Backward Feature Correction: How Deep Learning Performs Deep Learning

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
Microsoft Research Redmond

Yuanzhi Li
yuanzhil@andrew.cmu.edu
Carnegie Mellon University

Abstract

How does a 110-layer ResNet learn a high-complexity classifier using relatively few training examples and short training time? We present a theory towards explaining this in terms of *hierarchical learning*. We refer hierarchical learning as the learner *learns* to represent a complicated target function by decomposing it into a sequence of simpler functions to reduce sample and time complexity. This paper formally analyzes how multi-layer neural networks can perform such hierarchical learning *efficiently* and *automatically* simply by applying stochastic gradient descent (SGD) to the training objective.

On the conceptual side, we present, to the best of our knowledge, the *first* theory result indicating how *very deep* neural networks can still be sample and time efficient on certain hierarchical learning tasks, when ***no known*** non-hierarchical algorithms (such as kernel method, linear regression over feature mappings, tensor decomposition, sparse coding, and their simple combinations) are efficient. We establish a new principle called "backward feature correction", which we believe is the key to understand the hierarchical learning in multi-layer neural networks.

On the technical side, we show for regression and even for *binary classification*, for every input dimension $d > 0$, there is a concept class consisting of degree $\omega(1)$ multi-variate polynomials so that, using $\omega(1)$-layer neural networks as learners, SGD can learn any target function from this class in $\mathsf{poly}(d)$ time using $\mathsf{poly}(d)$ samples to any $\frac{1}{\mathsf{poly}(d)}$ regression or classification error, through *learning to represent it as a composition of $\omega(1)$ layers of quadratic functions*. In contrast, we present lower bounds stating that several non-hierarchical learners, including any kernel methods, neural tangent kernels, must suffer from super-polynomial $d^{\omega(1)}$ sample or time complexity to learn functions in this concept class even to any $d^{-0.01}$ error.

Prelude. Deep learning is sometimes also referred to as *hierarchical learning*.[1] In practice, multi-layer neural networks, as the most representative hierarchical learning model, often outperform non-hierarchical ones such as kernel methods, SVM over feature mappings, etc. However, from a theory standpoint,

Are multi-layer neural networks actually performing deep learning?

With huge non-convexity obstacles arising from the structure of neural networks, it is perhaps not surprising that existing theoretical works, to the best of our knowledge, have only been able to demonstrate that multi-layer neural networks can efficiently perform tasks that are *already known* solvable by non-hierarchical (i.e. shallow) learning methods. This is especially true for all recent results on based on neural tangent kernels [3, 5–8, 13, 15, 17, 18, 25, 28, 35, 40, 43, 57, 62]), which are just kernel methods instead of hierarchical learning. This is the motivation of our research to study the hierarchical learning process in multi-layer neural networks.

1 Introduction

How does a 110-layer ResNet [30] learn a high-complexity classifier for an image data set using *relatively few* training examples? How can the 100-th layer of the network discover a sophisticated function of the input image *efficiently* by simply applying stochastic gradient descent (SGD) on the training objective? In this paper, we present a theoretical result towards explaining this ***efficient deep learning*** process of such multi-layer neural networks in terms of *hierarchical learning*.

The term hierarchical learning in *supervised learning* refers to an algorithm that learns the target function (e.g. the labeling function) using a composition of simpler functions. The algorithm would first *learn* to represent the target function using simple functions of the input, and then aggregate these simple functions layer by layer to create more and more complicated functions to fit the target. Empirically, it has been discovered for a long time that hierarchical learning, in many applications, requires fewer training examples [11] when comparing to non-hierarchical learning methods that learn the target function in one shot.

Hierarchical learning is also everywhere around us. There is strong evidence that human brains perform learning in hierarchically organized circuits, which is the key for us to learn new concept class with relatively few examples [24]. Moreover, it is also observed that the human's decision making follows a hierarchical process as well: from "meta goals" to specific actions, which was the motivation for hierarchical reinforcement learning [38]. In machine learning, hierarchical learning is also the key to success for many models, such as hierarchical linear models [55], Bayesian networks/probabilistic directed acyclic graphical models [20], hierarchical Dirichlet allocation [52], hierarchical clustering [47], deep belief networks [39].

Hierarchical learning and multi-layer neural networks. Perhaps the most representative example of hierarchical learning is neural network. A multi-layer neural network is defined via layers $1, 2, \cdots, L$, where each layer represents a simple function (linear + activation) of the previous layers. Thus, multi-layer neural network defines a natural hierarchy: during learning, each network layer could use simple compositions of the *learnt functions* from previous layers to eventually represent the target function. Empirically, neural networks have shown great success across many different domains [26, 30, 37, 49]. Moreover, it is also well-known [59] that in learning tasks such as image recognition, each layer of the network indeed uses composition of previous layers to learn a function with an increasing complexity.

[1] `https://en.wikipedia.org/wiki/Deep_learning`.

In learning theory, however, little is known about hierarchical learning, especially for neural networks. Known results mostly focus on the representation power: there are functions that can be represented using 3-layer networks (under certain distributions), but requires an exponentially larger size to represent using 2-layer network [19]. However, the constructed function and data distribution in [19] separating the power of 2 and 3-layer networks are quite contrived.[2] When connecting it back to the actual learning process, to the best of our knowledge, there is no theoretical guarantee that training a 3-layer network from scratch (e.g. training via SGD from random initialization) can actually learn this separating function *efficiently*.[3] Hence, while the hierarchical structure of 3-layer network gives it more representation power than 2-layer ones, can the actual learning algorithm learn this "power of hierarchy" efficiently? In other words, from a theory point, these representation results can not answer the following question:

How can multi-layer neural networks perform efficient hierarchical learning when trained by SGD?

Before understanding "how" to this question, let us quickly mention to this date and to the best of our knowledge, it remains *even unclear* in theory whether for every $L \geq 3$, some L-layer neural network trained via SGD "can" actually use its hierarchical structure *as an advantage* to learn a function class *efficiently*, which is otherwise not efficiently learnable by non-hierarchical models. In fact, due to the extreme non-convexity in a multi-layer network, for theoretical purpose, the hierarchical structure is typically even a *disadvantage* for efficient training. One example of such "disadvantage" is deep linear network (DLN), whose hierarchical structure has no advantage over linear functions in representation power, but becomes an obstacle for training.[4]

In other words, not only "How can multi-layer neural networks perform efficient hierarchical learning?" is not answered in theory, but even the *significantly simpler question* "Can multi-layer neural networks efficiently learn simple functions that are already learnable by non-hierarchical models?" is non-trivial due to the extreme non-convexity caused by the hierarchical structures in multi-layer networks. For such reason, it is not surprising that most existing theoretical works on the efficient learning regime of neural networks either focus on (1) two-layer networks [9, 10, 12, 22, 23, 36, 41, 41, 42, 46, 50, 51, 53, 54, 56, 58, 61] which do not have any hierarchical structure, or (2) a multi-layer network but essentially only the last layer is trained [14, 33], or (3) reducing a multi-layer hierarchical neural network to non-hierarchical models such as kernel methods (a.k.a. the neural tangent kernel approach) [3, 5–8, 13, 13, 15, 17, 18, 25, 28, 35, 40, 43, 57, 62].

While the cited theoretical works shed great light on understanding the learning process of neural networks, *none of them* treat neural networks as a hierarchical learning model. Thus, we believe they are insufficient for understanding the ultimate power of neural networks. Motivated by such insufficiency, we propose to study the following fundamental question regarding the hierarchical learning in neural networks:

Question. *For every $L \geq 3$, can we prove that L-layer neural networks can efficiently learn a concept class, which is not learnable by any $(L-1)$ layer network of the same type (i.e. of the same activation function), and more importantly, not learnable by non-hierarchical methods such as the kernel methods (including neural tangent kernels defined by random initialized neural nets) or linear regression over feature mappings, given the same amount of sample and time complexity?*

[2] The function is heavily oscillating in the norm of the input, and the distribution is given by certain Bessel function with a lot of valleys and peaks.

[3] Indeed, since learning a 2-layer network with three hidden units is already NP-hard [27] due to the extreme non-convexity this, it could be challenging to show how such contrived concept class could be learnt by a even more non-convex multi-layer neural network efficiently from scratch.

[4] For instance, the theory for efficiently training DLNs only holds when the target linear function is well-conditioned [16, 29], and its running time bound is much worse than merely training a linear function as the learner.

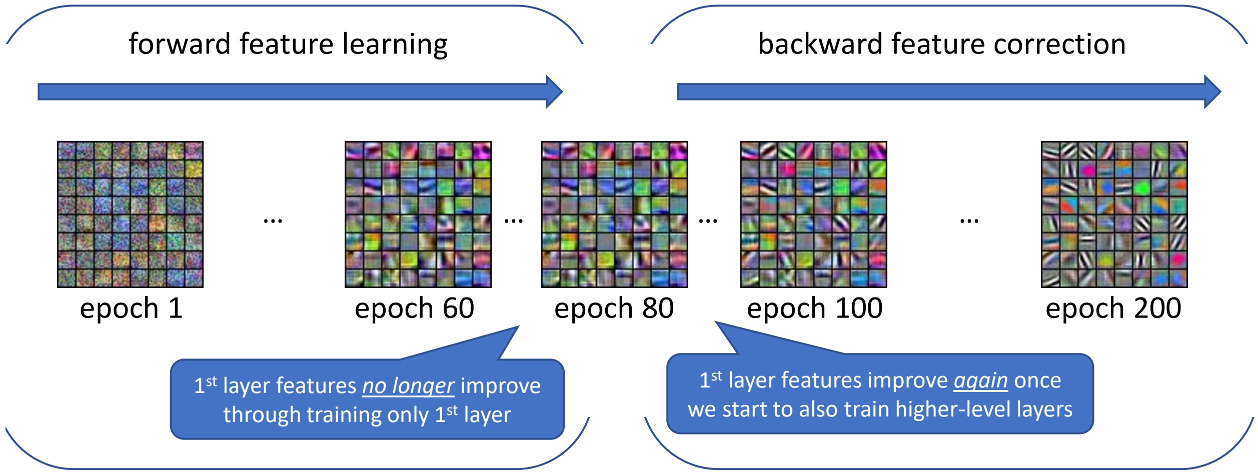

Figure 1: Convolutional features of the first layer in AlexNet. In the first 80 epochs, we train only the first layer, freezing layer 2 through 5; in the next 120 epochs, we train all the layers together (starting from the weights in epoch 80). Details in Appendix J.
Observation: In the first 80 epochs, when the first layer is trained until convergence, its features can already catch certain meaningful signals, but cannot get further improved. As soon as the 2nd through 5th layers are added to the training parameters, features of the first layer get improved again.

We consider a type of neural networks as the set of neural networks equipped with the same activation function σ. A positive answer to the first question indicates that going deeper in the network hierarchy can indeed learn a larger class of functions *efficiently*. A positive answer to the second question indicates that the hierarchical structure of the network is indeed used *as an advantage* comparing to non-hierarchical learning methods, hence the neural network is indeed performing hierarchical learning.

In this paper we give the first theoretical result towards answering this question: for every $L \geq 3$, there is certain type of L-layer neural networks equipped with quadratic activation functions so that training such networks by SGD indeed *efficiently* and *hierarchically* learns a concept class. Here, "efficient" means that to achieve any inverse polynomial generalization error, the number of training examples required to train the network is *polynomial* in the input dimension, and the total running time is also *polynomial*. Moreover, we also give lower bounds showing that this concept class is not learnable by non-hierarchical learning methods such as *any* kernel method (in particular, including the neural tangent kernel given by the initialization of the learner network) or linear regression over feature mappings, or even two-layer networks with certain polynomial activation functions, require super-polynomial sample or time complexity.

How can deep learning perform hierarchical learning? Our paper not only proves such a separation, but also gives, to the best of our knowledge, the first result showing *how* deep learning can actually perform hierarchical learning when trained by SGD. We identify *two critical steps* in the hierarchical learning process (see also Figure 1):

- The ***forward feature learning*** step, where a higher-level layer can learn its features using the simple combinations of the learnt features from lower-level layers.
- The ***backward feature correction*** step, where a lower-level layer can also learn to further *improve* the quality of its features using the learnt features in higher-level layers.

While "forward feature learning" is standard in theory, to the best of our knowledge, "backward feature correction" is not yet recorded anywhere in the theory literature. As we demonstrate both in theory and in experiment (see Figure 1), this is a most critical step in the hierarchical learning process of multi-layer neural networks, and we view it as the *main conceptual contribution* of this paper.

1.1 Our Theorem

Let us now go into notations. The type of networks we consider is DenseNet [32]:

$$G(x) = \sum_{\ell} u_\ell^\top G_\ell(x) \in \mathbb{R} \qquad \text{where} \quad G_0(x) = x \in \mathbb{R}^d, \qquad\qquad G_1(x) = \sigma(x) - \mathbb{E}[\sigma(x)] \in \mathbb{R}^d$$

$$G_\ell(x) = \sigma\left(\textstyle\sum_{j \in \mathcal{J}_\ell} \mathbf{W}_{\ell,j} G_j(x)\right) \in \mathbb{R}^{k_\ell} \qquad \text{for } \ell \geq 2 \text{ and } \mathcal{J}_\ell \subseteq \{0, 1, \cdots, \ell-1\}$$

Here, σ is the activation function, where we pick $\sigma(z) = z^2$ in this paper, $\mathbf{W}_{\ell,j}$ are weight matrices, and the final (1-dimensional) output $G(x)$ is a weighted summation of the outputs of all layers. The set $\mathcal{J}_\ell$ defines the connection graph (the structure of the network). For vanilla feed-forward network, it corresponds to $\mathcal{J}_\ell = \{\ell - 1\}$ so each layer only uses information from the immediate previous layer. ResNet [30] (with skip connection) corresponds to $\mathcal{J}_\ell = \{\ell-1, \ell-3\}$ with weight sharing (namely, $\mathbf{W}_{\ell,\ell-1} = \mathbf{W}_{\ell,\ell-3}$). In this paper, we can handle *any* connection graph with the only restriction being there is at least one "skip link," or in symbols, for every $\ell \geq 3$, we require $(\ell-1) \in \mathcal{J}_\ell$, $(\ell-2) \notin \mathcal{J}_\ell$ but $j \in \mathcal{J}_\ell$ for some $j \leq \ell - 3$.

One of the main reasons we pick quadratic activation $\sigma(z)$ is because it is easy to measure the network's representation power. Clearly, in quadratic DenseNet, each layer learns a quadratic function of the (weighted) summation of previous layers, so in layer ℓ, the hidden neurons represent a degree-2^ℓ multivariate polynomial of the input x. Hence, the concept class that can be represented by L-layer quadratic DenseNet is obviously increasing with L. The question remains to answer is: *Can L-layer quadratic DenseNet use its hierarchical structure* **as an advantage** *to learn certain class of degree-2^L polynomials more efficiently than non-hierarchical models?*

We answer this positively. Our main result can be sketched as the follows:

Theorem (informal)**.** *For every input dimension $d > 0$ and $L = o(\log\log d)$, there is a class of degree-2^L polynomials and input distributions such that,*

- *Given* $\mathsf{poly}(d/\varepsilon)$ *training samples and* $\mathsf{poly}(d/\varepsilon)$ *running time, by performing SGD over the training objective starting from random initialization, the L-layer quadratic DenseNet can learn any function in this concept class with any generalization error $\varepsilon > 0$.*
- *Any kernel method, any linear regression over feature mappings, or any two-layer neural networks equipped with arbitrary degree-2^L activations, require either $d^{\omega(1)}$ sample or time complexity, to achieve any non-trivial generalization error such as $\varepsilon = d^{-0.01}$, for any $L = \omega(1)$.*

The concept class (the class of functions to be learnt) considered in this paper is simply given by L-layer quadratic DenseNets with $\sim d^{1/2^\ell}$ neurons in the ℓ-th layer. Thus, each function in this concept class is equipped with a hierarchical structure defined by DenseNet, and our positive result is "using DenseNet to learn the hierarchical structure of an unknown DenseNet."

We also point out that in our result, the learner network has $\mathsf{poly}(d)$ neurons while the target network in the concept class has $\leq O(d^2)$ neurons. Thus, the learner network is *over-parameterized*, which is standard in deep learning. However, the necessity of over-parameterization here is for a very different reason comparing to existing theory work [6]. We will discuss more in Section 6.1.

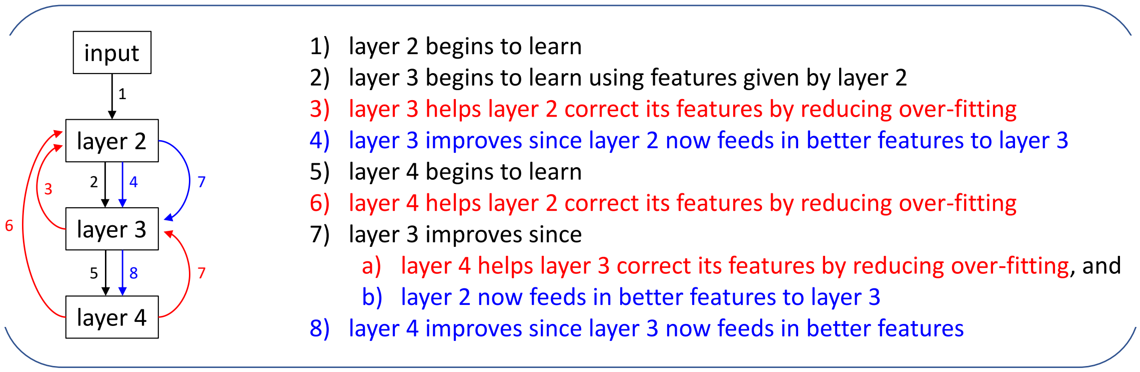

Figure 2: We explain the hierarchical learning process in a 4-layer example. The **back** and **blue** arrows correspond to "forward feature learning". The **red** arrows correspond to "backward feature correction".

1.2 Our Conceptual Message: How Deep Learning Performs Hierarchical Learning

Intuitively, the polynomials in our concept class are of degree 2^L and can depend on $d^{\Omega(1)}$ *unknown* directions of the input. Thus, when $L = \omega(1)$, using non-hierarchical learning methods, the typical sample or time complexity is at least $(d^{\Omega(2^L)}) = d^{\omega(1)}$ (and we have shown lower bounds for kernel methods and linear regression over feature mappings). Even if the learner performs hierarchical learning for $O(1)$ levels, it still cannot avoid learning in one level a degree-$\omega(1)$ polynomial that depends on $d^{\Omega(1)}$ variables, which typically requires sample/time complexity $d^{\omega(1)}$.

