \sum_{s_1 + \dots + s_r = k} c_{k-\tilde{P}}^2 \frac{1}{p_1! \dots p_r!} \frac{k!}{s_1! \dots s_r!} \left(\frac{s_1! \dots s_r!}{(s_1 - p_1)! \dots (s_r - p_r)!} \right)^2 \end{aligned}$$

$$\begin{aligned}
& \cdot (\beta_1^{s_1-p_1} \dots \beta_r^{s_r-p_r})^2 \mathbb{I}_{s_1-p_1 \geq 0, \dots, s_r-p_r \geq 0} \\
= & \sum_{q_1+\dots+q_r=k-\tilde{P}, q_j \geq 0} c_{k-\tilde{P}}^2 \frac{1}{p_1! \dots p_r!} \frac{k!}{(q_1+p_1)! \dots (q_r+p_r)!} \left(\frac{(q_1+p_1)! \dots (q_r+p_r)!}{q_1! \dots q_r!} \right)^2 \\
& \cdot (\beta_1^{q_1} \dots \beta_r^{q_r})^2 \\
= & \sum_{q_1+\dots+q_r=k-\tilde{P}, q_j \geq 0} c_{k-\tilde{P}}^2 \frac{(k-\tilde{P})!}{q_1! \dots q_r!} \frac{k!}{(k-\tilde{P})!} \left(\frac{(q_1+p_1)! \dots (q_r+p_r)!}{p_1! \dots p_r! q_1! \dots q_r!} \right) \\
& \cdot (\beta_1^{q_1} \dots \beta_r^{q_r})^2. \tag{C.5}
\end{aligned}$$

We can verify that there exists a constant \bar{C}_k depending only on P such that $\frac{k!}{(k-\tilde{P})!} \left(\frac{(q_1+p_1)! \dots (q_r+p_r)!}{p_1! \dots p_r! q_1! \dots q_r!} \right) \leq \bar{C}_k$ holds uniformly; for example, one can naively obtain

$$\frac{k!}{(k-\tilde{P})!} \left(\frac{(q_1+p_1)! \dots (q_r+p_r)!}{p_1! \dots p_r! q_1! \dots q_r!} \right) \leq k!(q_1+p_1)! \dots (q_r+p_r)! \leq k!k^k.$$

Here we used $(q_1+p_1) + \dots + (q_r+p_r) = k$.

Moreover, $\sum_{q_1+\dots+q_r=k-\tilde{P}, q_i \geq 0} \frac{(k-\tilde{P})!}{q_1! \dots q_r!} (\beta_1^{q_1} \dots \beta_r^{q_r})^2 = \|\beta^{\otimes k-\tilde{P}}\|_F^2 = 1$ holds. Hence (C.5) is upper bounded by $c_{k-\tilde{P}}^2 \bar{C}_k \leq R_c^2 \bar{C}_k$. \square

Corollary 44. *With high probability,*

$$\left| \frac{1}{N_2} \sum_{j=1}^{N_2} y_j h(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}}[\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle) h(\mathbf{x})] \right| = \tilde{O}\left(\frac{1}{\sqrt{N_2}}\right)$$

holds.

Proof. Plugging the results above and $t = \Theta\left(\frac{1}{\sqrt{N_2}}(\log d)^P\right)$ into Lemma 42 yields

$$\left| \frac{1}{N_2} \sum_{j=1}^{N_2} \sigma_*(\langle \mathbf{x}_j, \boldsymbol{\beta} \rangle) h(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}}[\sigma_*(\langle \mathbf{x}, \boldsymbol{\beta} \rangle) h(\mathbf{x})] \right| \lesssim \frac{1}{\sqrt{N_2}}(\log d)^P$$

with high probability. Moreover, we can easily verify $\left| \frac{1}{N_2} \sum_{j=1}^{N_2} \varsigma_j h(\mathbf{x}_j) \right| = \tilde{O}\left(\frac{1}{\sqrt{N_2}}\right)$ by the method used in the proof of Lemma 26. \square

C.3 Proof of Proposition 2

We prove the following Proposition 2 using the preparations above.

Proposition 2. *Let $\mathbf{X}^t = (\mathbf{x}_1^t, \dots, \mathbf{x}_{N_2}^t)$ and $\mathbf{y}^t = (y_1^t, \dots, y_{N_2}^t)^\top$ where $T_1 + 1 \leq t \leq T_1 + T_2$. With high probability over the data distribution and random initialization of the parameters, there exists $\bar{\Gamma}$ such that*

$$\left| f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \bar{\Gamma}, \mathbf{b}) - y^t \right| - \tau = \tilde{O}\left(\sqrt{\frac{r^{3P}}{m} + \frac{r^{2P}}{N_2}}\right)$$

holds for all $t \in \{T_1 + 1, \dots, T_2\}$, where $\mathbf{W}^{(1)}$ is obtained by line 4 of Algorithm 1 and \mathbf{b} is re-initialized in Line 5 of Algorithm 1, where the conditions on $\eta_1, \lambda_1, m, \gamma, N_1, T_1$ are identical to that in Theorem 1. Moreover, $\|\bar{\Gamma}\|_F = \tilde{O}(r^{2P}/m)$ is satisfied.

Proof. It suffices to show the proposition for each $T_1 + 1 \leq t \leq T_1 + T_2$, and then take the union bound. We drop the superscript t for simplicity.