In contrast, our quadratic DenseNet only uses sample and time complexity $\mathsf{poly}(d)$. The efficiency gain is due to the fact that the learning process is highly hierarchical: the network first learns a crude degree-4 approximation of the target function, and then it identifies hidden features and use them to fit a degree-8 approximation of the target function (using degree-2 polynomial over hidden features). Following this fashion, the learning task decomposes from learning a degree 2^L polynomial in one-shot which requires time $d^{\omega(1)}$, into learning one quadratic functions at a time for $\omega(1)$ times, which can be done individually in time $\mathsf{poly}(d)$. This is, from a high level, where the efficiency gain comes from, but there is more to say:

Critical observation: backward feature correction. In our quadratic DenseNet model, when training the first layer of the learner network, it tries to fit the target function using the best degree-4 polynomial. This polynomial might not be the one used by the target network due to over-fitting to higher-degree terms. As a concrete example, the best degree-4 polynomial to fit $x_1^4 + x_2^4 + 0.3(x_1^4 + x_2^4)^2$ is usually *not* $x_1^4 + x_2^4$, even under Gaussian distribution. However, through the hierarchical learning process, those higher-degree terms in $0.3(x_1^4 + x_2^4)^2$ will gradually get discovered by the *higher levels* of the learner network and "subtracted" from the training objective.

As a result, the *features* (i.e., intermediate outputs) of lower levels of the learner network can get improved due to less over-fitting. We explain this phenomenon more carefully in Figure 2 for the case of a 4-layer network. It provides theoretical explanation towards how lower levels get improved through hierarchical learning when we train lower and higher levels *together*.

Hierarchical learning is NOT layer-wise training. Our result also shed lights on the following critical observation in practice: typically layer-wise training (i.e. if we train layers one by one starting from lower levels)[5] performs much worse comparing to training all the layers together.

[5]Layer-wise training means first training the 1st hidden layer by setting other layers to zero, and then training the

The fundamental reason is due to the missing piece of "backward feature correction": *The function learnt by the lower levels is not accurate enough if we only train lower levels; by training lower and higher levels together, the functions generated by lower layers also get improved.*

Hierarchical learning is NOT simulating known (non-hierarchical) algorithms. To the best of our knowledge, this seems to be the *first theory result in the literature* for training a neural network via SGD, to solve an underlying problem not yet known solvable by existing algorithms, such as kernel methods (including applying kernel methods multiple times), tensor decomposition methods, etc. Thus, neural network training could be indeed performing hierarchical learning, instead of simulating known (non-hierarchical) algorithms.

1.3 More on Related Works

Learning Two-Layer Network [9, 10, 12, 22, 23, 36, 41, 41, 42, 46, 50, 51, 53, 54, 56, 58, 61]. There is a rich history of works considering the learnability of neural networks trained by SGD. However, as we mentioned before, many of these works only focus on network with 2 layers or only one layer in the network is trained. Hence, the learning process is not hierarchical.

Neural Tangent Kernel [3, 5–8, 13, 15, 17, 18, 25, 28, 35, 40, 43, 57, 62]. There is a rich literature approximating the learning process of over-parameterized networks using the neural tangent kernel (NTK) approach, where the kernel is defined by the gradient of a neural network at random initialization [35]. We stress that one *should not confuse* this hierarchically-defined kernel with a multi-layer network with hierarchical learning in the paper. As we pointed out, hierarchical learning means that each layer *learns* a combination of previous layers. In NTK, such combinations are *prescribed* by the random initializations of the neural network, which are not learnt during the training process. As our negative result shows, for certain learning tasks, hierarchical learning is indeed superior than any kernel method, including those hierarchical-defined kernels prescribed from any neural networks. Hence, in this task, the *learnt* combinations are indeed *superior* to the randomly prescribed ones given by the initialization of the network.

Three Layer Result [4]. This paper shows that 3-layer neural networks can learn the so-called "second-order NTK," which is not a linear model; however, second-order NTK is also learnable by doing a nuclear norm constraint linear regression over the feature mappings defined by the initialization of a neural network. Thus, the underlying learning process is still not truly hierarchical.

Three-Layer ResNet Result [2]. This recent paper shows that 3-layer ResNet can perform some *weaker* form of implicit hierarchical learning, with better sample or time complexity than any kernel method or linear regression over feature mappings. Our result is greatly inspired by [2], but with several major differences.

First and foremost, the result [2] can also be achieved by non-hierarchical methods such as *simply* applying kernel method twice— essentially layer-wise learning without backward feature correction.[6] Thus, the work [2] is *a weaker version of hierarchical learning without backward feature correction.*

2nd layer by fixing the 1st layer and setting others to zero. Such algorithm is used in theoretical works such as [45].

[6] Recall the target functions from [2] are of the form $F(x) + \alpha \cdot G(F(x))$ for $\alpha \ll 1$, and they were shown to be learnable by 3-layer ResNet up to generalization error α^2 in [2]. Here is a *simple* alternative two-step kernel method to achieve this same result. First, learn some $F'(x)$ that is α-close to $F(x)$ using kernel method. Then, use $(x, F'(x))$ as the input to learn $F(x) + \alpha G(F'(x))$ by kernel method, incurring a fixed generalization error of magnitude α^2. Note in particular, both this two-step kernel method as well as the 3-layer ResNet analysis from [2] *never* guarantees to learn any function $F''(x)$ that is α^2 close to $F(x)$, and therefore the "intermediate features" do not get improved. In other words, there is no backward feature correction.

Second, we prove in this paper a "poly vs. super-poly" running time separation, which is what one refers to as "efficient vs non-efficient" in the traditional theoretical computer science language. In contrast, the result in [2] is regarding "poly vs. bigger poly" (in the standard regime where the output dimension is constant).[7]

Third, without backward feature correction, the error incurred from lower layers in [2] cannot be improved through training (see Footnote 6), and thus their theoretical result does not lead to arbitrarily small generalization error like we do in this paper. This also prevents [2] from going beyond $L = 3$ layers; our result in this paper holds for *every* $L \geq 3$, demonstrating that going deeper in the hierarchy can actually have consistent advantage.

2 Target Network and Learner Network

We consider a target network defined as

$$G^\star_0(x) = x \in \mathbb{R}^d, \quad G^\star_1(x) = \sigma(x) - \mathbb{E}[\sigma(x)] \in \mathbb{R}^d, \quad G^\star_\ell(x) = \sigma\left(\sum_{j\in\mathcal{J}_\ell} \mathbf{W}^\star_{\ell,j} G^\star_j(x)\right) \in \mathbb{R}^{k_\ell} \quad \forall \ell \geq 2$$

where the weight matrices $\mathbf{W}^\star_{\ell,j} \in \mathbb{R}^{k_\ell \times k_j}$ for every ℓ, j. Each index set $\mathcal{J}_\ell$ is a subset of $\{0, 1, 2, \cdots, \ell-3\} \cup \{\ell - 1\}$. We assume that (1). $\ell - 1 \in \mathcal{J}$ (so there is a connection to the immediate previous layer) and (2). for every $\ell \geq 3$, $|\mathcal{J}_\ell| \geq 2$ (so there is at least one skip connection).

We consider target functions $G^\star \colon \mathbb{R}^d \to \mathbb{R}$ consisting of the coordinate summation of each layer:

$$G^\star(x) = \sum_{\ell=2}^L \sum_{i\in[k_\ell]} \alpha_\ell G^\star_{\ell,i}(x) = \sum_{\ell=2}^L \alpha_\ell \mathbf{Sum}(G^\star_\ell(x))$$

where $\mathbf{Sum}(v) = \sum_i v_i$,[8] and it satisfies $\alpha_2 = 1$ and $\alpha_{\ell+1} < \alpha_\ell$. We will provide more explanation of α_ℓ at Section 3. For analysis purpose, we use the convention $\mathbf{W}^\star_{\ell,j} = 0$ if $j \notin \mathcal{J}_\ell$, and define

$$S^\star_0(x) = G^\star_0(x) = x, \quad S^\star_1(x) = G^\star_1(x), \quad S^\star_\ell(x) \stackrel{\text{def}}{=} \sum_{j\in\mathcal{J}_\ell} \mathbf{W}^\star_{\ell,j} G^\star_j(x) = \sum_{j=0}^{\ell-1} \mathbf{W}^\star_{\ell,j} G^\star_j(x) \quad \forall \ell \geq 2$$

We remark here that for $\ell \geq 2$, $S^\star_\ell(x)$ is of degree $2^{\ell-1}$ and $G^\star_\ell(x) = \sigma(S^\star_\ell(x))$ is of degree 2^ℓ. It is convenient to think of $S^\star_\ell(x)$ as the "features" used by target network $G^\star(x)$.

2.1 Learner Network

Typically, for theory, by training a learner neural network, the objective is to construct network G of the same structure (possibly with over-parameterization) so it simulates $G^\star$:

$$G_\ell(x) = \sigma\left(\sum_{j\in\mathcal{J}_\ell} \mathbf{M}_{\ell,j} G_j(x)\right) \quad .$$

[7] The result [2] only works for a concept class whose functions contain *merely* networks with "number of hidden neurons = output dimension." Putting into the case of this paper, the output dimension is 1, so the result [2] only supports networks with one hidden neuron, and gives no separation between neural networks and kernel methods. When the output dimension is $O(1)$, they give separation between d and $d^{O(1)}$ which is "poly vs bigger poly". We shall illustrate in Section 6 that the *major* technical difficulty of this paper comes from extending the number of hidden neurons in the concept class *beyond* the output dimension.

For experts familiar with [2], they only proved that hierarchical learning happens when the output vector contains explicit information about the intermediate output. In symbols, their target network is $y = F(x) + \alpha \cdot G(F(x))$, so the output label y is a vector that has explicit information of the vector $F(x)$ up to error α. In this paper, we show that the network can discover hidden feature *vectors* from the target function, even if the output dimension is 1 such as $y = u^\top F(x) + \alpha \cdot v^\top G(F(x))$.

[8] Our result trivially extends to the case when $\mathbf{Sum}(v)$ is replaced with $\sum_i p_i v_i$ where $p_i \in \{\pm 1\}$ for half of the indices. We refrain from proving that version for notational simplicity.

Here, $G_0(x) = x, G_1 = G_1^\star(x)$ and we choose $\mathbf{M}_{\ell,0}, \mathbf{M}_{\ell,1} \in \mathbb{R}^{\binom{k_\ell+1}{2}\times d}$ and $\mathbf{M}_{\ell,j} \in \mathbb{R}^{\binom{k_\ell+1}{2}\times\binom{k_j+1}{2}}$ for every $2 \le j \le \ell - 1$. In other words, the amount of over-parameterization is quadratic (i.e., from $k_j \to \binom{k_j+1}{2}$) per layer. We want to construct the weight matrices so that

$$G(x) = \textstyle\sum_{\ell=2}^{L} \alpha_\ell \mathbf{Sum}(G_\ell(x)) \approx G^\star(x) \ .$$

Learner Network Re-parameterization. In this paper, it is more convenient to consider the re-parameterized network $F(x)$: We first re-parameterize the weight matrix $\mathbf{M}_{\ell,j} = \mathbf{R}_\ell \mathbf{K}_{\ell,j}$, where

- $\mathbf{R}_\ell \in \mathbb{R}^{\binom{k_\ell+1}{2}\times k_\ell}$ are randomly initialized for all $\ell \in \{2, 3, \cdots, L-1\}$, not changed during training;
- weights $\mathbf{W}_{\ell,j} \in \mathbb{R}^{m\times q}, \mathbf{K}_{\ell,j} \in \mathbb{R}^{k_\ell \times q}$ are trainable, for every $\ell \in [L]$ and $j \in \mathcal{J}_\ell$, and the dimension $q = \binom{k_j+1}{2}$ for $j \ge 2$ and $q = d$ for $j = 0, 1$.

Define functions $S_0(x) = G_0^\star(x)$, $S_1(x) = G_1^\star(x)$,[9] as well as (it is convenient to think of those $S(x)$ as the "features" used by learner network $F(x)$)

$$\forall \ell \ge 2: \qquad S_\ell(x) = \textstyle\sum_{j\in\mathcal{J}_\ell, j\ge 2} \mathbf{K}_{\ell,j}\sigma\left(\mathbf{R}_j S_j(x)\right) + \sum_{j\in\{0,1\}\cap\mathcal{J}_\ell} \mathbf{K}_{\ell,j} S_j(x) \in \mathbb{R}^{k_\ell} \tag{2.1}$$

$$\forall \ell \ge 2: \qquad F_\ell(x) = \sigma\Big(\textstyle\sum_{j\in\mathcal{J}_\ell, j\ge 2} \mathbf{W}_{\ell,j}\sigma\left(\mathbf{R}_j S_j(x)\right) + \sum_{j\in\{0,1\}\cap\mathcal{J}_\ell} \mathbf{K}_{\ell,j} S_j(x)\Big) \in \mathbb{R}^{m} \tag{2.2}$$

We define its final output

$$F(x) = \textstyle\sum_{\ell=2}^{L} \alpha_\ell \mathbf{Sum}(F_\ell(x))$$

and we shall use this function to fit the target $G^\star(x)$.

It is easy to verify that when $\mathbf{R}_\ell^\top \mathbf{R}_\ell = \mathbf{I}$ and when $\mathbf{W}_{\ell,j} = \mathbf{K}_{\ell,j}$, by defining $\mathbf{M}_{\ell,j} = \mathbf{R}_\ell \mathbf{K}_{\ell,j}$ we have that each $F_\ell(x) = G_\ell(x)$ and $F(x) = G(x)$. In this paper, we will mostly work with this re-parameterization F for *efficient training purpose*. As we shall see, we will impose regularizers on $\mathbf{W}, \mathbf{K}$ during the training to enforce that they are close to each other. The idea of using a larger unit (i.e., $\mathbf{W}$) for training and using a smaller unit (i.e., $\mathbf{K}$) to learn the larger one is called *knowledge distillation*, which is commonly used in practice [31].

Truncated Quadratic Activation. To make our analysis simpler, it would be easier to work with an activation function that has bounded derivatives in the entire space. For each layer ℓ, we consider a "truncated but smoothed" version of the square activation function $\widetilde{\sigma}_\ell(z)$ defined as follows. For some sufficiently large B'_ℓ (to be chosen later), and setting $B''_\ell = \Theta((B')^2_\ell))$, let

$$\widetilde{\sigma}_\ell(z) = \begin{cases} \sigma(z), & \text{if } |z| \le B'_\ell \\ B''_\ell & \text{if } |z| \ge 2B'_\ell \end{cases}$$

and in the range $[B'_\ell, 2B'_\ell]$, function $\widetilde{\sigma}(z)$ can be chosen as any monotone increasing function such that $|\widetilde{\sigma}_\ell(z)'|, |\widetilde{\sigma}_\ell(z)''|, |\widetilde{\sigma}_\ell(z)'''| = O(B'_\ell)$ are bounded for every z. Our final choice of B'_ℓ will make sure that when taking expectation over data, the difference between $\widetilde{\sigma}_\ell(z)$ and $\sigma(z)$ is negligible.

Accordingly, we define the network with respect to the truncated activation as follows.

$$\widetilde{S}_0(x) = G_0^\star(x) \ , \quad \widetilde{S}_1(x) = G_1^\star(x), \quad \widetilde{S}_\ell(x) = \textstyle\sum_{j\in\mathcal{J}_\ell, j\ge 2} \mathbf{K}_{\ell,j}\widetilde{\sigma}_j\left(\mathbf{R}_j \widetilde{S}_j(x)\right) + \sum_{j\in\{0,1\}\cap\mathcal{J}_\ell} \mathbf{K}_{\ell,j}\widetilde{S}_j(x)$$

$$\widetilde{F}(x) = \textstyle\sum_{\ell=2}^{L} \alpha_\ell \mathbf{Sum}(\widetilde{F}_\ell(x)) \ , \quad \widetilde{F}_\ell(x) = \sigma\left(\sum_{j\in\mathcal{J}_\ell, j\ge 2} \mathbf{W}_{\ell,j}\widetilde{\sigma}_j\left(\mathbf{R}_j \widetilde{S}_j(x)\right) + \sum_{j\in\{0,1\}\cap\mathcal{J}_\ell} \mathbf{K}_{\ell,j}\widetilde{S}_j(x)\right)$$

[9]Typically, during the training we only have access to the empirical expectation of $\sigma(x)$, however using $\mathsf{poly}(d/\varepsilon)$ samples, the empirical expectation would be $\frac{1}{\mathsf{poly}(d/\varepsilon)}$. For cleanness, we just write in both $G_1^\star$ and S_1 the true expectation, we the difference can be easily dealt by a Lipschitz argument in Section C.3

The truncated function $\widetilde{F}$ is only for *training propose* to ensure the network is Lipschitz, so we can obtain efficient running time. The original quadratic activation σ does not have an absolute Lipschitz bound. We also use $\widetilde{\sigma}$ instead of $\widetilde{\sigma}_j$ when its clear from content.

For notational simplicity, we concatenate the weight matrices used in the same layer ℓ as follows:

$$\mathbf{W}_\ell = (\mathbf{W}_{\ell,j})_{j\in\mathcal{J}_\ell} \qquad \mathbf{K}_\ell = (\mathbf{K}_{\ell,j})_{j\in\mathcal{J}_\ell} \qquad \mathbf{W}^\star_\ell = \left(\mathbf{W}^\star_{\ell,j}\right)_{j\in\mathcal{J}_\ell}$$
$$\mathbf{W}_{\ell\triangleleft} = (\mathbf{W}_{\ell,j})_{j\in\mathcal{J}_\ell, j\neq\ell-1} \qquad \mathbf{K}_{\ell\triangleleft} = (\mathbf{K}_{\ell,j})_{j\in\mathcal{J}_\ell, j\neq\ell-1} \qquad \mathbf{W}^\star_{\ell\triangleleft} = \left(\mathbf{W}^\star_{\ell,j}\right)_{j\in\mathcal{J}_\ell, j\neq=\ell-1}$$

2.2 Training Objective

For simplicity, we first state our result for ℓ_2 regression problem in the realizable case, where we simply want to minimize the difference between the output of the learner network $F(x)$ and the labels $G^\star(x)$, we will state the result for agnostic case and for classification in the next section.

Intuitively, we shall add a regularizer to ensure that $\mathbf{W}_\ell^\top \mathbf{W}_\ell \approx \mathbf{K}_\ell^\top \mathbf{K}_\ell$, that is $\mathbf{K}_\ell^\top \mathbf{K}_\ell$ is a low-rank approximation of $\mathbf{W}_\ell^\top \mathbf{W}_\ell$. This ensures that $\mathbf{Sum}(F_\ell(x)) \approx \mathbf{Sum}(\sigma(S_\ell(x)))$. The main reason for this "low rank approximation" is explained in Section 6. Furthermore, we shall add a regression loss to minimize $\left(G^\star(x) - F(x)\right)^2$. This ensures that

$$G^\star(x) \approx F(x) = \sum_\ell \alpha_\ell \mathbf{Sum}(F_\ell(x)) \approx \sum_\ell \alpha_\ell \mathbf{Sum}(\sigma(S_\ell(x))) \ .$$

Specifically, we use the following training objective:

$$\widetilde{\mathbf{Obj}}(x; \mathbf{W}, \mathbf{K}) = \widetilde{\mathbf{Loss}}(x; \mathbf{W}, \mathbf{K}) + \mathbf{Reg}(\mathbf{W}, \mathbf{K})$$

where the ℓ_2 loss is $\widetilde{\mathbf{Loss}}(x; \mathbf{W}, \mathbf{K}) = (G^\star(x) - \widetilde{F}(x))^2$ and

$$\mathbf{Reg}(\mathbf{W}, \mathbf{K}) = \textstyle\sum_{\ell=2}^L \lambda_{3,\ell} \left\|\mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell\triangleleft} - \mathbf{W}_{\ell,\ell-1}^\top \mathbf{W}_{\ell\triangleleft}\right\|_F^2 + \sum_{\ell=2}^L \lambda_{4,\ell} \left\|\mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1} - \mathbf{W}_{\ell,\ell-1}^\top \mathbf{W}_{\ell,\ell-1}\right\|_F^2$$
$$+ \textstyle\sum_{\ell=2}^L \lambda_{5,\ell} \left\|\mathbf{K}_\ell^\top \mathbf{K}_\ell - \mathbf{W}_\ell^\top \mathbf{W}_\ell\right\|_F^2 + \sum_{\ell=2}^L \lambda_{6,\ell} \left(\|\mathbf{K}_\ell\|_F^2 + \|\mathbf{W}_\ell\|_F^2\right)$$

and for a given set $\mathcal{Z}$ consisting of N i.i.d. samples from the true distribution $\mathcal{D}$, we minimize ($x \sim \mathcal{Z}$ denotes x is uniformly sampled from the training set $\mathcal{Z}$)

$$\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) = \mathop{\mathbb{E}}_{x\sim\mathcal{Z}}[\widetilde{\mathbf{Obj}}(x; \mathbf{W}, \mathbf{K})] \tag{2.3}$$

The other regularizers we used are just (squared) Frobenius norm on the weight matrices, which are used everywhere in practice. For the original quadratic activation network, we also denote by

$$\mathbf{Loss}(x; \mathbf{W}, \mathbf{K}) = (G^\star(x) - F(x))^2 \text{ and } \mathbf{Obj}(x; \mathbf{W}, \mathbf{K}) = \mathbf{Loss}(x; \mathbf{W}, \mathbf{K}) + \mathbf{Reg}(\mathbf{W}, \mathbf{K}).$$

3 Statement of Main Result

For simplicity, we only state here as a special case of our main theorem which is sufficiently interesting, and the full theorem can be found at Appendix A.

In this special case, there exists an absolute constants $C > C_1 > 1$ such that, for every $d, L > 0$, consider any target network $G^\star(x)$ and underlying data distribution $\mathcal{D}$ satisfying some properties (namely, properties defined in Section 5 with $\kappa \leq 2^{C_1^L}, B_\ell \leq 2^{C_1^\ell} k_\ell$). Suppose in addition that the network width is diminishing $k_\ell \leq d^{\frac{1}{C^{\ell+C_1}}}$ and there is an information gap $\frac{\alpha_{\ell+1}}{\alpha_\ell} \leq d^{-\frac{1}{C^\ell}}$ for $\ell \geq 2$. Moreover, we assume in the connection graph $\{2, 3, \cdots, \ell - C_1\} \cap \mathcal{J}_\ell = \varnothing$, meaning that the skip connections do not go very deep, unless directly connected to the input.

Theorem 3.1 (special case)**.** *For every $d > 0$, every $L = o(\log\log d)$, every $\varepsilon \in (0,1)$, and every target network satisfying the above parameters. Then, given $N = \mathsf{poly}(d/\varepsilon)$ i.i.d. samples $\mathcal{Z}$ from $\mathcal{D}$, by applying SGD over the training objective (2.3), with probability at least 0.99, we can find a learner network F in time $\mathsf{poly}(d/\varepsilon)$ such that:*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}} (G^\star(x) - F(x))^2 \leq \varepsilon^2 \quad \textit{and} \quad \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left(G^\star(x) - \widetilde{F}(x)\right)^2 \leq \varepsilon^2 \ .$$

Note $\alpha_{\ell+1} = \alpha_\ell d^{-\frac{1}{C^\ell}}$ implies $\alpha_L \geq d^{-\frac{1}{C}} \geq \frac{1}{d}$. Hence, when for instance $\varepsilon \leq \frac{1}{d^4}$, to achieve ε^2 regression error, the learning algorithm has to truly learn *all* the layers of $G^\star(x)$, as opposed to for instance ignoring the last layer which will incur error $\alpha_L \gg \varepsilon$. We give more details about the training algorithm in Section 4.

Comparing to Kernel Methods. The target function in $G^\star(x)$ is of degree 2^L. We show as a lower bound in Appendix H.1 that, any kernel method must suffer sample complexity $\Omega(d^{2^{L-1}}) = d^{\omega(1)}$ when $L = \omega(1)$, even when all $k_\ell = 1$. This is due to the fact that kernel methods *cannot* perform hierarchical learning so have to essentially "write down" all the monomials of degree 2^{L-1} in order to express the target function, which suffers a lot in the sample complexity.

On the other hand, one might hope for a "*sequential kernel*" learning of this target function, by first applying kernel method to identify degree-$O(1)$ polynomials used in $G^\star(x)$ (e.g., $G^\star_\ell(x)$ and $S^\star_\ell(x)$ for $\ell \leq O(1)$), and then use them as features to learn higher degrees. We point out:

- Even if we know all the degree-$O(1)$ polynomials, the network width k_ℓ at layer $\ell = O(1)$ can still be as large as $d^{\Omega(1)}$, so we still need to learn a degree $2^{\Omega(L)}$ polynomial over dimension $d^{\Omega(1)}$. This cannot be done by kernel method with $\mathsf{poly}(d)$ sample complexity.
- Even if we do "sequential kernel" for $\omega(1)$ rounds, this is similar to layer-wise training and misses the crucial "backward feature correction." As we pointed out in the intro, and shall later explain Section 6.1, this is unlikely to recover the target function to good accuracy.
- Even if we do "sequential kernel" together with "backward feature correction", this may not work since the backward correction may not lead to sufficient accuracy on intermediate features. Concretely, say we optimistically know the feature mappings $S^\star_\ell(x)$ up to error ε for $\ell \leq L-1$, and fit the target function by kernel method on top of features $\{S^\star_\ell\}_{\ell\leq L-1}$. This *does not* mean we can obtain $\mathbf{W}$ that is $O(\varepsilon)$ close to $\mathbf{W}^\star$ (sophisticated reasons deferred to Section 6). Thus, we cannot improve the quality of features $S^\star_\ell$ of previous layers.[10]

Significance of Our Result? To the best of our knowledge,

- We do not know any other simple algorithm that can learn the target functions considered in this paper within the same efficiency, the only simple learning algorithm we are aware of is to train a neural network to perform hierarchical learning.
- This seems to be the *only theory result in the literature* for training a neural network via SGD, to solve an underlying problem that is *not* known solvable by existing algorithms, such as kernel methods (including applying kernel methods multiple times), tensor decomposition methods, sparse coding, etc. Thus, the neural network is indeed performing hierarchical learning, instead of simulating known (non-hierarchical) algorithms.

Agnostic Learning. Our theorem also works in the agnostic setting, where the labeling function $Y(x)$ satisfies $\mathbb{E}_{x\sim\mathcal{D}}(G^\star(x) - Y(x))^2 \leq \mathsf{OPT}$ and $|G^\star(x) - Y(x)| \leq \mathsf{poly}(d)$ for some *unknown* $G^\star(x)$.

[10] One may want to connect this to [2]: according to Footnote 6, the analysis from [2] is analogous to doing "sequential kernel" for 2 rounds, but even if one wants to backward correct the features of the first hidden layer, its error remains to be α and cannot be improved to α^2 (although the regression output is already α^2).

The SGD algorithm can learn a function $F(x)$ with error at most $(1+\gamma)\mathsf{OPT}+\varepsilon^2$ for any constant $\gamma > 1$ given i.i.d. samples of $\{x, Y(x)\}$. Thus, the learner can *compete* with the performance of the best target network. We present the result in Appendix A.4 and state its special case below.

Theorem 3.2 (special case, agnostic)**.** *For every constant $\gamma > 0$, in the same setting Theorem 3.1, given $N = \mathsf{poly}(d/\varepsilon)$ i.i.d. samples $\mathcal{Z}$ from $\mathcal{D}$ and given their corresponding labels $\{Y(x)\}_{x\in\mathcal{Z}}$, by applying SGD over the training objective $\mathbb{E}_{x\sim\mathcal{Z}}\left(Y(x)-\widetilde{F}(x)\right)^2+\mathbf{Reg}(\mathbf{W},\mathbf{K})$, with probability at least 0.99, we can find a learner network F in time $\mathsf{poly}(d/\varepsilon)$ such that:*

$$\mathbb{E}_{x\sim\mathcal{D}}\left(F(x)-Y(x)\right)^2 \le \varepsilon^2 + (1+\gamma)\mathsf{OPT}$$

3.1 On Information Gap α_ℓ and classification problem

We have made a gap assumption $\frac{\alpha_\ell}{\alpha_\ell} \le d^{-\frac{1}{C^{\ell+1}}}$. We can view this "gap" as that in the target function $G^\star(x)$, *higher levels contribute less to its output.* This is typical for tasks such as image classification on CIFAR-10, where the first convolutional layer can already be used to classify $> 60\%$ of the data correctly. The higher-level layers have diminishing contributions to the signal (see Figure 3 for an illustration and we also refer to [60] for concrete measures). We emphasize, in practice, researchers do *fight for* even the final 0.1% performance gain by going for (much) larger networks, so those higher-level functions *can not be ignored.*

To formally justify this gap assumption, it is also beneficial to consider a *classification* problem. Let us w.l.o.g. scale $G^\star(x)$ so that $\mathbf{Var}_x[G^\star(x)] = 1$, and consider a two-class labeling function $Y(x_0, x)$ given as:

$$Y(x_0, x) = \mathsf{sgn}(x_0 + G^\star(x)) \in \{-1, 1\} \enspace ,$$

where $x_0 \sim \mathcal{N}(-\mathbb{E}_x[G^\star(x)], 1)$ is a Gaussian random variable independent of x. Here, x_0 can be viewed either a coordinate of the entire input $(x_0, x) \in \mathbb{R}^{d+1}$, or more generally as linear direction $x_0 = w^\top \widehat{x}$ for the input $\widehat{x} \in \mathbb{R}^{d+1}$. For notation simplicity, we focus on the former view.

Using probabilistic arguments, one can derive that except for α_ℓ fraction of the input $(x_0, x) \sim \mathcal{D}$, the label function $Y(x_0, x)$ is fully determined by the target function $G^\star(x)$ up to layer $\ell - 1$; or in symbols,[11]

$$\mathop{\mathbf{Pr}}_{(x_0,x)\sim\mathcal{D}}\left[Y(x_0,x) \neq \mathsf{sgn}\left(x_0 + \textstyle\sum_{s\le\ell-1}\alpha_s \mathbf{Sum}(G^\star_s(x))\right)\right] \approx \alpha_\ell \enspace .$$

In other words,

α_ℓ is (approximately) the increment in classification accuracy when we use an ℓ-layer network comparing to $\ell-1$-layer ones

Therefore, the gap assumption is equivalent to saying that harder data (which requires deeper networks to learn correctly) are fewer in the training set, which can be *very natural.* For instance, around 70% images of the CIFAR-10 data can be classified correctly by merely looking at their

[11]To be more precise, one can derive with probability at least α_ℓ (up to a small factor $d^{o(1)}$) it satisfies

$$x_0 + \textstyle\sum_{s\le\ell-1}\alpha_s\mathbf{Sum}(G^\star_s(x)) \in (-\frac{\alpha_\ell}{d^{o(1)}}, 0) \quad \text{and} \quad |\mathbf{Sum}(G^\star_\ell(x))| \ge \frac{1}{d^{o(1)}} \tag{3.1}$$

Indeed, there is probability at least 0.99 over x so that $\sum_{s\le\ell-1}\alpha_s\mathbf{Sum}(G^\star_s(x)) \le O(1)$, and at least 0.99 over x so that $\mathbf{Sum}(G^\star_\ell(x)) > \frac{1}{d^{o(1)}}$ (using the well-conditioned properties from Section 5 with $\kappa \le 2^{C_1^L}$ and $L = o(\log\log d)$). Then, using the property that x_0 is random Gaussian with variance 1 finishes the proof of (3.1). As a result, for at least $\alpha_\ell/d^{o(1)}$ fraction of the data, the label function is affected by the ℓ-th layer. One can do a similar argument to show that for at least $1-\alpha_\ell/d^{o(1)}$ fraction of the data, the label function is not affected by the ℓ-th layer and beyond.

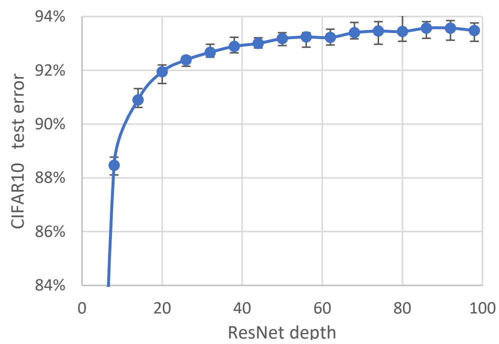

Figure 3: Performance of ResNet on CIFAR-10 dataset with various depths. One can confirm that deeper layers have diminishing contributions to the classification error. Experiment details in Appendix J.

rough colors and patterns using a one-hidden-layer network. Only the final $< 1\%$ accuracy gain requires much refined arguments such as whether there is a beak on the animal face which can only be detected using very deep networks. As another example, humans use much more training examples to learn counting, than to learn basic calculus, than to learn advanced calculus.

We refer the readers to Figure 3 which shows that indeed the increment in accuracy as we go deeper in neural networks is diminishing.

In this classification regime, our Theorem 3.1 still applies as follows. Recall the cross entropy (i.e., logistic loss) function $\mathsf{CE}(y,z) = -\log \frac{1}{1+e^{-yz}}$ where $y \in \{-1,1\}$ is the label and $z \in \mathbb{R}$ is the prediction. In this regime, we can choose a training loss function

$$\begin{aligned}\widetilde{\mathbf{Loss}}^{\mathsf{xE}}(x_0,x;\mathbf{W},\mathbf{K}) &\stackrel{\text{def}}{=} \mathsf{CE}(Y(x_0,x), v(x_0+\widetilde{F}(x;\mathbf{W},\mathbf{K}))) \\ &= \log\left(1+e^{-Y(x_0,x)\cdot v(x_0+\widetilde{F}(x;\mathbf{W},\mathbf{K}))}\right)\end{aligned}$$

where the parameter $v \geq 1$ is around $\frac{1}{\varepsilon}$ is for proper normalization and the training objective is

$$\widetilde{\mathbf{Obj}}^{\mathsf{xE}}(x_0,x;\mathbf{W},\mathbf{K}) = \widetilde{\mathbf{Loss}}^{\mathsf{xE}}(x_0,x;\mathbf{W},\mathbf{K}) + v\mathbf{Reg}(\mathbf{W},\mathbf{K}) \tag{3.2}$$

We have the following corollary of Theorem 3.1:

Theorem 3.3. *In the same setting Theorem 3.1, and suppose additionally* $\varepsilon > \frac{1}{d^{100\log d}}$. *Given* $N = \mathsf{poly}(d/\varepsilon)$ *i.i.d. samples* $\mathcal{Z}$ *from* $\mathcal{D}$ *and given their corresponding labels* $\{Y(x_0,x)\}_{(x_0,x)\in\mathcal{Z}}$, *by applying SGD over the training objective* $\widetilde{\mathbf{Obj}}^{\mathsf{xE}}(\mathcal{Z};\mathbf{W},\mathbf{K})$, *with probability at least 0.99, we can find a learner network* F *in time* $\mathsf{poly}(d/\varepsilon)$ *such that:*

$$\mathop{\mathbf{Pr}}_{(x_0,x)\sim\mathcal{D}}[Y(x_0,x) \neq \mathsf{sgn}(x_0+F(x))] \leq \varepsilon \enspace .$$

Intuitively, Theorem 3.3 is possible because under the choice of $v = 1/\varepsilon$, up to small multiplicative factors, " ℓ_2-loss equals ε^2 " becomes near identical to "cross-entropy loss equals ε". This is also why we need to add a factor v in from of the regularizers in (3.2). We make this more rigorous in Appendix G (see Proposition G.1).

4 Training algorithm

We describe our algorithm in Algorithm 1.[12] It is almost the vanilla SGD algorithm: in each iteration, it gets a random sample $z \sim \mathcal{D}$, computes (stochastic) gradient in $(\mathbf{W}, \mathbf{K})$, and moves in the negative gradient direction with step length $\eta > 0$.

Besides standard operations such as setting learning rates and regularizer weights, our only difference from SGD is to invoke (at most L times) the k-SVD decomposition algorithm to obtain a warm-start for each matrix $\mathbf{K}_\ell$ when it first becomes available. This warm-up is mainly for theoretical purpose to avoid singularities in $\mathbf{K}_\ell$ and it serves little role in actually learning $G^\star$. Essentially all of the learning is done by SGD.[13]

We emphasize once again that, when layer ℓ begins to train (by setting step length $\eta_\ell \leftarrow \eta$ to be nonzero), Algorithm 1 continues to train *all* layers $\ell' \leq \ell$. This helps to "correct" the error in layer ℓ' (recall "backward feature correction" and Figure 1). The algorithm does not work if one just trains layer ℓ and ignores others.

We specify the choices of thresholds $\mathsf{Thres}_{\ell,\triangle}$ and $\mathsf{Thres}_{\ell,\triangledown}$, and the choices of regularizer weights $\lambda_{3,\ell}, \lambda_{4,\ell}, \lambda_{5,\ell}$ in full in Section A. Below, we calculate their values in the special case Theorem 3.1.

$$\mathsf{Thres}_{\ell,\triangle} = \frac{\alpha_{\ell-1}^2}{d^{\frac{1}{3C^{\ell-1}}}} \;,\quad \mathsf{Thres}_{\ell,\triangledown} = \frac{\alpha_{\ell}^2}{d^{\frac{1}{6C^{\ell}}}} \;,\quad \lambda_{3,\ell} \leftarrow \frac{\alpha_{\ell}^2}{d^{\frac{1}{6C^{\ell}}}} \;,\quad \lambda_{4,\ell} \leftarrow \frac{\alpha_{\ell}^2}{d^{\frac{1}{3C^{\ell}}}} \;,\quad \lambda_{5,\ell} = \frac{\alpha_{\ell}^2}{d^{\frac{1}{2C^{\ell}}}} \tag{4.1}$$

As for the network width m, sample size N, and SGD learning rate η, in the special case Theorem 3.1 one can set $N = \mathsf{poly}(d/\varepsilon)$, $m = \mathsf{poly}(d/\varepsilon)$ and $\eta = \frac{1}{\mathsf{poly}(d/\varepsilon)}$.