Note that, from Lemma 8 and $\mathbb{E}_{\mathbf{x}}[h(\mathbf{x})h'(\mathbf{x})] = \mathbb{I}_{h=h'}$ for $h, h' \in \mathcal{H}$, we can expand $f_*(\mathbf{x})$ as $f_*(\mathbf{x}) = \sum_{n=1}^{B_P} \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}[f_*(\mathbf{x})h_n(\mathbf{x})]h_n(\mathbf{x})$. Now let $\mathbf{A} = [\mathbf{a}^1 \dots \mathbf{a}^{B_P}] \in \mathbb{R}^{m \times B_P}$, where

$\{\mathbf{a}^n\}_{n=1}^{B_P}$ is constructed in Proposition 33. Then, by setting $\bar{\Gamma} = \mathbf{A}\mathbf{A}^\top$,

$$\begin{aligned}
& \left\langle \frac{\mathbf{A}\mathbf{A}^\top \sigma(\mathbf{W}^{(1)}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top)\mathbf{y}}{N_2}, \begin{bmatrix} \sigma(\mathbf{w}_1^{(1)\top}\mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^{(1)\top}\mathbf{x} + b_m) \end{bmatrix} \right\rangle - y \\
&= \sum_{n=1}^{B_P} (h_n(\mathbf{x}) + \epsilon_n(\mathbf{x})) \left(\mathbb{E}[f_* h_n] + \left(\frac{1}{N_2} \sum_{j=1}^{N_2} y_j h_n(\mathbf{x}_j) - \mathbb{E}[f_* h_n] \right) + \frac{1}{N_2} \sum_{j=1}^{N_2} y_j \epsilon_n(\mathbf{x}_j) \right) \\
&\quad - f_*(\mathbf{x}) - \varsigma \\
&= \sum_{n=1}^{B_P} (h_n(\mathbf{x}) + \epsilon_n(\mathbf{x})) \left(\mathbb{E}[f_* h_n] + \left(\frac{1}{N_2} \sum_{j=1}^{N_2} y_j h_n(\mathbf{x}_j) - \mathbb{E}[f_* h_n] \right) + \frac{1}{N_2} \sum_{j=1}^{N_2} y_j \epsilon_n(\mathbf{x}_j) \right) \\
&\quad - \sum_{n=1}^{B_P} h_n(\mathbf{x}) \mathbb{E}[f_* h_n] - \varsigma
\end{aligned}$$

holds, where $\epsilon_n(\mathbf{x}) := \sum_{j=1}^m \mathbf{a}_j^n \sigma(\mathbf{w}_j^{(1)\top}\mathbf{x} + b_j) - h_n(\mathbf{x})$ satisfies $|\epsilon_n(\mathbf{x})| = \tilde{O}\left(\sqrt{r^P/m + 1/N_2}\right)$ from Proposition 33.

To bound the error, we note that the following holds for each n :

- As $h_n(\mathbf{x})$ is a polynomial whose degree is at most P , and $\mathbb{E}_{\mathbf{x}}[\nabla^k h_n(\mathbf{x})] = O(1)$, $|h_n(\mathbf{x})| \lesssim (\log d)^{P/2}$ holds with high probability from Lemma 15.
- From lemma 8, $|\mathbb{E}[f_*(\mathbf{x})h_n(\mathbf{x})]| = O(1)$ holds.
- From Corollary 17, $|y_j| \lesssim (\log d)^{P/2}$ holds with high probability.

Hence, for each $n \in [B_P]$, we obtain

$$\begin{aligned}
& \left| \left(\underbrace{h_n(\mathbf{x})}_{\tilde{O}(1)} + \underbrace{\epsilon_n(\mathbf{x})}_{\tilde{O}\left(\sqrt{\frac{r^P}{m} + \frac{1}{N_2}}\right)} \right) \left(\underbrace{\mathbb{E}[f_* h_n]}_{O(1)} + \underbrace{\left(\frac{1}{N_2} \sum_{j=1}^{N_2} y_j h_n(\mathbf{x}_j) - \mathbb{E}[f_* h_n] \right)}_{\tilde{O}\left(\sqrt{\frac{1}{N_2}}\right)} + \underbrace{\frac{1}{N_2} \sum_{j=1}^{N_2} y_j \epsilon_n(\mathbf{x}_j)}_{\tilde{O}\left(\sqrt{\frac{r^P}{m} + \frac{1}{N_2}}\right)} \right) \right. \\
& \quad \left. - h_n(\mathbf{x}) \mathbb{E}[f_* h_n] \right| \\
&= \tilde{O}\left(\sqrt{\frac{r^P}{m} + \frac{1}{N_2}} + \left(\frac{r^P}{m} + \frac{1}{N_2} \right) \right).
\end{aligned}$$

Combining this with $B_P = \Theta(r^P)$ and $m = \tilde{\Omega}(r^P)$, we arrive at the upper bound.

Finally for $\|\bar{\Gamma}\|_F$, we have

$$\|\bar{\Gamma}\|_F \leq \sum_{n=1}^{B_P} \|\mathbf{a}^n (\mathbf{a}^n)^\top\|_F = \sum_{n=1}^{B_P} \|\mathbf{a}^n\|^2 = \tilde{O}(r^{2P}/m).$$

□

D Generalization Error Analysis and Proof of Theorem 1

Note that after one-step gradient descent, $\|\mathbf{w}_j^{(1)}\| = \tilde{O}_{d,r}(1)$ is ensured with high probability from Lemma 18 and Corollary 32.

D.1 Rademacher Complexity Bound

Let $\mathbf{W} = \mathbf{W}^{(1)}$ and suppose b_1, \dots, b_m are biases obtained in line 5 of Algorithm 1. Let

$$\mathcal{F}_{N,G} = \left\{ (\mathbf{X}, \mathbf{y}, \mathbf{x}) \mapsto \left\langle \frac{\mathbf{\Gamma} \sigma(\mathbf{W}^{(1)} \mathbf{X} + \mathbf{b} \mathbf{1}_N^\top) \mathbf{y}}{N}, \begin{bmatrix} \sigma((\mathbf{w}_1^{(1)})^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma((\mathbf{w}_m^{(1)})^\top \mathbf{x} + b_m) \end{bmatrix} \right\rangle \middle| \|\mathbf{\Gamma}\|_F \leq G \right\}$$

be the set of transformers where the norm of the attention matrix is constrained, and let

$$\text{Rad}_T(\mathcal{F}_{N,G}) = \mathbb{E}_{\{\mathbf{X}^t\}, \{\mathbf{y}^t\}, \{\mathbf{x}^t\}, \{\epsilon^t\}} \left[\sup_{f \in \mathcal{F}_{N,G}} \frac{1}{T} \sum_{t=1}^T \epsilon^t f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t) \right]$$

be its Rademacher complexity, where $\epsilon^t \sim \text{Unif}(\{\pm 1\})$. Here contexts $\mathbf{X} \in \mathbb{R}^{d \times N}$ and $\mathbf{y} \in \mathbb{R}^N$ are of length N . We can evaluate the Rademacher complexity as follows.