5 Assumptions on Target Network and Distribution

Target Network. We assume the target network satisfies the following properties

1. (monotone) $d \geq k := k_2 \geq k_3 \geq \cdots \geq k_L$.
2. (normalized) $\mathbb{E}_{x\sim\mathcal{D}}\left[\mathbf{Sum}(G^\star_\ell(x))\right] \leq B_\ell$ for some $B_\ell \geq 1$ for all ℓ and $B = \max_\ell\{B_\ell\}$.
3. (well-conditioned) the singular values of $\mathbf{W}^\star_{\ell,j}$ are between $\frac{1}{\kappa}$ and κ for all ℓ, j pairs.

Properties $1, 3$ are standard and satisfied for many practical networks (in fact, many practical networks have weight matrices close to unitary, see e.g. [34]).

For property 2, although there exists worst case matrices $\mathbf{W}^\star_{\ell,j}$ with $B_\ell = \Theta(k^{2^L})$, we would like to point out when each $\mathbf{W}^\star_{\ell,j}$ is of the form $\mathbf{U}_{\ell,j}\mathbf{\Sigma}\mathbf{V}_{\ell,j}$ where $\mathbf{U}_{\ell,j}, \mathbf{V}_{\ell,j}$ are random row/column orthonormal matrices, then with probability at least 0.9999, it holds that $B_\ell = \kappa^{2^{O(\ell)}} k_\ell$ as long as $\mathbb{E}\left[\left(\frac{1}{d}\|x\|_2^2\right)^{2^L}\right] = 2^{2^{O(L)}}$.[14] Another view is that practical networks are all equipped with batch-normalization, which ensures that $B_\ell = O(k_\ell)$.

[12] We assume without loss of generality that the algorithm knows $k_\ell, \alpha_\ell, \mathcal{J}_\ell$. In fact, the algorithm requires knowing k_ℓ, which can be searched in time $\prod_\ell d^{1/C^\ell} = \mathsf{poly}(d)$ by trying all possibilities. Moreover, it seems to "require" knowing α_ℓ, but actually it suffices to know α_ℓ up to a constant factor α'_ℓ, since one can (positively) scale the weight matrices in $G^\star$ so that $G^\star$ actually uses precisely α'_ℓ. This will increase κ, B_ℓ by at most $2^{2^{O(\ell)}}$ at layer ℓ which does not change our result. The search process for all possible α'_ℓ can be done in time $O(\log(1/\varepsilon))^L = \mathsf{poly}(d)$ as long as $\varepsilon \geq 1/d^{\mathsf{poly}(\log d)}$. Moreover, knowing the set $\mathcal{J}_\ell$ also takes time $2^{O(L)} = \mathsf{poly}(d)$.

[13] For instance, after $\mathbf{K}_\ell$ is introduced and warmed up by SVD, the objective value is still around α_ℓ^2 (because deeper layers are not trained yet!) Therefore, it still requires SGD to update each $\mathbf{K}_\ell$ in order to eventualy decrease the objective to ε^2. The SVD warmup for $\mathbf{K}_\ell$ is to avoid the scenario for the minimum singular value of $\mathbf{K}_\ell$ being close to zero, which will significantly complicate analysis.

[14] We can see $\mathbb{E}_{\mathbf{W}^\star}\mathbb{E}_x[\mathbf{Sum}(G^\star_\ell(x))] = \mathbb{E}_x\mathbb{E}_{\mathbf{W}^\star}[\mathbf{Sum}(G^\star_\ell(x))]$ and then we can consider a fixed x and use the randomness of $\mathbf{W}^\star$ to prove the claim.

Algorithm 1 SGD for DenseNet

Input: Data set $\mathcal{Z}$ of size $N = |\mathcal{Z}|$, network size m, learning rate $\eta > 0$, target error ε.

current target error $\varepsilon_0 \leftarrow B^2$; $\quad \eta_\ell \leftarrow 0$; $\quad \lambda_{3,\ell}, \lambda_{4,\ell}, \lambda_{5,\ell}, \lambda_{6,\ell} \leftarrow 0$; $\quad [\mathbf{R}_\ell]_{i,j} \leftarrow \mathcal{N}(0, 1/(k_\ell)^2)$; $\mathbf{K}_\ell, \mathbf{W}_\ell \leftarrow \mathbf{0}$ for every $\ell = 2, 3, \dots, L$.
while $\varepsilon_0 \geq \varepsilon$ **do**
  while $\widetilde{\mathbf{Obj}} \geq \frac{1}{4}(\varepsilon_0)^2$ **do**
    for $\ell = 2, 3, \cdots, L$ **do**
      if $\eta_\ell = 0$ and $\widetilde{\mathbf{Obj}} \leq \mathsf{Thres}_{\ell,\triangle}$ **then**
        $\eta_\ell \leftarrow \eta$, $\lambda_{6,\ell} = \frac{(\varepsilon_0)^2}{(\overline{k}_\ell \cdot L \cdot \kappa)^8}$. $\quad \diamond$ $\overline{k}_\ell = \max\{k_j : j \in \mathcal{J}_\ell \wedge j \geq 2\}$
      end if
      if $\lambda_{3,\ell} = 0$ and $\widetilde{\mathbf{Obj}} \leq \mathsf{Thres}_{\ell,\triangledown}$ **then**
        set $\lambda_{3,\ell}, \lambda_{4,\ell}, \lambda_{5,\ell}$ according to (4.1)
        $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V} = k_\ell\text{-SVD}(\mathbf{W}_{\ell,\triangleleft}^\top \mathbf{W}_{\ell,\ell-1})$, $\mathbf{K}_{\ell,\triangleleft}^\top \leftarrow \mathbf{U}\mathbf{\Sigma}^{1/2}$, $\mathbf{K}_{\ell,\ell-1} \leftarrow \mathbf{\Sigma}^{1/2}\mathbf{V}$.
      end if
      $\mathbf{K}_\ell \leftarrow \mathbf{K}_\ell - \eta_\ell \nabla_{\mathbf{K}_\ell} \widetilde{\mathbf{Obj}}(x; \mathbf{W}, \mathbf{K})$. $\quad \diamond$ *for a random sample $x \sim \mathcal{Z}$*
      $\mathbf{W}_\ell \leftarrow \mathbf{W}_\ell - \eta_\ell \nabla_{\mathbf{W}_\ell} \widetilde{\mathbf{Obj}}(x; \mathbf{W}, \mathbf{K}) + \mathsf{noise}$. $\quad \diamond$ noise *is any polynomially-small Gaussian noise;*
    end for $\quad \diamond$ noise *is for theory purpose to escape saddle points [21].*
  end while
  $\varepsilon_0 \leftarrow \varepsilon_0/2$ and $\lambda_{6,\ell} \leftarrow \lambda_{6,\ell}/4$ for every $\ell = 2, 3, \dots, L$.
end while
return $\mathbf{W}$ and $\mathbf{K}$, representing $F(x; \mathbf{W}, \mathbf{K})$.

Input Distribution. We assume the input distribution $\mathcal{D}$ has the following property:

1. (isotropic). There is an absolute constant $c_6 > 0$ such that for every w, we have that

$$\mathbb{E}_{x \sim \mathcal{D}}[|\langle w, x \rangle|^2] \leq c_6 \|w\|_2^2 \quad \text{and} \quad \mathbb{E}_{x \sim \mathcal{D}}[|\langle w, S_1(x) \rangle|^2] \leq c_6 \|w\|_2^2 \tag{5.1}$$

2. (degree-preserving). For every positive integer q, there exists positive value $c_1(q)$ such that for every polynomial P over x with maximum degree q, let P_q be the polynomial consists of all the degree exactly q monomials of P, then the following holds

$$\mathcal{C}_x(P_q) \leq c_1(q) \mathbb{E}_{x \sim \mathcal{D}} P(x)^2 \tag{5.2}$$

   For $\mathcal{D} = N(0, \mathbf{I})$, such inequality holds with $c_1(q) \leq q!$ (can be easily proved using Hermite polynomial expansion). [15]

3. (hyper-contractivity). There exists absolute constant $c_2 > 0$ such that, for some value $c_4(q) \geq q$, we have: for every degree q polynomial $f(x)$.

$$\mathbf{Pr}_x \left[|f(x) - \mathbb{E}[f(x)]| \geq \lambda\right] \leq c_4(q) \cdot e^{-\left(\frac{\lambda^2}{c_2 \cdot \mathbf{Var}[f(x)]}\right)^{1/c_4(q)}} \tag{5.3}$$

   If $\mathcal{D} = \mathcal{N}(0, \mathbf{I})$, we have $c_4(q) \leq O(q)$ (see Lemma I.2b). This implies that for some value $c_3(q) > 0$, we also have, for every degree q polynomial $f(x)$, for every integer $p \leq 6$,

$$\mathbb{E}_{x \sim \mathcal{D}}\left[(f(x))^{2p}\right] \leq c_3(q) \mathbb{E}\left[(f(x))^2\right]^p \tag{5.4}$$

[15] We can also replace this degree-preserving assumption by directly assuming that the minimal singular value of $\mathbb{E}_{x \sim \mathcal{D}}[(\widehat{S}_{\ell'}^\star * \widehat{S}_{\ell'}^\star) \otimes (\widehat{S}_\ell^\star * \widehat{S}_\ell^\star)]$ defined in Lemma D.1 is large for $\ell' \neq \ell$ (and the corresponding "symmetric version" is large for $\ell' = \ell$), as well as $\mathbb{E}_{x \sim \mathcal{D}}[\|\widehat{S}_\ell^\star\|_2^2] \leq B$ for every $\ell \geq 2, \ell' \geq 0$.

If $\mathcal{D} = \mathcal{N}(0, \mathbf{I})$, we have $c_3(q) \leq O((6q)!)$; and more generally we have $c_3(q) \leq O(c_4(q))^{c_4(q)}$.

Assumptions 1 and 3 are very common assumptions for distributions, and they are satisfied for sub-gaussian distributions or even heavy-tailed distributions such as $p(x) \propto e^{-x^{0.1}}$. Assumption 2 says that the data has certain variance along every "high-degree directions", which is also typical for distributions such like Gaussian or heavy-tailed distributions.

We would like to point out that it is possible to have a distribution satisfying assumption 2 to be a mixture of C-distributions, where non of the individual distribution satisfies Assumption 2. For example, the distribution can be a mixture of d-distributions, the i-th distribution satisfies that $x_i = 0$ and other coordinates are i.i.d. standard Gaussian. Thus, non of the individual distribution is degree-preserving, however, the mixture of them is as long as $q \leq d - 1$.

It is easy to check that simple distributions satisfying the following parameters.

Proposition 5.1. *Our distributional assumption is satisfied for $c_6 = O(1), c_1(q) = O(q)^q, c_4(q) = O(q), c_3(q) = q^{O(q)}$ when $\mathcal{D} = \mathcal{N}(0, \mathbf{\Sigma}^2)$, where $\mathbf{\Sigma}$ has constant singular values, it is also satisfied for a mixture of arbitrarily many $\mathcal{D}_i = \mathcal{N}(0, \mathbf{\Sigma}_i^2)$'s as long as each $\mathbf{\Sigma}_i$ has constant singular values and for each j, the j-th row: $\|[\mathbf{\Sigma}_i]_j\|_2$ has the same norm for every i.*

In the special case of the main theorem stated in Theorem 3.1, we work with the above parameters. In our full Theorem A.1, we shall make the dependency of those parameters transparent.

6 Proof Intuitions

In this high-level intuition let us first ignore the difference between truncated activations and the true quadratic activation. We shall explain at the end why we need to do truncation.

6.1 A Though Experiment

We provide intuitions about the proof by first considering the following extremely simplified example: $L = 3$, $d = 4$, and $G^\star(x) = x_1^4 + x_2^4 + \alpha((x_1^4 + x_3)^2 + (x_2^4 + x_4)^2)$ for some $\alpha = o(1)$. In our language, due to notational convenience, $L = 3$ refers to having only two trainable layers, that we refer to as the second and third layers.

Richer representation by over-parameterization. Since $\alpha < 1$, one would hope for the second layer of the network to learn x_1^4 and x_2^4 (by some representation of its neurons), and feed this as an input to the third layer. If so, the third layer could learn a quadratic function over x_1^4, x_2^4, x_3, x_4 to fit the remainder $\alpha((x_1^4 + x_3)^2 + (x_2^4 + x_4)^2)$ in the objective. This logic has a critical flaw:

- *Instead of learning x_1^4, x_2^4, the second layer may as well learn $\frac{1}{5}(x_1^2 + 2x_2^2)^2, \frac{1}{5}(2x_1^2 - x_2^2)^2$.*

Indeed, it is easy to verify that $\frac{1}{5}(x_1^2 + 2x_2^2)^2 + \frac{1}{5}(2x_1^2 - x_2^2)^2 = x_1^4 + x_2^4$. However, *no* quadratic function over $\frac{1}{5}(x_1^2 + 2x_2^2)^2, \frac{1}{5}(2x_1^2 - x_2^2)^2$ and x_3, x_4 can reproduce $(x_1^4 + x_3)^2 + (x_2^4 + x_4)^2$. Therefore, the second layer not only needs to learn a function to fit $x_1^4 + x_2^4$, but also has to learn the "correct basis" x_1^4, x_2^4 for the next layer.

To achieve this goal, we let the learner network to use (quadratically-sized) over-parameterization with random initialization. Instead of having only two hidden neurons, we will let the network have $m_2 > 2$ hidden neurons. We then show a critical lemma that the neurons in the second layer of the network can learn *a richer representation* of the same function $x_1^4 + x_2^4$, given by:

$$\{(\alpha_i x_1^2 + \beta_i x_2^2)^2\}_{i=1}^{m_2}$$

In each hidden neuron, the coefficients α_i, β_i behave like i.i.d. gaussian random variables. Indeed, $\mathbb{E}[(\alpha_i x_1^2 + \beta_i x_2^2)^2] \approx x_1^4 + x_2^4$, and w.h.p. when $m_2 \geq 3$, we can show that a quadratic function of $\{(\alpha_i x_1^2 + \beta_i x_2^2)^2\}_{i=1}^{m_2}, x_3, x_4$ can be used to fit $(x_1^4 + x_3)^2 + (x_2^4 + x_4)^2$. The algorithm can proceed. Here we present a completely different view of such over-parameterization:

The role of this over-parameterization is not to make the training easier in the current layer, instead, it enforces the network to learn a richer set of features (to represent the same target function) that can be better used for the next layer.

Improvement in lower layers after learning higher layers. The second obstacle in this thought experiment is that the second layer might not learn the function $x_1^4 + x_2^4$ *exactly*. Indeed, it is possible to come up with a distribution where the best quadratic function of x_i^2 to fit $G^\star(x)$ is $(x_1^2 + \alpha x_3^2)^2 + (x_2^2 + \alpha x_4^2)^2$. This is *over-fitting*, and the error $\alpha x_3, \alpha x_4$ *cannot* be corrected by over-parameterization, it can only be corrected via learning higher layers (i.e., backward feature correction).

Now, since the second layer feeds $(x_1^2 + \alpha x_3^2)^2, (x_2^2 + \alpha x_4^2)^2$ to the third layer, the third layer might start to learn $\alpha((x_1^2+\alpha x_3^2)^2+x_3)^2+\alpha((x_2^2+\alpha x_4^2)^2+x_4)^2$ to fit the remaining terms in $G^\star(x)$. A very neat observation is that $\alpha((x_1^2+\alpha x_3^2)^2+x_3)^2+\alpha((x_2^2+\alpha x_4^2)^2+x_4)^2$ is actually α^2 close to $\alpha((x_1^4+x_3)^2+(x_2^4+x_4)^2)$. Therefore, the first layer can now be correct to e.g. $(x_1^2+\alpha^2 x_3^2)^2+(x_2^2+\alpha^2 x_4^2)^2$ due to the existence of the third layer to over-fitting to the higher order term reduced from $\alpha((x_1^4+x_3)^2+(x_2^4+x_4)^2) \sim \alpha$ to $\alpha((x_1^4+x_3)^2+(x_2^4+x_4)^2)-\alpha((x_1^2+\alpha x_3^2)^2+x_3)^2-\alpha((x_2^2+\alpha x_4^2)^2+x_4)^2 \sim \alpha^2$. We call this "backward feature correction" (see Figure 1).

This process can keep going and the network can thus gradually improve the function in the second layer via the reduction of over-fitting from the third layer to eventually learn $G^\star$ up to arbitrarily small error $\varepsilon > 0$. We summarize the *hierarchical* learning process in Figure 2 when more than 2 layers are present, which is slightly more involved.

6.2 Details on Implementing the Intuition

Following this intuition, we would like to show that when the loss is ε^2, then for each $\ell \in [L]$, the function S_ℓ learnt by the network at the ℓ-th layer is correct up to error ε/α_ℓ, thus, when we use the output of the ℓ-th layer to learn the output $\alpha_{\ell+1} G^\star_{\ell+1}$ on the $\ell+1$-th layer, in principle, we can learn it up to accuracy $\varepsilon/\alpha_\ell \times \alpha_{\ell+1} < \varepsilon$, thus, the function output of the higher level layers can be learnt up to a *smaller* error than ε, which will in turn reduce the error in the lower layers.

There are several major obstacles for implementing the above intuition, as we summarized blow.

Function value v.s. coefficients. To actually implement the approach, we first notice that $F_{\ell+1}$ is a polynomial of *maximum* degree $2^{\ell+1}$, however, it also has a lot of lower-degree monomials. Obviously, the monomials of degree less than $2^\ell+1$ can also be learnt at layers ℓ through F_ℓ, which means that it is *impossible* to show that $F_{\ell+1} \approx G^\star_{\ell+1}$ simply from $F \approx G^\star$. As an example, in principle, the learner network could instead learn $F_{\ell+1}(x) \approx G^\star_{\ell+1}(x) - F'(x)$ for some error function $F'(x)$ of degree 2^ℓ, while satisfying $F_\ell(x) \approx G^\star_\ell(x) + \frac{\alpha_{\ell+1}}{\alpha_\ell} F'(x)$.

Our critical lemma (see Lemma E.1) proves that this *cannot* happen when we train the network using SGD. We prove it by first focusing on all the monomials in $F_{\ell+1}$ of degree $2^\ell+1, \dots, 2^{\ell+1}$, which are not learnable at lower-level layers. One might hope to use this observation to show that it must be the case $\widehat{F}_{\ell+1}(x) \approx \widehat{G^\star}_{\ell+1}(x)$, where the $\widehat{F}_{\ell+1}$ contains all the monomials in $F_{\ell+1}$ of degree $2^\ell+1, \dots, 2^{\ell+1}$ and similarly for $\widehat{G^\star}_{\ell+1}$.

Unfortunately, this approach fails again. Even in the ideal case when we already have $F_{\ell+1} \approx G^\star_{\ell+1} \pm \varepsilon'$ where $\varepsilon' = \varepsilon/\alpha_\ell$, it still *does not* imply that $\widehat{F}_{\ell+1} \approx \widehat{G^\star}_{\ell+1} \pm \varepsilon'$. One counter example is the polynomial $\sum_{i\in[d]} \frac{\varepsilon'}{\sqrt{d}}(x_i^2 - 1)$ where $x_i \sim \mathcal{N}(0,1)$. This polynomial is ε'-close to zero, however, its degree-2 terms $\frac{\varepsilon'}{\sqrt{d}} x_i^2$ when added up is actually $\sqrt{d}\varepsilon'$, loosing a dimension factor. In worst case, such difference could end up leading to a complexity of $d^{\Omega(2^L)}$ large for degree 2^L polynomials, leading to an unsatisfying bound.

To correct this, we count the monomial *coefficients* instead of the actual function value. The main observation is that the top-degree (i.e., degree-$2^{\ell+1}$) coefficients of the monomials in $F_{\ell+1}$ is in fact ε' close to that of $G^\star_{\ell+1}$ in terms of ℓ_2-norm, without sacrificing a dimension factor (and only sacrificing a factor that depends on the degree). Taking the above example, the ℓ_2 norm of the coefficients of $\frac{\varepsilon'}{\sqrt{d}} x_i^2$ is indeed ε', which does not grow with the dimension d. The closeness in coefficient is used in (E.7) in our proof of Lemma E.1.

Symmetrization. Following the aforementioned step, one would like to show that when the network learns a function $F_{\ell+1}$ whose coefficients of the degree $2^{\ell+1}$ monomials matches that of $G^\star_{\ell+1}$, it must imply that $\mathbf{W}_{\ell+1,\ell}$ is close to $\mathbf{W}^\star_{\ell+1,\ell}$ in some measure. Indeed, all of the top-degree (i.e., degree $2^{\ell+1}$) monomials in $F_{\ell+1}$ must come from $\sigma(\mathbf{W}_{\ell+1,\ell}\sigma(\mathbf{R}_\ell \widehat{S}_\ell))$, where $\widehat{S}_\ell$ consists of all the top-degree (i.e., degree-$2^{\ell-1}$) monomials in S_ℓ. Now, using the assumption that S_ℓ is ε' close to $S^\star_\ell$, we be able to show the coefficients of $\widehat{S}_\ell$ is ε'-close to that of $\widehat{S}^\star_\ell$ (in coefficients). Now, we arrive at the following question:

If the coefficients of $\widehat{S}_\ell(x)$, in ℓ_2-norm, are ε'-close to that of $\widehat{S}^\star_\ell(x)$, and the coefficients of $\sigma(\mathbf{W}_{\ell+1,\ell}\sigma(\mathbf{R}_\ell \widehat{S}_\ell))$, in ℓ_2-norm, are ε'-close to that of $\sigma(\mathbf{W}^\star_{\ell+1,\ell}\sigma(\widehat{S}^\star_\ell))$, then, does it mean that $\mathbf{W}^\star_{\ell+1,\ell}$ is ε'-close to $\mathbf{W}^\star_{\ell+1,\ell}$ in some measure?

The answer to this question is intricate, due to the huge amount of "symmetricity" in a high-degree polynomial. For example, when $x \in \mathbb{R}^d$ and $\mathbf{M}, \mathbf{M}' \in \mathbb{R}^{d^2 \times d^2}$, suppose $(x \otimes x)^\top \mathbf{M}(x \otimes x)$ is ε'-close to $(x \otimes x)^\top \mathbf{M}'(x \otimes x)$ in terms of coefficients when we view them as degree 4 polynomials, this *does not* imply that $\mathbf{M}$ is close to $\mathbf{M}'$ at all. Indeed, if we increase $\mathbf{M}_{(1,2),(3,4)}$ by 10^{10} and decrease $\mathbf{M}_{(1,3),(2,4)}$ by 10^{10}, then $(x \otimes x)^\top \mathbf{M}(x \otimes x)$ remains the same.

To solve this issue, we first consider a symmetric version of tensor product, the $*$ product defined in Definition B.2, which makes sure that $x * x$ only has $\binom{d+1}{2}$ dimensions, each corresponding to $\{i,j\}$-th entry for $i \le j$. This makes sure that the $\mathbf{M}_{\{1,2\},\{3,4\}}$ is the same entry as $\mathbf{M}_{\{2,1\},\{4,3\}}$. However, this simple fix does not resolve all the "symmetricity" issues. Indeed, $\mathbf{M}_{\{1,2\},\{3,4\}}$ and $\mathbf{M}_{\{1,3\},\{2,4\}}$ are still difference entries of $\mathbf{M}$.

Therefore, for fundamental reasons, we *cannot* derive that $\mathbf{W}_{\ell+1,\ell}$ and $\mathbf{W}^\star_{\ell+1,\ell}$ are ε'-close. However, they should still be close after somehow "twice symmetrizing" their entries. For this to hold, we introduce a "twice symmetrization" operator $\mathbf{Sym}$ on matrices defined in Definition B.3, and eventually derive that $\mathbf{W}_{\ell+1,\ell}$ and $\mathbf{W}^\star_{\ell,\ell}$ are close under the following notation:

$$\mathbf{Sym}\left((\mathbf{R}_\ell * \mathbf{R}_\ell)(\mathbf{W}_{\ell+1,\ell})^\top \mathbf{W}_{\ell+1,\ell}(\mathbf{R}_\ell * \mathbf{R}_\ell)^\top\right) \approx \mathbf{Sym}\left((\mathbf{I} * \mathbf{I})^\top (\mathbf{W}_{\ell+1,\ell})^\top \mathbf{W}^\star_{\ell+1,\ell}(\mathbf{I} * \mathbf{I})\right) \pm \varepsilon' \tag{6.1}$$

Although $\mathbf{W}_{\ell+1,\ell}$ and $\mathbf{W}^\star_{\ell,\ell}$ themselves are not close, fortunately, we can still use this to non-trivially derive that $\sigma(\mathbf{W}_{\ell+1,\ell}\sigma(\mathbf{R}_\ell S_\ell))$ is close to $\sigma(\mathbf{W}^\star_{\ell+1,\ell}\sigma(S^\star_\ell))$, since S_ℓ is close to $S^\star_\ell$ as we have assumed.

Once we have that, we can move to the second-highest degree monomials of $F_{\ell+1}(x)$. Without loss of generality, we assume it is of degree $2^\ell + 2^{\ell-2}$. (Note it cannot be $2^\ell + 2^{\ell-1}$ since we assumed skip links.) Such degree monomials must either come from $\sigma(\mathbf{W}^\star_{\ell+1,\ell}\sigma(S^\star_\ell))$— which we have just shown that it is close to $\sigma(\mathbf{W}_{\ell+1,\ell}\sigma(\mathbf{R}_\ell S_\ell))$— come from the cross term

$$(S^\star_\ell * S^\star_\ell)\left(\mathbf{W}^\star_{\ell+1,\ell}\right)^\top \mathbf{W}^\star_{\ell+1,\ell-2}(S^\star_{\ell-2} * S^\star_{\ell-2})$$

Thus, we can proceed by showing that the coefficients of the above cross term is close to its counterpart in $\mathbf{W}_{\ell+1,\ell}^\top \mathbf{W}_{\ell+1,\ell-2}$.

Fortunately, this time the matrix $(\mathbf{W}^\star_{\ell+1,\ell})^\top \mathbf{W}^\star_{\ell+1,\ell-2}$ is not symmetric, and therefore we do not have the "twice symmetrization" issue as argued above. Therefore, we can directly conclude that the non-symmetrized closeness, or in symbols,

$$(\mathbf{R}_{\ell-2} * \mathbf{R}_{\ell-2})\,(\mathbf{W}_{\ell+1,\ell-2})^\top \mathbf{W}_{\ell+1,\ell}\,(\mathbf{R}_\ell * \mathbf{R}_\ell)^\top \approx (\mathbf{I} * \mathbf{I})^\top\,(\mathbf{W}^\star_{\ell+1,\ell-2})^\top \mathbf{W}^\star_{\ell+1,\ell}\,(\mathbf{I} * \mathbf{I}) \tag{6.2}$$

We can continue in this fashion for all the remaining degrees until degree $2^\ell + 1$.

Moving from W to K: Part I. We have so far shown that $\mathbf{W}_{\ell+1,j}$ and $\mathbf{W}^\star_{\ell+1,j}$ are close in some measure. We hope to use this to show that the function $S_{\ell+1}$ is close to $S^\star_{\ell+1}$ and proceed the induction. However, if we use the matrix $\mathbf{W}_{\ell+1}$ to define $S_{\ell+1}$ (instead of introducing the notation $\mathbf{K}_{\ell+1}$), then $S_{\ell+1}$ might have a huge error compare to $S^\star_{\ell+1}$.

Indeed, even if in the ideal case that $(\mathbf{W}_{\ell+1,\ell})^\top \mathbf{W}_{\ell+1,\ell} \approx (\mathbf{W}^\star_{\ell+1,\ell})^\top \mathbf{W}^\star_{\ell+1,\ell} + \varepsilon'$, this only guarantees that $\mathbf{W}_{\ell+1,\ell} \approx \mathbf{U}\mathbf{W}^\star_{\ell+1,\ell} + \sqrt{\varepsilon'}$ for some column orthonormal matrix $\mathbf{U}$. This is because the inner dimension m of $\mathbf{W}_{\ell+1,\ell}$ is much larger than that the inner dimension $k_{\ell+1}$ of $\mathbf{W}^\star_{\ell+1,\ell}$.[16] This $\sqrt{\varepsilon'}$ error can lie in the orthogonal complement of $\mathbf{U}$.

To fix this issue, we need to "reduce" the dimension of $\mathbf{W}_{\ell+1,\ell}$ back to $k_{\ell+1}$, which also reduces the error. This is why we need to introduce the $\mathbf{K}_{\ell+1,\ell}$ matrix of rank $k_{\ell+1}$, and add a regularizer to ensure that $\mathbf{K}_{\ell+1,\ell}^\top \mathbf{K}_{\ell+1,\ell}$ approximates $(\mathbf{W}_{\ell+1,\ell})^\top \mathbf{W}_{\ell+1,\ell}$. (This can be reminiscent of knowledge distillation used in practice [31].) This "SVD" type of update further decreases the error back to ε', so now $\mathbf{K}_{\ell+1,\ell}$ becomes ε' close to $\mathbf{W}^\star_{\ell+1,\ell}$ up to column orthonormal transformation.[17] We use this to proceed and conclude the closeness of $S_{\ell+1}$. This is done in Section E.6.

Moving from W to K: Part II. Now suppose the leading term (6.1) holds without the **Sym** operator (see Footnote 17 for how to get rid of it), and suppose the cross term (6.2) also holds. The former means "$(\mathbf{W}_{\ell+1,\ell})^\top \mathbf{W}_{\ell+1,\ell}$ is close to $(\mathbf{W}^\star_{\ell+1,\ell})^\top \mathbf{W}^\star_{\ell+1,\ell}$ (up to transformations)" and the latter means "$(\mathbf{W}_{\ell+1,\ell-2})^\top \mathbf{W}_{\ell+1,\ell}$ is close to $(\mathbf{W}^\star_{\ell+1,\ell-2})^\top \mathbf{W}^\star_{\ell+1,\ell}$". These two together, still does not imply that "$(\mathbf{W}_{\ell+1,\ell-2})^\top \mathbf{W}_{\ell+1,\ell-2}$ is close to $(\mathbf{W}^\star_{\ell+1,\ell-2})^\top \mathbf{W}^\star_{\ell+1,\ell-2}$", since the error of $\mathbf{W}_{\ell+1,\ell-2}$ can also lie on the orthogonal complement of $\mathbf{W}_{\ell+1,\ell}$. This error can in fact be arbitrary large when $\mathbf{W}_{\ell+1,\ell}$ is not full rank.

This means, the learner network can still make a lot of error on the $\ell + 1$ layer, *even when it already learns all degree $> 2^\ell$ monomials correctly.* To resolve this, we again need to use the regularizer to ensure closeness between $\mathbf{W}_{\ell,\ell-2}$ to $\mathbf{K}_{\ell,\ell-2}$. It "reduces" the error because by enforcing $\mathbf{W}_{\ell+1,\ell-2}$ being close to $\mathbf{K}_{\ell+1,\ell}$, it must be of low rank— thus the "arbitrary large error" from the orthogonal complement cannot exist. Thus, *it is important that we keep $\mathbf{W}_\ell$ being close to the low rank counterpart $\mathbf{K}_\ell$, and update them together gradually.*

Remark 6.1. If we have "weight sharing", meaning forcing $\mathbf{W}_{\ell+1,\ell-2} = \mathbf{W}_{\ell+1,\ell}$, then we immediately have $(\mathbf{W}_{\ell+1,\ell-2})^\top \mathbf{W}_{\ell+1,\ell-2}$ is close to $(\mathbf{W}^\star_{\ell+1,\ell-2})^\top \mathbf{W}^\star_{\ell+1,\ell-2}$, so we do not need to rely on "$\mathbf{W}_{\ell+1,\ell-2}$ is close to $\mathbf{K}_{\ell+1,\ell}$" and this can make the proof much simpler.

Empirical v.s. Population loss. So far we only focus on the case when F is close to $G^\star$ in

[16]Recall that without RIP-type of strong assumptions, such over-parameterization m is somewhat necessary for a neural network with quadratic activations to perform optimization without running into saddle points, and is also used in [4].

[17]In fact, things are still trickier than one would expect. To show "$\mathbf{K}_{\ell+1,\ell}$ close to $\mathbf{W}^\star_{\ell+1,\ell}$," one needs to first have "$\mathbf{W}_{\ell+1,\ell}$ close to $\mathbf{W}^\star_{\ell+1,\ell}$", but we do not have that due to the twice symmetrization issue from (6.1). Instead, our approach is to first (6.2) to derive that there exists some matrix $\mathbf{P}$ satisfying "$\mathbf{P}\mathbf{K}_{\ell+1,\ell}$ is close to $\mathbf{P}\mathbf{W}^\star_{\ell+1,\ell}$" and "$\mathbf{P}^{-1}\mathbf{K}_{\ell+1,\ell-2}$ is close to $\mathbf{P}\mathbf{W}^\star_{\ell+1,\ell-2}$". Then, we plug this back to (6.1) to derive that $\mathbf{P}$ must be close to $\mathbf{I}$. This is precisely why we need a skip connection.

population case (i.e., under the true distribution $\mathcal{D}$), since properties such as degree preserving Property 5.2 is *only* true for the population loss. Indeed, if we only have $\mathsf{poly}(d)$ samples, the empirical distribution can not be degree-preserving at all for any $2^\ell = \omega(1)$.

One would like to get around it by showing that, when F is close to $G^\star$ only on the *training* data set $\mathcal{Z}$, then the aforementioned closeness between S_ℓ and $S^\star_\ell$ still holds for the *population* case. This turns out to be a challenging task.

One naive idea would be to show that $\mathbb{E}_{x\sim\mathcal{Z}}\left(F(x) - G^\star(x)\right)^2$ is close to $\mathbb{E}_{x\sim\mathcal{D}}\left(F(x) - G^\star(x)\right)^2$ for any networks weights $\mathbf{W}, \mathbf{K}$. However, this *cannot* work at all. Since $F(x) - G^\star(x)$ is a degree 2^L polynomial, we know that for a fixed F, $\mathbb{E}_{x\sim\mathcal{Z}}\left(F(x) - G^\star(x)\right)^2 \approx \mathbb{E}_{x\sim\mathcal{D}}\left(F(x) - G^\star(x)\right)^2 \pm \varepsilon$ only holds with probability $e^{-(N\log(1/\varepsilon))^{\frac{1}{2^L}}}$, where $|\mathcal{Z}| = N$. This implies, in order for it to hold *for all* possible $\mathbf{W}, \mathbf{K}$, we need at least $N = \Omega(d^{2^L})$ many samples, which is too bad.

We took an alternative approach. We truncated the learner network from F to $\widetilde{F}$ using truncated quadratic activations (recall 2.2): if the intermediate value of some layers becomes larger than some parameter B', then we truncate it to $\Theta(B')$. Using this operation, we can show that the function output of $\widetilde{F}$ is always bounded by a small value. Using this, one could show that $\mathbb{E}_{x\sim\mathcal{Z}}\big(\widetilde{F}(x) - G^\star(x)\big)^2 \approx \mathbb{E}_{x\sim\mathcal{D}}\big(\widetilde{F}(x) - G^\star(x)\big)^2 \pm \varepsilon$.

But, why is $F(x)$ necessarily close to $\widetilde{F}(x)$, especially on the training set $\mathcal{Z}$? If some of the $x \in \mathcal{Z}$ is too large, then $\big(\widetilde{F}(x) - F(x)\big)^2$ can be large as well. Fortunately, we show during the training process, the neural network actually has *implicit self-regularization* (as shown in Corollary E.3e): the *intermediate values* such as $\|S_\ell(x)\|^2$ stay away from $2B$ for most of the $x \sim \mathcal{D}$. This ensures that $\mathbb{E}_{x\sim\mathcal{D}}(F(x) - \widetilde{F}(x))^2$ is small in the population loss.

This implicit regularization is elegantly maintained by SGD where the weight matrix does not move too much at each step, this is another place where we need gradual training instead of one-shot learning.

Using this property we can conclude that

$$\mathbb{E}_{x\sim\mathcal{Z}}\left(\widetilde{F}(x) - G^\star(x)\right)^2 \text{ is small} \iff \mathbb{E}_{x\sim\mathcal{D}}\left(\widetilde{F}(x) - G^\star(x)\right)^2 \text{ is small} \iff \mathbb{E}_{x\sim\mathcal{D}}\left(F(x) - G^\star(x)\right)^2 \text{ is small},$$

which allows us to interchangeably apply all the aforementioned arguments both on the empirical truncated loss and on the population loss.

Appendix: Complete Proofs

We provide clear roadmap of what is included in this appendix. Note that a full statement of our theorem and its high-level proof plan begin on the next page.

- Section A : In this section, we first state the general version of the main theorem, including agnostic case in Section A.4.

- Section B : In this section, we introduce notations including defining the symmetric tensor product $*$ and the twice symmetrization operator $\mathbf{Sym}(\mathbf{M})$.

- Section C : In this section, we show useful properties of our loss function. To mention a few:

  1. In Section C.1 we show the truncated version $\widetilde{S}_\ell$ is close to S_ℓ in the population loss.
  2. In Section C.3 we show S_ℓ is Lipschitz continuous in the population loss. We need this to show that when doing a gradient update step, the quantity $\mathbb{E}_{x\sim\mathcal{D}}[\|S_\ell\|^2]$ does not move too much in population loss. This is important for the self-regularization property we discussed in Section 6 to hold.
  3. In Section C.4 we show the empirical truncated loss is Lipschitz w.r.t. $\mathbf{K}$.
  4. In Section C.5 we show the empirical truncated loss satisfies higher-order Lipschitz smoothness w.r.t. $\mathbf{K}$ and $\mathbf{W}$. We need this to derive the time complexity of SGD.
  5. In Section C.6 we show empirical truncated loss is close to the population truncated loss. We need this together with Section C.1 to deriv the final generalization bound.