Proposition 45.

$$\text{Rad}_T(\mathcal{F}_{N,G}) = \tilde{O}_{d,r} \left(\frac{Gm}{\sqrt{T}} \right)$$

holds, when $|b_j| = \tilde{O}_{d,r}(1)$ and $\|\mathbf{w}_j^{(1)}\| = \tilde{O}_{d,r}(1)$ for all $j \in [m]$.

Proof. First, it holds that

$$\begin{aligned} & \text{Rad}_T(\mathcal{F}_{N,G}) \\ &= \mathbb{E}_{\{\mathbf{X}^t\}, \{\mathbf{y}^t\}, \{\mathbf{x}^t\}, \{\epsilon^t\}} \left[\sup_{f \in \mathcal{F}_{N,G}} \frac{1}{T} \sum_{t=1}^T \epsilon^t f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t) \right] \\ &= \mathbb{E} \left[\sup_{\|\mathbf{\Gamma}\|_F \leq G} \frac{1}{T} \sum_{t=1}^T \epsilon^t \left\langle \frac{\mathbf{\Gamma} \sigma(\mathbf{W}^{(1)} \mathbf{X}^t + \mathbf{b} \mathbf{1}_N^\top) \mathbf{y}^t}{N}, \begin{bmatrix} \sigma((\mathbf{w}_1^{(1)})^\top \mathbf{x}^t + b_1) \\ \vdots \\ \sigma((\mathbf{w}_m^{(1)})^\top \mathbf{x}^t + b_m) \end{bmatrix} \right\rangle \right] \\ &\stackrel{(a)}{\leq} \mathbb{E} \left[\frac{1}{T} \sup_{\|\mathbf{\Gamma}\|_F \leq G} \|\mathbf{\Gamma}\|_F \left\| \sum_{t=1}^T \epsilon^t \frac{\sigma(\mathbf{W}^{(1)} \mathbf{X}^t + \mathbf{b} \mathbf{1}_N^\top) \mathbf{y}^t}{N} \begin{bmatrix} \sigma((\mathbf{w}_1^{(1)})^\top \mathbf{x}^t + b_1) \\ \vdots \\ \sigma((\mathbf{w}_m^{(1)})^\top \mathbf{x}^t + b_m) \end{bmatrix} \right\|_F \right] \\ &= \frac{G}{T} \mathbb{E} \left[\left\| \sum_{t=1}^T \epsilon^t \frac{\sigma(\mathbf{W}^{(1)} \mathbf{X}^t + \mathbf{b} \mathbf{1}_N^\top) \mathbf{y}^t}{N} \begin{bmatrix} \sigma((\mathbf{w}_1^{(1)})^\top \mathbf{x}^t + b_1) \\ \vdots \\ \sigma((\mathbf{w}_m^{(1)})^\top \mathbf{x}^t + b_m) \end{bmatrix} \right\|_F \right] \\ &\leq \frac{G}{T} \sqrt{\mathbb{E} \left[\left\| \sum_{t=1}^T \epsilon^t \frac{\sigma(\mathbf{W}^{(1)} \mathbf{X}^t + \mathbf{b} \mathbf{1}_N^\top) \mathbf{y}^t}{N} \begin{bmatrix} \sigma((\mathbf{w}_1^{(1)})^\top \mathbf{x}^t + b_1) \\ \vdots \\ \sigma((\mathbf{w}_m^{(1)})^\top \mathbf{x}^t + b_m) \end{bmatrix} \right\|_F^2 \right]} \\ &\stackrel{(b)}{=} \frac{G}{T} \sqrt{T \mathbb{E}_{\mathbf{X}, \mathbf{y}, \mathbf{x}} \left[\left\| \frac{\sigma(\mathbf{W}^{(1)} \mathbf{X} + \mathbf{b} \mathbf{1}_N^\top) \mathbf{y}}{N} \begin{bmatrix} \sigma((\mathbf{w}_1^{(1)})^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma((\mathbf{w}_m^{(1)})^\top \mathbf{x} + b_m) \end{bmatrix} \right\|_F^2 \right]} \\ &= \frac{G}{\sqrt{T}} \sqrt{\sum_{k,l=1}^m \mathbb{E}_{\mathbf{X}, \mathbf{y}, \mathbf{x}} \left[\left(\frac{1}{N} \sum_{j=1}^N y_j \sigma((\mathbf{w}_k^{(1)})^\top \mathbf{x}_j + b_k) \sigma((\mathbf{w}_l^{(1)})^\top \mathbf{x} + b_l) \right)^2 \right]} \end{aligned}$$