- Section D : In this section, we prove the critical result about the "coefficient preserving" property of $\widehat{S}^\star_\ell(x)$, as we discussed in Section 6. This is used to show that if the output of F is close to $G^\star$ in population, then the high degree coefficient must match, thus $\mathbf{W}$ must be close to $\mathbf{W}^\star$ in some measure.

- Section E : In this section, we present our main technical lemma for hierarchical learning. It says as long as the (population) objective is as small as ε^2, then the following properties hold: loosely speaking, for every layer ℓ,

  1. (hierarchical learning): $S_\ell(x)$ close to $S^\star_\ell(x)$ by error $\sim \varepsilon/\alpha_\ell$, up to unitary transformation.
  2. (boundedness): each $\mathbb{E}[\|S_\ell(x)\|_2^2]$ is bounded. (This is needed in self-regularization.)

  We emphasize that these properties are maintained *gradually*. In the sense that we need to start with a case where these properties are already *approximately* satisfied, and then we show that the network will *self-regularize* to improve these properties. It does not mean, for example in the "hierarchical learning" property above, any network with loss smaller than ε^2 satisfies this property; we need to conclude from the fact that this network is obtained via a (small step) gradient update from an earlier network that has this property with loss $\leq 2\varepsilon$.

- Section F : In this section, we use the main technical lemma to show that there is a descent direction of the training objective, as long as the objective value is not too small. Specifically, we show that there is a gradient update direction of $\mathbf{K}$ and a second order Hessian update direction of $\mathbf{W}$, which guarantees to decrease the objective. This means, in the non-convex optimization language, there is no second-order critical points, so one can apply SGD to sufficiently decrease the objective.

- Section G : We show how to extend our theorems to classification.
- Section H : This section contains our lower bounds.

A Main Theorem and Proof Plan

Let us recall that d is the input dimension and $x \in \mathbb{R}^d$ is the input. We use L to denote the total number of layers in the network, and use k_ℓ to denote the width (number of neurons) of the hidden layer ℓ. Throughout the appendix, we make the following conventions:

- $k = \max_\ell\{k_\ell\}$ and $\overline{k}_\ell = \max\{k_j : j \in \mathcal{J}_\ell \wedge j \geq 2\}$.
- $B = \max_\ell\{B_\ell\}$ and $\overline{B}_\ell = \max\{B_j : j \in \mathcal{J}_\ell \wedge j \geq 2\}$.

Our main theorem in its full generalization can be stated as follows.

Theorem A.1. *For any desired accuracy $\varepsilon \in (0,1)$, suppose the following gap assumption is satisfied*

$$\frac{\alpha_\ell}{\alpha_{\ell+1}} \geq (c_4(2^\ell)\log(dL/\varepsilon))^{c_4(2^\ell)} \cdot (\kappa \cdot c_1(2^\ell) \cdot c_3(2^\ell))^{2^{c_0 \cdot L}} \prod_{j=\ell}^{L} \left(\overline{k}_\ell \overline{B}_\ell\right)^{L 2^{c_0(j-\ell)}}$$

Then, there exist choices of parameters (i.e., regularizer weight, learning rate, over parameterization) so that using

$$N \geq d^2 \cdot \log^{\Omega(1)} \frac{d}{\delta} + \frac{d \log d}{\varepsilon^6} \cdot \mathsf{poly}(B, k, \kappa) \cdot \left(c_4(2^L) \log \frac{BkL\kappa d}{\delta\varepsilon}\right)^{\Omega(c_4(2^L))}$$

samples. With probability at least 0.99 *over the randomness of* $\{\mathbf{R}_\ell\}_\ell$, *with probability at least* $1-\delta$ *over the randomness of* $\mathcal{Z}$, *in at most time complexity*

$$T \leq \mathsf{poly}\left(\kappa^L, \prod_\ell \overline{k}_\ell \overline{B}_\ell, (c_4(2^L))^{c_4(2^L)}, \log^{c_4(2^L)} \frac{1}{\delta}, \frac{d}{\varepsilon}\right)$$

SGD converges to a point with

$$\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) \leq \varepsilon^2 \quad \widetilde{\mathbf{Obj}}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \leq \varepsilon^2 \quad \mathbf{Obj}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \leq \varepsilon^2$$

Corollary A.2. *In the typical setting when $c_3(q) \leq q^{O(q)}$, $c_1(q) \leq O(q^q)$, and $c_4(q) \leq O(q)$, Theorem A.1 simplifies to*

$$\frac{\alpha_\ell}{\alpha_{\ell+1}} \geq \left(\log \frac{d}{\varepsilon}\right)^{c_0 \cdot 2^\ell} (\kappa)^{2^{c_0 \cdot L}} \prod_{j=\ell}^{L} \left(\overline{k}_\ell \overline{B}_\ell\right)^{L 2^{c_0(j-\ell)}}$$

$$N \geq d^2 \cdot \log^{\Omega(1)} \frac{d}{\delta} + \frac{d \log d}{\varepsilon^6} \cdot \mathsf{poly}(B, k, \kappa) \cdot \left(2^L \log \frac{Bk\kappa d}{\delta\varepsilon}\right)^{\Omega(2^L)}$$

$$T \leq \mathsf{poly}\left(\kappa^L, \prod_\ell \overline{k}_\ell \overline{B}_\ell, 2^{L 2^L}, \log^{2^L} \frac{1}{\delta}, \frac{d}{\varepsilon}\right)$$

Finally, in the special case Theorem 3.1, we have additional assumed $\delta = 0.01$, $L = o(\log\log d)$, $\kappa \leq 2^{C_1^L}$, $B_\ell \leq 2^{C_1^\ell} k_\ell$, and $k_\ell \leq d^{\frac{1}{C^{\ell+C_1}}}$. This further simplifies the notations:

Corollary A.3. *In the typical setting when $c_3(q) \le q^{O(q)}$, $c_1(q) \le O(q^q)$, and $c_4(q) \le O(q)$, and in the special case of Theorem 3.1, we have that Theorem A.1 simplifies to*

$$\frac{\alpha_{\ell+1}}{\alpha_\ell} \le d^{-\frac{1}{C^\ell}}, \quad N \ge \mathsf{poly}(d/\varepsilon), \quad \text{and} \quad T \le \mathsf{poly}(d/\varepsilon)$$

A.1 Parameter Choices

Definition A.4. *In our analysis, let us introduce a few more notations.*

- *With the following notation we can write $\mathsf{poly}(\widetilde{\kappa}_\ell)$ instead of $\mathsf{poly}(\overline{k}_\ell, L, \kappa)$ whenever needed.*
$$\widetilde{\kappa}_\ell = (\overline{k}_\ell \cdot L \cdot \kappa)^4 \text{ and } \tau_\ell = (\overline{B}_\ell \cdot \overline{k}_\ell \cdot L \cdot \kappa)^4 .$$

- *The next one is our final choice of the truncation parameter for $\widetilde{\sigma}_\ell(x)$ at each layer ℓ.*
$$B'_\ell \stackrel{\text{def}}{=} \mathsf{poly}(\tau_\ell) \cdot \Omega(c_4(2^\ell)\log(dL/\varepsilon))^{c_4(2^\ell)} \text{ and } \overline{B}'_\ell = \max\{B'_j \,:\, j \in \mathcal{J}_\ell \wedge j \ge 2\}$$

- *The following can simplify our notations.*
$$k = \max_\ell\{k_\ell\},\ B = \max_\ell\{B_\ell\},\ \widetilde{\kappa} = \max_\ell\{\widetilde{\kappa}_\ell\},\ \tau = \max_\ell\{\tau_\ell\},\ B' = \max_\ell\{B'_\ell\}$$

- *The following is our main "big polynomial factors" to carry around, and it satisfies*
$$D_\ell \stackrel{\text{def}}{=} \left(\tau_\ell \cdot \kappa^{2^\ell} \cdot (2^\ell)^{2^\ell} \cdot c_1(2^\ell) \cdot c_3(2^\ell)\right)^{c_0 \ell} \text{ and } \Upsilon_\ell = \prod_{j=\ell}^{L} (D_j)^{20 \cdot 2^{6(j-\ell)}}$$
Note it satisfies $\Upsilon_\ell \ge (D_\ell)^{20}(\Upsilon_{\ell+1}\Upsilon_{\ell+2}\cdots\Upsilon_L)^6$.

- *The following is our gap assumption.*
$$\frac{\alpha_{\ell+1}}{\alpha_\ell} \le \frac{1}{(\Upsilon_{\ell+1})^6 \overline{B}'_{\ell+1}}$$

- *Our thresholds*
$$\mathsf{Thres}_{\ell,\triangle} = \left(\frac{\alpha_{\ell-1}}{(D_{\ell-1})^9 \Upsilon_{\ell-1}}\right)^2 , \quad \mathsf{Thres}_{\ell,\triangledown} = \frac{1}{4}\left(\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}\right)^2$$

- *The following is our choice of the regularizer weights*[18]
$$\lambda_{6,\ell} = \frac{\varepsilon^2}{(\widetilde{\kappa}_\ell)^2}, \quad \lambda_{3,\ell} = \frac{\alpha_\ell^2}{D_\ell \cdot \Upsilon_\ell}, \quad \lambda_{4,\ell} = \frac{\alpha_\ell^2}{(D_\ell)^7 \Upsilon_\ell^2}, \quad \lambda_{5,\ell} = \frac{\alpha_\ell^2}{(D_\ell)^{13}\Upsilon_\ell^3}$$

- *The following is our amount of the over-parametrization*
$$m \ge \mathsf{poly}(\widetilde{\kappa}, B')/\varepsilon^2$$

- *The following is our final choice of the sample complexity*
$$N \ge d^2 \cdot \log^{\Omega(1)} \frac{d}{\delta} + \frac{md\log d}{\varepsilon^4} \cdot \mathsf{poly}(\tau)\left(2^L c_4(2^L)\log\frac{\tau d}{\delta\varepsilon}\right)^{c_4(2^L)+\Omega(1)}$$

[18] Let us make a comment on $\lambda_{6,\ell} = \frac{\varepsilon^2}{(\widetilde{\kappa}_\ell)^2}$. In Algorithm 1, we have in fact chosen $\lambda_{6,\ell} = \frac{(\varepsilon_0)^4}{(\widetilde{\kappa}_\ell)^2}$, where ε_0 is the current "target error", that is guaranteed to be within a factor of 2 comparing to the true ε (that comes from $\varepsilon^2 = \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})$). To make the notations simpler, we have ignored this constant factor 2.

A.2 Algorithm Description For Analysis Purpose

For analysis purpose, it would be nice to divide our Algorithm 1 into stages for $\ell = 2, 3, \ldots, L$.

- Stage $\ell^\triangle$ begins with $\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) \leq \mathsf{Thres}_{\ell,\triangle} = \left(\frac{\alpha_{\ell-1}}{(D_{\ell-1})^9 \Upsilon_{\ell-1}}\right)^2$.

  Our algorithm satisfies $\eta_j = 0$ for $j > \ell$ and $\lambda_{3,j} = \lambda_{4,j} = \lambda_{5,j} = 0$ for $j \geq \ell$. In other words, only the matrices $\mathbf{W}_2, \ldots, \mathbf{W}_\ell, \mathbf{K}_2, \ldots, \mathbf{K}_{\ell-1}$ are training parameters and the rest of the matrices stay at zeros. Our analysis will ensure that applying (noisy) SGD one can decrease this objective to $\frac{1}{4}\left(\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}\right)^2$, and when this point is reached we move to stage $\ell^\diamond$.

- $\ell^\diamond$ begins with $\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) \leq \mathsf{Thres}_{\ell,\triangledown} = \frac{1}{4}\left(\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}\right)^2$.

  In this stage, our analysis will guarantee that $\mathbf{W}_{\ell,\ell-1}^\top \mathbf{W}_{\ell\triangleleft}$ is extremely close to a rank k_ℓ matrix, so we can apply k-SVD decomposition to get some warm-up choice of $\mathbf{K}_\ell$ satisfying

  $$\|\mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell\triangleleft} - \mathbf{W}_{\ell,\ell-1}^\top \mathbf{W}_{\ell\triangleleft}\|_F$$

  being sufficiently small. Then, we set $\lambda_{3,\ell}, \lambda_{4,\ell}, \lambda_{5,\ell}$ from Definition A.4, and our analysis will ensure that the objective increases to at most $\left(\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}\right)^2$. We move to stage $\ell^\triangledown$.

- $\ell^\triangledown$ begins with $\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) \leq 4\mathsf{Thres}_{\ell,\triangle} = \left(\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}\right)^2$.

  Our algorithm satisfies $\eta_j = 0$ for $j > \ell$ and $\lambda_{3,j} = \lambda_{4,j} = \lambda_{5,j} = 0$ for $j > \ell$. In other words, only the matrices $\mathbf{W}_2, \ldots, \mathbf{W}_\ell, \mathbf{K}_2, \ldots, \mathbf{K}_\ell$ are training parameters and the rest of the matrices stay at zeros. Our analysis will ensure that applying (noisy) SGD one can decrease this objective to $\left(\frac{\alpha_\ell}{(D_\ell)^9\Upsilon_\ell}\right)^2$, so we can move to stage $(\ell+1)^\triangle$.

A.3 Proof of Theorem A.1

We begin by noting that our truncated empirical objective $\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})$ is in fact lip-bounded, lip-Lipschitz continuous, lip-Lipschitz smooth, and lip-second-order smooth for some parameter $\mathsf{lip} = (\widetilde{\kappa}, B')^{O(L)} \cdot \mathsf{poly}\left(B, (c_4(2^L))^{c_4(2^L)}, \log^{c_4(2^L)} \frac{1}{\delta}, d\right)$ that is sufficiently small (see Claim C.5). This parameter lip will eventually go into our running time, but not anywhere else.

Throughout this proof, we assume as if $\lambda_{6,\ell}$ is always set to be $\frac{\varepsilon^2}{(\widetilde{\kappa}_\ell)^2}$ for convenience, where $\varepsilon^2 = \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})$ is the current objective value. (We can assume so because Algorithm 1 will iteratively shrink the target error ε_0 by a factor of 2.)

Stage $\ell^\triangle$. Suppose we begin this stage with the promise that (guaranteed by the previous stage)

$$\varepsilon^2 = \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) \leq \left(\frac{\alpha_{\ell-1}}{(D_{\ell-1})^9 \Upsilon_{\ell-1}}\right)^2 \quad \text{and} \quad \left\{\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2] \leq \tau_j\right\}_{j<\ell} \tag{A.1}$$

and Algorithm 1 will ensure that $\mathbf{W}_\ell = 0$ is now added to the trainable parameters.

Our main difficulty is to prove (see Theorem F.10) that whenever (A.1) holds, for every small $\eta_1 > 0$, there must exist some update direction $(\mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})})$ satisfying

- $\|\mathbf{K}^{(\mathsf{new})} - \mathbf{K}\|_F \leq \eta_1 \cdot \mathsf{poly}(\widetilde{\kappa})$,
- $\mathbb{E}_{\mathbf{D}} \|\mathbf{W}^{(\mathsf{new})} - \mathbf{W}\|_F^2 \leq \eta_1 \cdot \mathsf{poly}(\widetilde{\kappa})$,
- $\mathbb{E}_{\mathbf{D}}\left[\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})})\right] \leq \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) - \eta_1(0.7\varepsilon^2 - 2\alpha_{\ell+1}^2)$.

Therefore, as long as $\varepsilon^2 > 4\alpha_{\ell+1}^2$, by classical theory from optimization (see Fact I.11 for completeness), we know that

$$\text{either } \|\nabla\widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K})\|_F > \frac{\varepsilon^2}{\mathsf{poly}(\widetilde{\kappa})} \text{ or } \lambda_{\min}\left(\nabla^2\widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K})\right) \leq -\frac{\varepsilon^2}{\mathsf{poly}(\widetilde{\kappa})} \ . \tag{A.2}$$

This means, the current point cannot be an (even approximate) second-order critical point. Invoking known results on stochastic non-conex optimization [21], we know starting from this point, (noisy) SGD can decrease the objective. Note the objective will continue to decrease at least until $\varepsilon^2 \leq 8\alpha_{\ell+1}^2$, but we do not need to wait until the objective is this small, and whenever ε hits $\frac{1}{2}\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}$, we can go into stage $\ell^\diamond$.

Remark A.5. In order to apply SGD to decrease the objective, we need to maintain that the boundedness $\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2] \leq \tau_j$ in (A.1) always holds. This is ensured because of *self-regularization*: we proved that (1) whenever (A.1) holds it must satisfy a tighter bound $\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2] \leq 2B_j \ll \tau_j$, and (2) the quantity $\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2]$ satisfies a Lipschitz continuity statement (see Claim C.3). Specifically, if we move by η in step length, then $\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2]$ is affected by at most $\eta \cdot \left(\prod_{j=\ell}^{i}\mathsf{poly}(\tau_j, c_3(2^j))\right)$. If we choose the step length of SGD to be smaller than this amount, then the quantity $\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2]$ self-regularizes. (This Lipschitz continuity factor also goes into the running time.)

Stage $\boldsymbol{\ell^\diamond}$. Using $\varepsilon^2 \leq \frac{1}{4}\left(\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}\right)^2$, we shall have a theorem to derive that[19]

$$\left\|\mathbf{W}_{\ell,\ell-1}^\top\mathbf{W}_{\ell\triangleleft} - \mathbf{M}\right\|_F^2 \leq \frac{\mathsf{poly}(\widetilde{\kappa}_\ell)}{(D_\ell)^4\Upsilon_\ell}$$

for some matrix $\mathbf{M}$ with rank k_ℓ and singular values between $[\frac{1}{\mathsf{poly}(k,L)}, \mathsf{poly}(k,L)]$. Therefore, applying k-SVD decomposition on $\mathbf{W}_{\ell,\ell-1}^\top\mathbf{W}_{\ell\triangleleft}$, one can derive a warm-up solution of $\mathbf{K}_\ell$ satisfying

$$\|\mathbf{K}_{\ell,\ell-1}^\top\mathbf{K}_{\ell\triangleleft} - \mathbf{W}_{\ell,\ell-1}^\top\mathbf{W}_{\ell\triangleleft}\|_F^2 \leq \frac{\mathsf{poly}(\widetilde{\kappa}_\ell)}{(D_\ell)^4\Upsilon_\ell} \ .$$

Note that, without loss of generality, we can assume $\|\mathbf{K}_\ell\|_F \leq \mathsf{poly}(k,L) \leq \widetilde{\kappa}_\ell/100$ and

$$\|\mathbf{K}_{\ell,\ell-1}^\top\mathbf{K}_{\ell,\ell-1} - \mathbf{W}_{\ell,\ell-1}^\top\mathbf{W}_{\ell,\ell-1}\|_F^2 \leq \mathsf{poly}(\widetilde{\kappa}_\ell) \quad \text{and} \quad \|\mathbf{K}_\ell^\top\mathbf{K}_\ell - \mathbf{W}_\ell^\top\mathbf{W}_\ell\|_F^2 \leq \mathsf{poly}(\widetilde{\kappa}_\ell)$$

(This can be done by left/right multiplying the SVD solution as the solution is not unique.

Since we have chosen regularizer weights (see Definition A.4)

$$\lambda_{6,\ell} = \frac{\varepsilon^2}{(\widetilde{\kappa}_\ell)^2}, \quad \lambda_{3,\ell} = \frac{\alpha_\ell^2}{D_\ell\cdot\Upsilon_\ell}, \quad \lambda_{4,\ell} = \frac{\alpha_\ell^2}{(D_\ell)^7\Upsilon_\ell^2}, \quad \lambda_{5,\ell} = \frac{\alpha_\ell^2}{(D_\ell)^{13}\Upsilon_\ell^3}$$

with the introduction of new trainable variables $\mathbf{K}_\ell$, our objective has increased by at most

$$\begin{aligned}&\lambda_{6,\ell}\frac{(\widetilde{\kappa}_\ell)^2}{100} + \lambda_{3,\ell}\cdot\frac{\mathsf{poly}(\widetilde{\kappa}_\ell)}{(D_\ell)^4\Upsilon_\ell} + \lambda_{4,\ell}\cdot\mathsf{poly}(\widetilde{\kappa}_\ell) + \lambda_{5,\ell}\cdot\mathsf{poly}(\widetilde{\kappa}_\ell)\\ &\leq \frac{\varepsilon^2}{100} + \frac{\alpha_\ell^2}{\Upsilon_\ell^2(D_\ell)^4} + +\frac{\alpha_\ell^2}{\Upsilon_\ell^2(D_\ell)^6} + +\frac{\alpha_\ell^2}{\Upsilon_\ell^3(D_\ell)^{12}} \leq \frac{1}{4}\left(\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}\right)^2\end{aligned}$$

[19] In the language of later sections, Corollary E.4a implies

$$\left\|\mathbf{Q}_{\ell-1}^\top\overline{\mathbf{W}}_{\ell,\ell-1}^\top\overline{\mathbf{W}}_{\ell\triangleleft}\mathbf{Q}_{\ell\triangleleft} - \overline{\mathbf{W}^\star}_{\ell,\ell-1}^\top\overline{\mathbf{W}^\star}_{\ell\triangleleft}\right\|_F^2 \leq \frac{1}{(D_\ell)^4\Upsilon_\ell} \ .$$

Since $\overline{\mathbf{W}^\star}_{\ell,\ell-1}^\top\overline{\mathbf{W}^\star}_{\ell\triangleleft}$ is of rank k_ℓ, this means $\mathbf{Q}_{\ell-1}^\top\overline{\mathbf{W}}_{\ell,\ell-1}^\top\overline{\mathbf{W}}_{\ell\triangleleft}\mathbf{Q}_{\ell\triangleleft}$ is close to rank k_ℓ. Since our notation $\overline{\mathbf{W}}_{\ell,j}\mathbf{Q}_j$ is only an abbreviation of $\mathbf{W}_{\ell,j}(\mathbf{R}_j\mathbf{U}_j * \mathbf{R}_j\mathbf{U}_j)$ for some well conditioned matrix $(\mathbf{R}_j\mathbf{U}_j * \mathbf{R}_j\mathbf{U}_j)$, this also implies $\mathbf{W}_{\ell,\ell-1}^\top\mathbf{W}_{\ell\triangleleft}$ is close to being rank k_ℓ.

This means we can move to stage $\ell^\triangledown$.

Stage $\boldsymbol{\ell}^\triangledown$. We begin this stage with the promise

$$\varepsilon^2 = \widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K}) \le \left(\frac{\alpha_\ell}{(D_\ell)^3\sqrt{\Upsilon_\ell}}\right)^2 \quad \text{and} \quad \left\{\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2] \le \tau_j\right\}_{j<\ell} \tag{A.3}$$

and our trainable parameters are $\mathbf{W}_1,\dots,\mathbf{W}_\ell,\mathbf{K}_1,\dots,\mathbf{K}_\ell$. This time, we have another Theorem F.11 to guarantee that as long as (A.3) is satisfied, then (A.2) still holds (namely, it is not an approximate second-order critical point). Therefore, one can still apply standard (noisy) SGD to sufficiently decrease the objective at least until $\varepsilon^2 \le 8\alpha_{\ell+1}^2$ (or until arbitrarily small $\varepsilon^2 > 0$ if $\ell = L$). This is much smaller than the requirement of stage $(\ell+1)^\vartriangle$.

For similar reason as Remark A.5, we have self-regularization so $\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2] \le \tau_j$ (for $j<\ell$) holds throughout the optimization process. In addition, this time Theorem F.11 also implies that whenever we exit this stage, namely when $\varepsilon \le \frac{\alpha_\ell}{(D_\ell)^9\Upsilon_\ell}$ is satisfied, then $\mathbb{E}_{x\sim\mathcal{D}}[\|S_\ell(x)\|_2^2] \le 2B_\ell$.

End of Algorithm. Note in the last $L^\triangledown$ stage, we can decrease the objective until arbitrarily small $\varepsilon^2 > 0$ and thus we have $\widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K}) \le \varepsilon^2$. Applying Proposition C.7 (relating empirical and population losses) and Claim C.1 (relating truncated and quadratic losses), we have

$$\widetilde{\mathbf{Obj}}(\mathcal{D};\mathbf{W},\mathbf{K}) \le 2\varepsilon^2 \quad \text{and} \quad \mathbf{Obj}(\mathcal{D};\mathbf{W},\mathbf{K}) \le 3\varepsilon^2 \ .$$

Time Complexity. As for the time complexity, since our objective satisfies lip-Lipschitz property until second-order smoothness, the time complexity of SGD depends only on $\mathsf{poly}(\mathsf{lip}, \frac{1}{\varepsilon}, d)$ (see [21]).

Quadratic Activation. We used the truncated quadratic activation $\widetilde{\sigma}_j(x)$ only for the purpose to make sure the training objective is sufficiently smooth. Our analysis will ensure that, in fact, when substituting $\widetilde{\sigma}_j(x)$ back with the vanilla quadratic activation, the objective is also small (see (F.8) and (F.9)).

A.4 Our Theorem on Agnostic Learning

For notational simplicity, throughout this paper we have assumed that the exact true label $G^\star(x)$ is given for every training input $x \sim \mathcal{Z}$. This is called *realizable learning.*

In fact, our proof trivially generalizes to the *agnostic learning* case at the expense of introducing extra notations. Suppose that $Y(x) \in \mathbb{R}$ is a label function (not necessarily a polynomial) and is OPT close to some target network, or in symbols,

$$\mathbb{E}_{x\sim\mathcal{D}}\left[(G^\star(x) - Y(x))^2\right] \le \mathsf{OPT} \ .$$

Suppose the algorithm is given training set $\{(x, Y(x)) : x \in \mathcal{Z}\}$, so the loss function now becomes

$$\mathbf{Loss}(x;\mathbf{W},\mathbf{K}) = (F(x;\mathbf{W},\mathbf{K}) - Y(x))^2$$

Suppose in addition that $|Y(x)| \le B$ almost surely. Then,[20]

Theorem A.6. *For every constant $\gamma > 1$, for any desired accuracy $\varepsilon \in (\sqrt{\mathsf{OPT}}, 1)$, in the same setting as Theorem A.1, Algorithm 1 can find a point with*

$$\widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K}) \le (1+\frac{1}{\gamma})\mathsf{OPT}+\varepsilon^2 \quad \widetilde{\mathbf{Obj}}(\mathcal{D};\mathbf{W},\mathbf{K}) \le (1+\frac{1}{\gamma})\mathsf{OPT}+\varepsilon^2 \quad \mathbf{Obj}(\mathcal{D};\mathbf{W},\mathbf{K}) \le (1+\frac{1}{\gamma})\mathsf{OPT}+\varepsilon^2$$

[20] The proof is nearly identical. The main difference is to replace the use of $\mathsf{OPT}_{\le\ell} \le 2\alpha_{\ell+1}^2$ with $\mathsf{OPT}_{\le\ell} \le O(\alpha_{\ell+1}^2) + (1+\frac{1}{\gamma})\mathsf{OPT}$ (when invoking Lemma F.8) in the final proofs of Theorem F.10 and Theorem F.11.

B Notations and Preliminaries

We denote by $\|w\|_2$ and $\|w\|_\infty$ the Euclidean and infinity norms of vectors w, and $\|w\|_0$ the number of non-zeros of w. We also abbreviate $\|w\| = \|w\|_2$ when it is clear from the context. We use $\|\mathbf{W}\|_F, \|\mathbf{W}\|_2$ to denote the Frobenius and spectral norm of matrix $\mathbf{W}$. We use $\mathbf{A} \succeq \mathbf{B}$ to denote that the difference between two symmetric matrices $\mathbf{A} - \mathbf{B}$ is positive semi-definite. We use $\sigma_{\min}(\mathbf{A}), \sigma_{\max}(\mathbf{A})$ to denote the minimum and maximum singular values of a rectangular matrix, and $\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})$ for the minimum and maximum eigenvalues.

We use $\mathcal{N}(\mu, \sigma)$ to denote Gaussian distribution with mean μ and variance σ; or $\mathcal{N}(\mu, \Sigma)$ to denote Gaussian vector with mean μ and covariance Σ. We use $\mathbb{1}_{event}$ or $\mathbb{1}[event]$ to denote the indicator function of whether *event* is true.

We denote $\mathbf{Sum}(x) = \sum_i x_i$ as the sum of the coordinate of this vector.

We use $\sigma(x) = x^2$ as the quadratic activation function.

Definition B.1. *Given any degree-q homogenous polynomial $f(x) = \sum_{I \in \mathbb{N}^n \colon \|I\|_1 = q} a_I \prod_{j \in [n]} x_j^{I_j}$, recall we have defined*

$$\mathcal{C}_x(f) \stackrel{\text{def}}{=} \sum_{I \in \mathbb{N}^n \colon \|I\|_1 = q} a_I^2$$

When it is clear from the context, we also denote $\mathcal{C}(f) = \mathcal{C}_x(f)$.

B.1 Symmetric Tensor

When it is clear from the context, in this paper sets can be multisets. This allows us to write $\{i, i\}$. We also support notation $\forall \{i, j\} \in \binom{n+1}{2}$ to denote all possible (unordered) sub multi-sets of $[n]$ with cardinality 2.

Definition B.2 (symmetric tensor)**.** *The symmetric tensor $*$ for two vectors $x, y \in \mathbb{R}^n$ is given as:*

$$[x * y]_{\{i,j\}} = a_{i,j} x_i x_j, \quad \forall 1 \le i \le j \le p$$

for $a_{i,i} = 1$ and $a_{i,j} = \sqrt{2}$ for $j \neq i$. Note $x * y \in \mathbb{R}^{\binom{n+1}{2}}$. The symmetric tensor $*$ for two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}_{m \times n}$ is given as:

$$[\mathbf{X} * \mathbf{Y}]_{p,\{i,j\}} = a_{i,j} \mathbf{X}_{p,i} \mathbf{X}_{p,j}, \quad \forall p \in [m], 1 \le i \le j \le p$$

and it satisfies $\mathbf{X} * \mathbf{Y} \in \mathbb{R}^{m \times \binom{n+1}{2}}$.

It is a simple exercise to verify that $\langle x, y \rangle^2 = \langle x * x, y * y \rangle$.

Definition B.3 (**Sym**)**.** *For any $\mathbf{M} \in \mathbb{R}^{\binom{n+1}{2} \times \binom{n+1}{2}}$, define $\mathbf{Sym}(\mathbf{M}) \in \mathbb{R}^{\binom{n+1}{2} \times \binom{n+1}{2}}$ to be the "twice-symmetric" version of $\mathbf{M}$. For every $1 \le i \le j \le n$ and $1 \le k \le l \le n$, define*[21]

$$\mathbf{Sym}(\mathbf{M})_{\{i,j\},\{k,l\}} \stackrel{\text{def}}{=} \frac{\sum_{\{p,q\},\{r,s\} \in \binom{n+1}{2} \wedge \{p,q,r,s\} = \{i,j,k,l\}} a_{p,q} a_{r,s} \mathbf{M}_{\{p,q\},\{r,s\}}}{a_{i,j} a_{k,l} \cdot \left| \left\{ \{p,q\},\{r,s\} \in \binom{n+1}{2} \colon \{p,q,r,s\} = \{i,j,k,l\} \right\} \right|}$$

Fact B.4. $\mathbf{Sym}(\mathbf{M})$ *satisfies the following three properties.*

[21] For instance, when $i, j, k, l \in [n]$ are distinct, this means

$$\mathbf{Sym}(\mathbf{M})_{\{i,j\},\{k,l\}} = \frac{\mathbf{M}_{\{i,j\},\{k,l\}} + \mathbf{M}_{\{i,k\},\{j,l\}} + \mathbf{M}_{\{i,l\},\{j,k\}} + \mathbf{M}_{\{j,k\},\{i,l\}} + \mathbf{M}_{\{j,l\},\{i,k\}} + \mathbf{M}_{\{k,l\},\{i,j\}}}{6} .$$

- $(z * z)^\top \mathbf{Sym}(\mathbf{M})(z * z) = (z * z)^\top \mathbf{M}(z * z)$ *for every* $z \in \mathbb{R}^n$;
- *If* $\mathbf{M}$ *is symmetric and satisfies* $\mathbf{M}_{\{i,j\},\{k,l\}} = 0$ *whenever* $i \neq j$ *or* $k \neq l$, *then* $\mathbf{Sym}(\mathbf{M}) = \mathbf{M}$.
- $O(1)\|\mathbf{M}\|_F^2 \geq \mathcal{C}_z\big((z * z)^\top \mathbf{M}(z * z)\big) \geq \|\mathbf{Sym}(\mathbf{M})\|_F^2$

It is not hard to derive the following important property (proof see Appendix I.3)

Lemma B.5. *If* $\mathbf{U} \in \mathbb{R}^{p\times p}$ *is unitary and* $\mathbf{R} \in \mathbb{R}^{s\times p}$ *for* $s \geq \binom{p+1}{2}$, *then there exists some* unitary *matrix* $\mathbf{Q} \in \mathbb{R}^{\binom{p+1}{2}\times\binom{p+1}{2}}$ *so that* $\mathbf{RU} * \mathbf{RU} = (\mathbf{R} * \mathbf{R})\,\mathbf{Q}$.

B.2 Network Initialization and Network Tensor Notions

We show the following lemma on random initialization (proved in Appendix I.2).

Lemma B.6. *Let* $\mathbf{R}_\ell \in \mathbb{R}^{\binom{k_\ell+1}{2}\times k_\ell}$ *be a random matrix such that each entry is i.i.d. from* $\mathcal{N}\left(0, \frac{1}{k_\ell^2}\right)$, *then with probability at least* $1-p$, $\mathbf{R}_\ell * \mathbf{R}_\ell$ *has singular values between* $[\frac{1}{O(k_\ell^4 p^2)}, O(1 + \frac{1}{k_\ell^2}\log\frac{k_\ell}{p})]$, *and* $\|\mathbf{R}_\ell\|_2 \leq O(1 + \frac{\sqrt{\log(1/p)}}{k_\ell})$.

As a result, with probability at least 0.99, it satisfies for all $\ell = 2, 3, \dots, L$, *the square matrices* $\mathbf{R}_\ell * \mathbf{R}_\ell$ *have singular values between* $[\frac{1}{O(k_\ell^4 L^2)}, O(1 + \frac{\log(Lk_\ell)}{k_\ell})]$ *and* $\|\mathbf{R}_\ell\|_2 \leq O(1 + \frac{\sqrt{\log L}}{k_\ell})$.

Through out the analysis, it is more convenient to work on the matrix symmetric tensors. For every $\ell = 2, 3, 4, \dots, L$ and every $j \in \mathcal{J}_\ell \setminus \{0, 1\}$, we define

$$\begin{aligned}
\overline{\mathbf{W}^\star}_{\ell,j} &\stackrel{\text{def}}{=} \mathbf{W}^\star_{\ell,j}\big(\mathbf{I} * \mathbf{I}\big) = \mathbf{W}^\star_{\ell,j} * \mathbf{W}^\star_{\ell,j} && \in \mathbb{R}^{k_\ell \times \binom{k_j+1}{2}} \\
\overline{\mathbf{W}}_{\ell,j} &\stackrel{\text{def}}{=} \mathbf{W}_{\ell,j}\big(\mathbf{R}_j * \mathbf{R}_j\big) = \mathbf{W}_{\ell,j}\mathbf{R}_j * \mathbf{W}_{\ell,j}\mathbf{R}_j && \in \mathbb{R}^{m \times \binom{k_j+1}{2}} \\
\overline{\mathbf{K}}_{\ell,j} &\stackrel{\text{def}}{=} \mathbf{K}_{\ell,j}(\mathbf{R}_j * \mathbf{R}_j) = \mathbf{K}_{\ell,j}\mathbf{R}_j * \mathbf{K}_{\ell,j}\mathbf{R}_j && \in \mathbb{R}^{k_\ell \times \binom{k_j+1}{2}}
\end{aligned}$$

so that

$$\forall z \in \mathbb{R}^{k_j}: \quad \begin{aligned} \overline{\mathbf{W}^\star}_{\ell,j}(z * z) &= \mathbf{W}^\star_{\ell,j}\sigma(z) \\ \overline{\mathbf{W}}_{\ell,j}\,(z * z) &= \mathbf{W}_{\ell,j}\sigma(\mathbf{R}_j z) \\ \overline{\mathbf{K}}_{\ell,j}\,(z * z) &= \mathbf{K}_{\ell,j}\sigma(\mathbf{R}_j z) \end{aligned}$$

For convenience, whenever $j \in \mathcal{J}_\ell \cap \{0, 1\}$, we also write

$$\overline{\mathbf{W}^\star}_{\ell,j} = \mathbf{W}^\star_{\ell,j} \quad \overline{\mathbf{W}}_{\ell,j} = \mathbf{W}_{\ell,j} \quad \overline{\mathbf{K}}_{\ell,j} = \mathbf{K}_{\ell,j}$$

We define

$$\begin{aligned}
\overline{\mathbf{W}^\star}_\ell &= \big(\overline{\mathbf{W}^\star}_{\ell,j}\big)_{j\in\mathcal{J}_\ell} \in \mathbb{R}^{k_\ell\times *}, &\quad \overline{\mathbf{W}}_\ell &= \big(\overline{\mathbf{W}}_{\ell,j}\big)_{j\in\mathcal{J}_\ell} \in \mathbb{R}^{m\times *}, &\quad \overline{\mathbf{K}}_\ell &= \big(\overline{\mathbf{K}}_{\ell,j}\big)_{j\in\mathcal{J}_\ell} \in \mathbb{R}^{k_\ell\times *} \\
\overline{\mathbf{W}^\star}_{\ell\triangleleft} &= \big(\overline{\mathbf{W}^\star}_{\ell,j}\big)_{j\in\mathcal{J}_\ell, j\neq\ell-1}, & \overline{\mathbf{W}}_{\ell\triangleleft} &= \big(\overline{\mathbf{W}}_{\ell,j}\big)_{j\in\mathcal{J}_\ell, j\neq\ell-1}, & \overline{\mathbf{K}}_{\ell\triangleleft} &= \big(\overline{\mathbf{K}}_{\ell,j}\big)_{j\in\mathcal{J}_\ell, j\neq\ell-1}
\end{aligned}$$

Fact B.7. *Singular values of* $\mathbf{W}^\star_{\ell,j}$ *are in* $[1/\kappa, \kappa]$. *Singular values of* $\overline{\mathbf{W}^\star}_\ell$ *are in* $[1/\kappa, \ell\kappa]$.