$$\begin{aligned}
&\stackrel{(c)}{\leq} \frac{G}{\sqrt{T}} \sqrt{\sum_{k,l=1}^m \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d), \mathbf{x}' \sim \mathcal{N}(0, \mathbf{I}_d), f_*, \varsigma} \left[(f_*(\mathbf{x}) + \varsigma)^2 \sigma((\mathbf{w}_k^{(1)})^\top \mathbf{x} + b_k)^2 \sigma((\mathbf{w}_l^{(1)})^\top \mathbf{x}' + b_l)^2 \right]} \\
&\stackrel{(d)}{\leq} \frac{G}{\sqrt{T}} \sqrt{\sum_{k,l=1}^m \mathbb{E}_{\mathbf{x}, f_*, \varsigma} \left[(2f_*(\mathbf{x})^2 + 2\varsigma^2)(2((\mathbf{w}_k^{(1)})^\top \mathbf{x})^2 + 2b_k^2) \right] \mathbb{E}_{\mathbf{x}'} \left[(2((\mathbf{w}_l^{(1)})^\top \mathbf{x}')^2 + 2b_l^2) \right]},
\end{aligned}$$

where the annotated equality/inequalities are obtained by (a) Cauchy–Schwarz inequality, (b) $\mathbb{E}_{\epsilon_t \sim \text{Unif}\{\pm 1\}} \left[\left\| \sum_t \epsilon_t \mathbf{A}_t \right\|_F^2 \right] = \sum_t \|\mathbf{A}_t\|_F^2$, (c) convexity of $f(z) = z^2$, and (d) $(\sigma(\alpha + \beta))^2 \leq (\alpha + \beta)^2 \leq 2\alpha^2 + 2\beta^2$.

Let us further evaluate the last line: from the assumption and Lemma 7, we know that $\|\mathbf{w}_k\|$, $|b_k|$ and $\mathbb{E}[f_*(\mathbf{x})^2]$ are all $\tilde{O}(1)$: in particular, we can easily obtain that $\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[((\mathbf{w}_k^{(1)})^\top \mathbf{x})^2 \right] = \tilde{O}_{d,r}(1)$.

Consequently, the last line is $\frac{G}{\sqrt{T}} \sqrt{\sum_{k,l=1}^m \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d), f_*} \left[(2f_*(\mathbf{x})^2)(2((\mathbf{w}_k^{(1)})^\top \mathbf{x})^2) \right]} + \tilde{O}_{d,r}(1)$, and remains to evaluate $\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d), f_*} \left[f_*(\mathbf{x})^2 ((\mathbf{w}_k^{(1)})^\top \mathbf{x})^2 \right]$. On this term, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}, f_*} \left[f_*(\mathbf{x})^2 ((\mathbf{w}_k^{(1)})^\top \mathbf{x})^2 \right] &\leq \mathbb{E}_{\mathbf{x}, f_*} \left[f_*(\mathbf{x})^4 \right]^{1/2} \mathbb{E}_{\mathbf{x}} \left[((\mathbf{w}_k^{(1)})^\top \mathbf{x})^4 \right]^{1/2} \\
&= \tilde{O}(1) \quad (\because \text{Lemma 7}).
\end{aligned}$$

Hence we can conclude $\text{Rad}_T(\mathcal{F}_{N,G}) = O\left(\frac{Gm}{\sqrt{T}}\right)$. \square

Corollary 46. *Let*

$$\begin{aligned}
\tilde{\mathcal{F}}_{N,G} &= \left\{ (\mathbf{X}, \mathbf{y}, \mathbf{x}, y) \mapsto \left\langle \frac{\mathbf{\Gamma} \sigma(\mathbf{W}^{(1)} \mathbf{X} + \mathbf{b} \mathbf{1}_N^\top) \mathbf{y}}{N}, \begin{bmatrix} \sigma((\mathbf{w}_1^{(1)})^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma((\mathbf{w}_m^{(1)})^\top \mathbf{x} + b_m) \end{bmatrix} \right\rangle \middle| \|\mathbf{\Gamma}\|_F \leq G \right\} \\
&\cup \left\{ (\mathbf{X}, \mathbf{y}, \mathbf{x}, y) \mapsto y \right\}.
\end{aligned}$$

Then, $\text{Rad}_T(\tilde{\mathcal{F}}_{N,G}) = \tilde{O}_{d,r}\left(\frac{Gm}{\sqrt{T}}\right)$ holds when $|b_j| = \tilde{O}_{d,r}(1)$ and $\|\mathbf{w}_j^{(1)}\| = \tilde{O}_{d,r}(1)$ for all $j \in [m]$.

Proof. Note that

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{X}^t\}, \{\mathbf{y}^t\}, \{\mathbf{x}^t\}, \{y^t\}, \{\epsilon^t\}} \left[\sup_{f \in \tilde{\mathcal{F}}_{N,G}} \frac{1}{T} \sum_{t=1}^T \epsilon^t f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t) \right] \\
&\leq \mathbb{E}_{\{\mathbf{X}^t\}, \{\mathbf{y}^t\}, \{\mathbf{x}^t\}, \{\epsilon^t\}} \left[\sup_{f \in \mathcal{F}_{N,G}} \frac{1}{T} \sum_{t=1}^T \epsilon^t f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t) \right] + \mathbb{E}_{\{y^t\}, \{\epsilon^t\}} \left[\left| \frac{1}{T} \sum_{t=1}^T \epsilon^t y^t \right| \right].
\end{aligned}$$

This follows from $\sup_{f \in \mathcal{F}_{N,G}} \frac{1}{T} \sum_{t=1}^T \epsilon^t f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t) \geq 0$, as $\mathcal{F}_{N,G}$ contains zero function corresponding to $\mathbf{\Gamma} = \mathbf{O}$. The second term in the last line can be bounded as

$$\begin{aligned}
\mathbb{E}_{\{y^t\}, \{\epsilon^t\}} \left[\left| \frac{1}{T} \sum_{t=1}^T \epsilon^t y^t \right| \right] &\leq \sqrt{\mathbb{E}_{\{y^t\}, \{\epsilon^t\}} \left[\left(\frac{1}{T} \sum_{t=1}^T \epsilon^t y^t \right)^2 \right]} \\
&= \frac{1}{\sqrt{T}} \mathbb{E}[(y^t)^2] = O\left(\frac{1}{\sqrt{T}}\right)
\end{aligned}$$

by Lemma 7. \square

D.2 Prompt Length-free Generalization Bound

The ICL risk for prompt length N was defined as

$$\mathcal{R}_N(f) = \mathbb{E}_{\mathbf{X}, \mathbf{y}, \mathbf{x}, \mathbf{y}} [|f(\mathbf{X}_{1:N}, \mathbf{y}_{1:N}, \mathbf{x}; \mathbf{W}, \mathbf{\Gamma}, \mathbf{b}) - y|],$$

where the length of $\mathbf{X}_{1:N}$ and $\mathbf{y}_{1:N}$ is fixed to N . In this section, we show that the ICL risk “converges” when the context length is sufficiently large.

Proposition 47. *Assume that $\|\mathbf{w}_j\| = \tilde{O}_{d,r}(1)$ and $|b_j| = \tilde{O}_{d,r}(1)$ for each $j \in [m]$. Then,*

$$|\mathcal{R}_N(f) - \mathcal{R}_M(f)| = \tilde{O}\left(\|\mathbf{\Gamma}\|_F m \sqrt{1/N + 1/M}\right)$$

holds.

Proof. Note that

$$\begin{aligned} & |\mathcal{R}_N(f) - \mathcal{R}_M(f)| \\ & \leq \mathbb{E} [|f(\mathbf{X}_{1:N}, \mathbf{y}_{1:N}, \mathbf{x}; \mathbf{W}, \mathbf{\Gamma}, \mathbf{b}) - f(\mathbf{X}_{1:M}, \mathbf{y}_{1:M}, \mathbf{x}; \mathbf{W}, \mathbf{\Gamma}, \mathbf{b})|] \\ & = \mathbb{E} \left[\left\| \left\langle \mathbf{\Gamma} \left(\frac{\sigma(\mathbf{W}\mathbf{X}_{1:N} + \mathbf{b}\mathbf{1}_N^\top)\mathbf{y}_{1:N}}{N} - \frac{\sigma(\mathbf{W}\mathbf{X}_{1:M} + \mathbf{b}\mathbf{1}_M^\top)\mathbf{y}_{1:M}}{M} \right), \begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \end{bmatrix} \right\rangle \right\| \right] \\ & \stackrel{(a)}{\leq} \|\mathbf{\Gamma}\|_F \mathbb{E} \left[\left\| \left(\frac{\sigma(\mathbf{W}\mathbf{X}_{1:N} + \mathbf{b}\mathbf{1}_N^\top)\mathbf{y}_{1:N}}{N} - \frac{\sigma(\mathbf{W}\mathbf{X}_{1:M} + \mathbf{b}\mathbf{1}_M^\top)\mathbf{y}_{1:M}}{M} \right) \right\| \left\| \begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \end{bmatrix} \right\| \right] \\ & \leq \|\mathbf{\Gamma}\|_F \mathbb{E} \left[\left\| \left(\frac{\sigma(\mathbf{W}\mathbf{X}_{1:N} + \mathbf{b}\mathbf{1}_N^\top)\mathbf{y}_{1:N}}{N} - \frac{\sigma(\mathbf{W}\mathbf{X}_{1:M} + \mathbf{b}\mathbf{1}_M^\top)\mathbf{y}_{1:M}}{M} \right) \right\|^2 \right]^{1/2} \\ & \quad \cdot \mathbb{E} \left[\left\| \begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \end{bmatrix} \right\|^2 \right]^{1/2}, \end{aligned}$$

where (a) holds because $\mathbf{a}^\top \mathbf{\Gamma} \mathbf{b} = \mathbf{\Gamma} \circ (\mathbf{a} \mathbf{b}^\top) \leq \|\mathbf{\Gamma}\|_F \|\mathbf{a} \mathbf{b}^\top\|_F = \|\mathbf{\Gamma}\|_F \|\mathbf{a}\| \|\mathbf{b}\|$.

Let us bound $\mathbb{E} \left[\left\| \left(\frac{\sigma(\mathbf{W}\mathbf{X}_{1:N} + \mathbf{b}\mathbf{1}_N^\top)\mathbf{y}_{1:N}}{N} - \frac{\sigma(\mathbf{W}\mathbf{X}_{1:M} + \mathbf{b}\mathbf{1}_M^\top)\mathbf{y}_{1:M}}{M} \right) \right\|^2 \right]$, which is to $\sum_{j=1}^m \mathbb{E}_{f_*} \left[\mathbb{E}_{\mathbf{X}, \mathbf{y}} \left[\left(\frac{1}{N} \sum_{i=1}^N \sigma(\mathbf{w}_j^\top \mathbf{x}_i + b_j) y_i - \frac{1}{M} \sum_{i=1}^M \sigma(\mathbf{w}_j^\top \mathbf{x}_i + b_j) y_i \right)^2 \right] \right]$. For simplicity, we drop the subscript j and obtain

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}, \mathbf{y}} \left[\left(\frac{1}{N} \sum_{i=1}^N \sigma(\mathbf{w}^\top \mathbf{x}_i + b) y_i - \frac{1}{M} \sum_{i=1}^M \sigma(\mathbf{w}^\top \mathbf{x}_i + b) y_i \right)^2 \right] \\ & \leq 2 \mathbb{E} \left[\left(\frac{1}{N} \sum_{i=1}^N \sigma(\mathbf{w}^\top \mathbf{x}_i + b) y_i - \mathbb{E}[\sigma(\mathbf{w}^\top \mathbf{x} + b) y] \right)^2 \right] \\ & \quad + 2 \mathbb{E} \left[\left(\frac{1}{M} \sum_{i=1}^M \sigma(\mathbf{w}^\top \mathbf{x}_i + b) y_i - \mathbb{E}[\sigma(\mathbf{w}^\top \mathbf{x} + b) y] \right)^2 \right] \\ & = \frac{2}{N} \mathbb{V}[\sigma(\mathbf{w}^\top \mathbf{x} + b) y] + \frac{2}{M} \mathbb{V}[\sigma(\mathbf{w}^\top \mathbf{x} + b) y] \\ & \leq \left(\frac{2}{N} + \frac{2}{M} \right) \mathbb{E}[\sigma(\mathbf{w}^\top \mathbf{x} + b)^2 (f_*(\mathbf{x}) + \varsigma)^2] \\ & \stackrel{(a)}{\leq} \left(\frac{2}{N} + \frac{2}{M} \right) \mathbb{E}[(2(\mathbf{w}^\top \mathbf{x})^2 + 2b^2)(2f_*(\mathbf{x})^2 + 2\varsigma^2)]. \end{aligned}$$

Here (a) is obtained in the same vein as the derivation of (d) in the proof of Proposition 45. Moreover, in the same way as that proof, we can show that $\mathbb{E}[(2(\mathbf{w}^\top \mathbf{x})^2 + 2b^2)(2f_*(\mathbf{x})^2 + 2\varsigma^2)] = \tilde{O}_{d,r}(1)$.

Consequently we get $\mathbb{E}\left[\left\|\left(\frac{\sigma(\mathbf{W}\mathbf{X}_{1:N} + \mathbf{b}\mathbf{1}_N^\top)\mathbf{y}_{1:N}}{N} - \frac{\sigma(\mathbf{W}\mathbf{X}_{1:M} + \mathbf{b}\mathbf{1}_M^\top)\mathbf{y}_{1:M}}{M}\right)\right\|^2\right] = \left(\frac{2}{N} + \frac{2}{M}\right)\tilde{O}_{d,r}(m)$.

Next, we observe that

$$\mathbb{E}\left[\left\|\begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \end{bmatrix}\right\|^2\right] \leq 2 \sum_{j=1}^m \mathbb{E}[(\mathbf{w}_j^\top \mathbf{x})^2 + b_j^2] = \tilde{O}_{d,r}(m).$$

Putting all things together, we arrive at

$$|\mathcal{R}_N(f) - \mathcal{R}_M(f)| = \tilde{O}\left(\|\mathbf{\Gamma}\|_F m \sqrt{1/N + 1/M}\right).$$

□

D.3 Proof of Theorem 1

Finally we are ready to prove our main theorem.

Proof. [Proof of Theorem 1] Let $\bar{\mathbf{\Gamma}}$ be the attention matrix constructed in Proposition 2 and let $\mathbf{\Gamma}^*$ be the minimizer of the ridge regression problem (line 8 in Algorithm 1). By the equivalence between optimization with L_2 regularization and norm-constrained optimization, there exists λ_2 such that

$$\|\mathbf{\Gamma}^*\|_F \leq \|\bar{\mathbf{\Gamma}}\|_F = \tilde{O}(r^{2P}/m),$$

$$\begin{aligned} & \left(\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} |y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \mathbf{\Gamma}^*, \mathbf{b})|\right)^2 \\ & \leq \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} (y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \mathbf{\Gamma}^*, \mathbf{b}))^2 \\ & \leq \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} (y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \bar{\mathbf{\Gamma}}, \mathbf{b}))^2. \end{aligned}$$

Then, from Proposition 2,

$$\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} |y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \mathbf{\Gamma}^*, \mathbf{b})| - \tau = \tilde{O}\left(\sqrt{\frac{r^{3P}}{m} + \frac{r^{2P}}{N_2}}\right)$$

holds.

We evaluate $\mathcal{R}_{N_2}(f) - \tau$ where $f = f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \mathbf{\Gamma}^*, \mathbf{b})$ using the Rademacher complexity.

$$\begin{aligned} & \mathcal{R}_{N_2}(f) - \tau \\ & = \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} |y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \mathbf{\Gamma}^*, \mathbf{b})| \\ & \quad + \left(\mathcal{R}_{N_2}(f) - \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} |y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t; \mathbf{W}^{(1)}, \mathbf{\Gamma}^*, \mathbf{b})|\right) - \tau \\ & \leq \tilde{O}\left(\sqrt{\frac{r^{3P}}{m} + \frac{r^{2P}}{N_2}}\right) \\ & \quad + \sup_{f \in \tilde{\mathcal{F}}_{N, \|\bar{\mathbf{\Gamma}}\|_F}} \left(\mathcal{R}_{N_2}(f) - \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} |y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)|\right) \end{aligned}$$

holds, where $\tilde{\mathcal{F}}$ is defined in Corollary 46. We can evaluate the expectation value of the second term of the last line as, using $\epsilon^t \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{\pm 1\}$,

$$\begin{aligned}
& \mathbb{E}_{\{\mathbf{X}^t\}, \{\mathbf{y}^t\}, \{\mathbf{x}^t\}, \{y^t\}} \left[\sup_{f \in \tilde{\mathcal{F}}_{N, \|\bar{\Gamma}\|_F}} \left(\mathcal{R}_{N_2}(f) - \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} |y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)| \right) \right] \\
& \stackrel{(a)}{\leq} 2\mathbb{E}_{\{\mathbf{X}^t\}, \{\mathbf{y}^t\}, \{\mathbf{x}^t\}, \{y^t\}, \{\epsilon^t\}} \left[\sup_{f \in \tilde{\mathcal{F}}_{N, \|\bar{\Gamma}\|_F}} \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \epsilon^t |y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)| \right] \\
& \stackrel{(b)}{\lesssim} \text{Rad}_{T_2}(\tilde{\mathcal{F}}_{N_2, \|\bar{\Gamma}\|_F}) + \mathbb{E}_{y, \epsilon} \left[\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \epsilon^t y^t \right] \\
& \leq \text{Rad}_{T_2}(\tilde{\mathcal{F}}_{N_2, \|\bar{\Gamma}\|_F}) + \mathbb{E}_{y, \epsilon} \left[\left(\frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \epsilon^t y^t \right)^2 \right]^{1/2} \\
& \leq \text{Rad}_{T_2}(\tilde{\mathcal{F}}_{N_2, \|\bar{\Gamma}\|_F}) + \frac{1}{\sqrt{T_2}} \mathbb{E}[y^2]^{1/2} \stackrel{(c)}{=} \tilde{O} \left(\sqrt{\frac{r^{4P}}{T_2}} \right).
\end{aligned}$$

Here, (a) holds from the standard symmetrization argument (see eq. (4.17) in [Wai19] for example), (b) is derived using eq. (1) in [Mau16], noting that $(x, y)^\top \mapsto |x - y|$ is $\sqrt{2}$ -Lipschitz continuous, and (c) follows from the bound of $\|\bar{\Gamma}\|_F$ in Proposition 2, together with Lemma 7 and Proposition 45.

Note that

$$\sup_{f \in \tilde{\mathcal{F}}_{N, \|\bar{\Gamma}\|_F}} \left(\mathcal{R}_{N_2}(f) - \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} |y^t - f(\mathbf{X}^t, \mathbf{y}^t, \mathbf{x}^t, y^t)| \right) \quad (\text{D.1})$$

is nonnegative because $f(\mathbf{X}, \mathbf{y}, \mathbf{x}, y) = y$ is included in $\tilde{\mathcal{F}}_{N, \|\bar{\Gamma}\|_F}$. Hence from Markov's inequality, (D.1) is $\tilde{O} \left(\sqrt{\frac{r^{4P}}{T_2}} \right)$ with probability at least 0.995. Hence we obtain

$$\mathcal{R}_{N_2}(f) - \tau = \tilde{O} \left(\sqrt{\frac{r^{3P}}{m} + \frac{r^{2P}}{N_2}} + \sqrt{\frac{r^{4P}}{T_2}} \right).$$

Now we have done the upper bound for $\mathcal{R}_{N_2}(f)$. For $\mathcal{R}_{N^*}(f)$, we can use Proposition 47.

Note that all the desired events occur with high probability, except for the one described above, which occurs with probability 0.995. This completes the proof of Theorem 1. \square

E Derivation of Simplified Self-attention Module

We derive equation (2.6), following the same line of argument as [ZFB23]. As we assumed that $\mathbf{W}^{PV} \mathbf{W}^{KQ}$ are in the form as (2.4), the output of the attention layer is written as

$$\tilde{f}_{\text{Attn}}(\mathbf{E}; \mathbf{W}^{PV}, \mathbf{W}^{KQ}) = \mathbf{E} + \begin{bmatrix} * & * \\ \mathbf{0}_{1 \times m} & v \end{bmatrix} \mathbf{E} \cdot \frac{\mathbf{E}^\top \begin{bmatrix} \mathbf{K} & * \\ \mathbf{0}_{1 \times m} & * \end{bmatrix} \mathbf{E}}{N}. \quad (\text{E.1})$$

Note that we adopt the right-bottom entry $\left(\tilde{f}_{\text{Attn}}(\mathbf{E}; \mathbf{W}^{PV}, \mathbf{W}^{KQ}) \right)_{m+1, N+1}$ as a prediction for output of the query. By (E.1). Hence we have

$$\left(\tilde{f}_{\text{Attn}}(\mathbf{E}; \mathbf{W}^{PV}, \mathbf{W}^{KQ}) \right)_{m+1, N+1}$$

$$\begin{aligned}
& \mathbf{E}^\top \begin{bmatrix} \mathbf{K} & * \\ \mathbf{0}_{1 \times m} & * \end{bmatrix} \begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \\ 0 \end{bmatrix} \\
&= [\mathbf{0}_{1 \times m} \quad v] \mathbf{E} \cdot \frac{\begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \\ 0 \end{bmatrix}}{N} \\
&= v[y_1 \quad \cdots \quad y_N \quad 0] \cdot \frac{\mathbf{E}^\top \begin{bmatrix} \mathbf{K} \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \\ 0 \end{bmatrix}}{N} \\
&= v[y_1 \quad \cdots \quad y_N] \cdot \frac{\sigma(\mathbf{W}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top)^\top \mathbf{K} \begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \end{bmatrix}}{N} \\
&= \left\langle \frac{v \mathbf{K}^\top \sigma(\mathbf{W}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top) \mathbf{y}}{N}, \begin{bmatrix} \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) \\ \vdots \\ \sigma(\mathbf{w}_m^\top \mathbf{x} + b_m) \end{bmatrix} \right\rangle.
\end{aligned}$$

Setting $\Gamma = v \mathbf{K}^\top$ yields equation (2.6).

F Details of Experiments

F.1 Detailed Experimental Settings

For the experiments in Section 4, test loss was averaged over 128 independent tasks, with 2048 and 128 independent queries generated for each task in the experiment of Figure 1 and 2, respectively. During the validation of the experiment of Figure 1, the coefficients $\{c_i\}$ in the single-index model were fixed to be $(c_2, c_3) = (\sqrt{2} \cdot 2! / 2, \sqrt{2} \cdot 3! / 2)$ to reduce the variance in the baseline methods.

GPT-2 model. The experiments in Section 4 were based on the architecture used in [GTLV22]¹, whose backbone is the GPT-2 model [RWC⁺19]. Given the prompt $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N+1}$, the embedding $\mathbf{E} \in \mathbb{R}^{d \times (2N+2)}$ was constructed as $\mathbf{E} = [\mathbf{x}_1, \tilde{\mathbf{y}}_1, \dots, \mathbf{x}_N, \tilde{\mathbf{y}}_N]$ where $\tilde{\mathbf{y}}_i = [y_i, 0, \dots, 0]^\top$. This embedding was mapped to $\tilde{\mathbf{E}} \in \mathbb{R}^{256 \times (2N+2)}$ through the read-in layer. The transformed embedding $\tilde{\mathbf{E}}$ was then passed through the 12-layer GPT-2 backbone with 8 heads, with dropout disabled. Finally, the output was converted into a $2N + 2$ -dimensional vector through the read-out layer, where the $(2k - 1)$ -th entry ($k \in [N + 1]$) corresponds to the prediction $\hat{y}_k(\mathbf{x}_1^t, y_1^t, \dots, \mathbf{x}_{k-1}^t, y_{k-1}^t, \mathbf{x}_k^t)$ for y_k , that is, the answer for the query \mathbf{x}_k^t reading the context of length $k - 1$. We used the Adam optimizer [KB15] with a learning rate of 0.0001. The training loss at each step was set as $\frac{1}{B} \sum_{t=1}^B \sum_{k=1}^{N+1} (y_k^t - \hat{y}_k(\mathbf{x}_1^t, y_1^t, \dots, \mathbf{x}_{k-1}^t, y_{k-1}^t, \mathbf{x}_k^t))^2$ where $B = 8$ is the minibatch size.

Baseline algorithms. Baseline algorithms used in the comparison experiment are as follows:

- Two-layer neural network $f(\mathbf{x}) = \sum_{j=1}^m a_j \sigma(\mathbf{w}_j^\top \mathbf{x})$ where $\sigma = \text{ReLU}$ and $m = 256$. The initialization scale followed the mean-field regime, with $a_j \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{\pm 1/m\})$ and $\mathbf{w}_j \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1})$. The architecture was trained using the Adam optimizer, with a learning rate of 0.1 for the first layer and $0.1/m$ for the second layer, along with a weight decay rate of 0.0001. Early stopping was applied to prevent overfitting.
- Kernel ridge regression using the Gaussian RBF kernel. The ridge regression parameter λ was set to $\lambda = 0.01$.

Again, note that these baseline algorithms were trained from scratch for each validation task, whereas the parameters of the pretrained GPT-2 model remained unchanged across all validation tasks.

¹<https://github.com/dtsip/in-context-learning>, MIT License.

F.2 Experiment on Our Architecture and Algorithm 1

In addition to the GPT-2 experiments, we also conducted comparison between our Algorithm 1 on our simplified architecture and baseline algorithms. Data is generated as $y_i^t = \text{He}_2(\langle \mathbf{x}_i^t, \boldsymbol{\beta}^t \rangle)$, where $\boldsymbol{\beta}^t \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{\boldsymbol{\beta} \in \mathbb{R}^d \mid \boldsymbol{\beta} = [\boldsymbol{\beta}_{1:r}, \mathbf{0}]^\top, \|\boldsymbol{\beta}\| = 1\}$ with $d = 32$ and $r = 2$. We pretrained our transformer architecture (2.6) with $m = 8,000$, using Algorithm 1 with $T_1 = 10^5, N_1 = 10^4, \eta_1 = 10^5$ (recall that $\eta_1 \asymp m^{\frac{3}{2}} r d^{2Q - \frac{1}{2}} \cdot (\log d)^{-C_n}$ should be large) and $T_2 = 3,000, N_2 = 512, \lambda_2 = 0.1$. We used mini-batch SGD to find an approximate optimizer Γ^* of the ridge regression. For validation, we altered the context length from $2^{4.5}$ to 2^9 and compared its test error with one-step GD (training the first layer via one-step gradient descent, and then conducting ridge regression on the second layer) on a two-layer NN of width 8,000 and kernel ridge regression with RBF kernel working directly on the test prompt. In Figure 3 we see that our simplified transformer also outperforms neural network and kernel method, especially in short context regime.

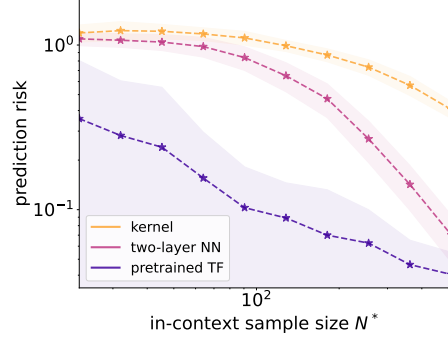


Figure 3: In-context generalization error of kernel ridge regression, neural network + one-step gradient descent, and Algorithm 1 on our Transformer (2.6).

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: They are presented accurately.

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: We mentioned the limitation of our result and the future challenge in Section 5.

Guidelines:

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

3. Theory Assumptions and Proofs

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

Answer: [\[Yes\]](#)

Justification: We presented full-proofs for our results in the appendix, and sketched the proof in Section 3.2.

Guidelines:

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

4. Experimental Result Reproducibility

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

Answer: [Yes]

Justification: See Appendix F.

Guidelines:

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

5. Open access to data and code

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

Answer: [No]

Justification: This is a theory paper where the toy experiments are not central to our contribution.

Guidelines:

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

6. Experimental Setting/Details

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

Answer: [Yes]

Justification: See Section 4 and Appendix F.

Guidelines:

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

7. Experiment Statistical Significance

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

Answer: [Yes]

Justification: See Section 4 and Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

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

Answer: [No]

Justification: We only have toy experiments and it is not a central part of this paper.

Guidelines:

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

9. Code Of Ethics

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

Answer: [Yes]

Justification: Yes, and we checked the code.

Guidelines:

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

10. Broader Impacts

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

Answer: [NA]

Justification: Our results are purely theoretical working on a small simplified model, thus there seems no societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [NA]

Justification: We don't have any models or datasets to release.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: We included the information in Appendix F.

Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New Assets**

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

Answer: [NA]

Justification: We don’t release any assets.

Guidelines:

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

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

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

Answer: [NA]

Justification: Our paper does not include such subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: Our paper does not include such subjects.

Guidelines:

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