C Useful Properties of Our Objective Function

C.1 Closeness: Population Quadratic vs. Population Truncated Loss

Claim C.1. *Suppose for every* $\ell \in [L]$, $\|\mathbf{K}_\ell\|_2, \|\mathbf{W}_\ell\|_2 \leq \widetilde{\kappa}_\ell$ *for some* $\widetilde{\kappa}_\ell \geq k_\ell + L + \kappa$ *and* $\mathbb{E}_{x\sim\mathcal{D}}[\|S_\ell(x)\|^2] \leq \tau_\ell$ *for some* $\tau_\ell \geq \widetilde{\kappa}_\ell$. *Then, for every* $\varepsilon \in (0, 1]$, *when choosing*

$$\textit{truncation parameter:} \qquad B'_\ell \geq \tau_\ell^2 \cdot \mathsf{poly}(\widetilde{\kappa}_\ell) \cdot \Omega(2^\ell c_4(2^\ell)\log(dL/\varepsilon))^{c_4(2^\ell)} \ ,$$

we have for every integer constant $p \le 10$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\widetilde{F}(x) - F(x)\right)^p\right] \le \varepsilon \quad \textit{and} \quad \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\|\widetilde{S}_\ell(x) - S_\ell(x)\|_2\right)^p\right] \le \varepsilon$$

Proof of Claim C.1. We first focus on $\widetilde{S}_\ell(x) - S_\ell(x)$. We first note that for every $S_\ell(x), \widetilde{S}_\ell(x)$, there is a crude (but absolute) upper bound:

$$\|S_\ell(x)\|_2, \|\widetilde{S}_\ell(x)\|_2 \le (\widetilde{\kappa}_\ell k_\ell \ell)^{O(2^\ell)} \|x\|_2^{2^\ell} =: C_1 \|x\|_2^{2^\ell} \ .$$

By the isotropic property of x (see (5.1)) and the hyper-contractivity (see (5.3)), we know that for R_1 is as large as $R_1 = (d \log(C_1/\varepsilon))^{\Omega(2^\ell)}$, it holds that

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathbb{1}_{\|x\|_2^{2^\ell} \ge R_1} \|x\|_2^{p\cdot 2^\ell}\right] \le \frac{\varepsilon}{2C_1^p}$$

This implies

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\|\widetilde{S}_\ell(x) - S_\ell(x)\|_2\right)^p \mathbb{1}_{\|x\|_2^{2^\ell} \ge R_1}\right] \le \frac{\varepsilon}{2} \tag{C.1}$$

Next, we consider the remaining part, since $\mathbb{E}_{x\sim\mathcal{D}}[\|S_\ell(x)\|^2] \le \tau_\ell$, we know that when $B'_\ell \ge \tau_\ell \cdot \Omega(c_4(2^\ell))^{c_4(2^\ell)} \log^{c_4(2^\ell)}(C_1 R_1 L/\varepsilon)$, by the hyper-contractivity Property 5.3, we have for every fixed ℓ,

$$\mathbf{Pr}[\|\mathbf{R}_\ell S_\ell(x)\|_2 \ge B'_\ell] \le \frac{\varepsilon}{2(2C_1R_1)^p L}$$

Therefore, with probability at least $1 - \frac{\varepsilon}{2(2C_1R_1)^p}$, at every layer ℓ, the value plugged into $\widetilde{\sigma}$ and σ are the same. As a result,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\|\widetilde{S}_\ell(x) - S_\ell(x)\|_2\right)^p \mathbb{1}_{\|x\|_2^{2^\ell} \le R_1}\right] \le (2C_1R_1)^p \,\mathbf{Pr}\left[\exists \ell' \le \ell, \|\mathbf{R}_{\ell'} S_{\ell'}(x)\|_2 \ge B'_\ell\right] \le \varepsilon/2 \tag{C.2}$$

Putting together (C.1) and (C.2) we complete the proof that

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\|\widetilde{S}_\ell(x) - S_\ell(x)\|_2\right)^p\right] \le \varepsilon$$

An identical proof also shows that

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\|\mathbf{Sum}(\widetilde{F}_\ell(x)) - \mathbf{Sum}(F_\ell)(x)\|_2\right)^p\right] \le \varepsilon$$

Thus, scaling down by a factor of Lp we can derive the bound on $\mathbb{E}_{x\sim\mathcal{D}}\left[\left(\widetilde{F}(x) - F(x)\right)^p\right]$. $\square$

C.2 Covariance: Empirical vs. Population

Recall that our isotropic Property 5.1 says for every $w \in \mathbb{R}^d$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\langle w, x\rangle^2] \le O(1) \cdot \|w\|^2 \quad \text{and} \quad \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\langle w, S_1(x)\rangle^2] \le O(1) \cdot \|w\|^2 \ .$$

Below we show that this also holds for the empirical dataset as long as enough samples are given.

Proposition C.2. *As long as* $N = d^2 \cdot \log^{\Omega(1)} \frac{d}{\delta}$, *with probability at least* $1 - \delta$ *over the random choice of* $\mathcal{Z}$, *for every vector* $w \in \mathbb{R}^d$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{Z}}[\langle w, x\rangle^4] \le O(1) \cdot \|w\|^2 \quad \textit{and} \quad \mathop{\mathbb{E}}_{x\sim\mathcal{Z}}[\langle w, S_1(x)\rangle^4] \le O(1) \cdot \|w\|^2$$

$$\forall x \in \mathcal{Z}\colon\ \max\{\|x\|^2, \|S_1(x)\|^2\} \le d \log^{O(1)} \frac{d}{\delta}$$

Proof of Proposition C.2. Our isotropic Property 5.1 together with the hyper-contractivity Property 5.3 implies if $N \geq d \log^{\Omega(1)} \frac{d}{\delta}$, then with probability at least $1 - \delta/4$,

$$\forall x \in \mathcal{Z}: \quad \|x\|^2 \leq R_3 \quad \text{and} \quad \|S_1(x)\|^2 \leq R_3$$

Where $R_3 = d \cdot \log^{O(1)} \frac{d}{\delta}$. Next, conditioning on this event, we can apply Bernstein's inequality to derive that as long as $N \geq \Omega(R_3 \cdot \log \frac{1}{\delta_0})$ with probability at least $1 - \delta_0$, for every fixed $w \in \mathbb{R}^d$,

$$\mathbf{Pr}_{x \sim \mathcal{D}} \left[\langle w, x \rangle^4 \geq \Omega(1) \right] \geq 1 - \delta_0$$

Taking an epsilon-net over all possible w finishes the proof. $\square$

C.3 Lipschitz Continuity: Population Quadratic

Claim C.3. *Suppose* $\mathbf{K}$ *satisfies* $\|\mathbf{K}_j\|_2 \leq \tau_j$ *for every* $j \in \{2, 3, \cdots, L\}$ *where* $\tau_j \geq \overline{k}_j + \kappa + L$, *and suppose for some* $\ell \in \{2, 3, \cdots, L\}$, $\mathbf{K}_\ell$ *replaced with* $\mathbf{K}'_\ell = \mathbf{K}_\ell + \mathbf{\Delta}_\ell$ *with any* $\|\mathbf{\Delta}_\ell\|_F \leq \left(\prod_{j=\ell}^{L} \mathsf{poly}(\tau_j, c_3(2^j)) \right)^{-1}$, *then for every* $i \geq \ell$

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\left| \|S'_i(x)\|^2 - \|S_i(x)\|^2 \right| \right] \leq \eta \cdot \left(\prod_{j=\ell}^{i} \mathsf{poly}(\tau_j, c_3(2^j)) \right)$$

and for every $i < \ell$ *obviously* $S_i(x) = S'_i(x)$.

Proof of Claim C.3. We first check the stability with respect to $\mathbf{K}$, and suppose without loss of generality that only one $\mathbf{W}_\ell$ is changed for some ℓ. For notation simplicity, suppose we do an update $\mathbf{K}'_\ell = \mathbf{K}_\ell + \eta \mathbf{\Delta}_\ell$ for $\|\mathbf{\Delta}_\ell\|_F = 1$. We use S' to denote the sequence of S after the update, and we have $S'_j(x) = S_j(x)$ for every $j < \ell$. As for $S'_\ell(x)$, we have

$$\begin{aligned}
\|S'_\ell(x) - S_\ell(x)\| &\leq \eta \left(\sum_{j \geq 2}^{\ell-1} \|\mathbf{\Delta}_{\ell,j}\|_2 \|\sigma(\mathbf{R}_j S_j(x))\| + \|\mathbf{\Delta}_{\ell,1} S_1(x)\| + \|\mathbf{\Delta}_{\ell,0} x\| \right) \\
&\leq \eta \mathsf{poly}(\overline{k}_\ell, \kappa, L) \left(\sum_{j < \ell} \|S_j(x)\|^2 + \|\mathbf{\Delta}_{\ell,1} S_1(x)\| + \|\mathbf{\Delta}_{\ell,0} x\| \right)
\end{aligned}$$

so using $\mathbb{E}_{x \sim \mathcal{D}}[\|S_j(x)\|^2] \leq \tau_j$, the isotropic Property 5.1 and the hyper-contractivity Property 5.4, we can write

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\|S'_\ell(x) - S_\ell(x)\|^2] \leq \eta^2 \mathsf{poly}(\tau_\ell, c_3(2^\ell)) =: \theta_\ell$$

As for later layers $i > \ell$, we have

$$\|S'_i(x) - S_i(x)\| \leq 4 \sum_{j \geq 2}^{i-1} \|\mathbf{K}_{i,j}\|_2 \|\mathbf{R}_j\|_2^2 (\|S_j(x)\| \|S'_j(x) - S_j(x)\| + \|S'_j(x) - S_j(x)\|^2)$$

so taking square and expectation, and using hyper-contractivity Property 5.4 again, (and using our assumption on η)[22]

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \|S'_i(x) - S_i(x)\|^2 \leq \mathsf{poly}(\tau_i, c_3(2^i)) \cdot \theta_{i-1} =: \theta_i$$

[22]This requires one to repeatedly apply the trivial inequality $ab \leq \eta a^2 + b^2/\eta$.

by recursing $\theta_i = \mathsf{poly}(\tau_i, c_3(2^i)) \cdot \theta_{i-1}$ we have

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \|S'_i(x) - S_i(x)\|^2 \le \left(\prod_{j=\ell}^{i} \mathsf{poly}(\tau_j, c_3(2^j))\right)$$

□

C.4 Lipschitz Continuity: Empirical Truncated Loss in K

Claim C.4. *Suppose the sampled set $\mathcal{Z}$ satisfies the event of Proposition C.2. For every $\mathbf{W}, \mathbf{K}$ satisfying*

$$\forall j = 2, 3, \dots, L: \qquad \|\mathbf{W}_j\|_2 \le \widetilde{\kappa}_j,\ \|\mathbf{K}_j\|_2 \le \widetilde{\kappa}_j$$

for some $\widetilde{\kappa}_j \ge k_j + \kappa + L$. Then, for any $\ell \in \{2, 3, \cdots, L-1\}$ and consider $\mathbf{K}_\ell$ replaced with $\mathbf{K}'_\ell = \mathbf{K}_\ell + \mathbf{\Delta}_\ell$ for any $\|\mathbf{\Delta}_\ell\|_F \le \frac{1}{\mathsf{poly}(\widetilde{\kappa}_\ell, \overline{B}'_\ell, d)}$. Then,

$$|\widetilde{\mathbf{Loss}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) - \widetilde{\mathbf{Loss}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}')| \le \alpha_{\ell+1}\sqrt{\widetilde{\mathbf{Loss}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})} \cdot \mathsf{poly}(\widetilde{\kappa}_j, \overline{B}'_j) \cdot \|\mathbf{\Delta}_\ell\|_F$$

Proof of Claim C.4. Let us denote $\varepsilon^2 = \widetilde{\mathbf{Loss}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})$. For notation simplicity, suppose we do an update $\mathbf{K}'_\ell = \mathbf{K}_\ell + \eta\mathbf{\Delta}_\ell$ for $\eta > 0$ and $\|\mathbf{\Delta}_\ell\|_F = 1$. We use $\widetilde{S}'$ to denote the sequence of $\widetilde{S}$ after the update, and we have $\widetilde{S}'_j(x) = \widetilde{S}_j(x)$ for every $j < \ell$. As for $\widetilde{S}'_\ell(x)$, we have (using the boundedness of $\widetilde{\sigma}$)

$$\begin{aligned}\|\widetilde{S}'_\ell(x) - \widetilde{S}_\ell(x)\| &\le \eta\left(\sum_{j\ge 2}^{\ell-1} \|\mathbf{\Delta}_{\ell,j}\|_2 \|\widetilde{\sigma}(\widetilde{S}_j(x))\| + \|\mathbf{\Delta}_{\ell,1} S_1(x)\| + \|\mathbf{\Delta}_{\ell,0} x\|\right) \\ &\le \eta L \overline{B}'_\ell + \eta\left(\|\mathbf{\Delta}_{\ell,1} S_1(x)\| + \|\mathbf{\Delta}_{\ell,0} x\|\right)\end{aligned}$$

As for later layers $i > \ell$, we have (using the Lipschitz continuity of $\widetilde{\sigma}$)

$$\begin{aligned}\|\widetilde{S}'_i(x) - \widetilde{S}_i(x)\| &\le \sum_{j\ge 2}^{i-1} \|\mathbf{K}_{i,j}\|_2 B'_j \|\mathbf{R}_j\|_2 \|\widetilde{S}'_j(x) - \widetilde{S}_j(x)\| \\ &\le \cdots \le \prod_{j=\ell+1}^{i} (\widetilde{\kappa}_j \overline{B}'_j L^2)\left(\eta L \overline{B}'_\ell + \eta\left(\|\mathbf{\Delta}_{\ell,1} S_1(x)\| + \|\mathbf{\Delta}_{\ell,0} x\|\right)\right) =: p_i\end{aligned}$$

As for $\widetilde{F}(x)$, recall

$$\widetilde{F}(x) = \sum_i \alpha_i \left\|\mathbf{W}_{i,0} x + \mathbf{W}_{i,1} S_1(x) + \sum_{j\in\{2,3,\cdots,i-1\}} \mathbf{W}_{i,j}\sigma\left(\mathbf{R}_j \widetilde{S}_j(x)\right)\right\|^2 =: \sum_i \alpha_i \|A_i\|^2 \ .$$

Using the bound $\|A_i\| \le \|\mathbf{W}_{i,0} x\| + \|\mathbf{W}_{i,1} S_1(x)\| + \mathsf{poly}(\widetilde{\kappa}_i, \overline{B}'_i)$, one can carefully verify[23]

$$\begin{aligned}|\widetilde{F}'(x) - \widetilde{F}(x)| &\le \sum_{i\ge\ell+1} \alpha_i \left(\|A_i\| \cdot p_{i-1} + p_{i-1}^2\right) \cdot \mathsf{poly}(\widetilde{\kappa}_i, \overline{B}'_i) \\ &\le \alpha_{\ell+1}\eta \mathsf{poly}(\widetilde{\kappa}_\ell, \overline{B}'_\ell) \cdot \left(1 + (\|\mathbf{W}_{\ell,0} x\| + \|\mathbf{W}_{\ell,1} S_1(x)\|)(\|\mathbf{\Delta}_{\ell,1} S_1(x)\| + \|\mathbf{\Delta}_{\ell,0} x\|)\right)\end{aligned}$$

[23] This requires us to use the gap assumption between α_{i+1} and α_i, and the sufficient small choice of $\eta > 0$. For instance, the $\eta^2\|\mathbf{\Delta}_{\ell,0} x\|^2$ term diminishes because η is sufficiently small and $\|x\|$ is bounded for every $x \sim \mathcal{Z}$ (see Proposition C.2).

Therefore, we know that

$$
\begin{aligned}
&\left|\left(G^\star(x)-\widetilde{F}(x)\right)^2-\left(G^\star(x)-\widetilde{F}'(x)\right)^2\right| \\
&\le 2\left|G^\star(x)-\widetilde{F}(x)\right|\cdot|\widetilde{F}'(x)-\widetilde{F}(x)|+|\widetilde{F}'(x)-\widetilde{F}(x)|^2 \\
&\le \frac{\alpha_{\ell+1}\eta}{\varepsilon}\cdot\left|G^\star(x)-\widetilde{F}(x)\right|^2+\varepsilon\frac{|\widetilde{F}'(x)-\widetilde{F}(x)|^2}{\alpha_{\ell+1}\eta}+|\widetilde{F}'(x)-\widetilde{F}(x)|^2 \\
&\le \frac{\alpha_{\ell+1}\eta}{\varepsilon}\cdot\left|G^\star(x)-\widetilde{F}(x)\right|^2 \\
&\quad+\varepsilon\alpha_{\ell+1}\eta\,\mathsf{poly}(\widetilde{\kappa}_\ell,\overline{B}'_\ell)\left(1+(\|\mathbf{W}_{\ell,0}x\|^2+\|\mathbf{W}_{\ell,1}S_1(x)\|^2)(\|\mathbf{\Delta}_{\ell,1}S_1(x)\|^2+\|\mathbf{\Delta}_{\ell,0}x\|)^2\right)
\end{aligned}
$$

Note that $2a^2b^2 \le a^4+b^4$ and:

- From Proposition C.2 we have $\mathbb{E}_{x\sim\mathcal{Z}}\|\mathbf{W}_{\ell,0}x\|^4, \mathbb{E}_{x\sim\mathcal{Z}}\|\mathbf{W}_{\ell,1}S_1(x)\|^4 \le \widetilde{\kappa}_\ell$.
- From Proposition C.2 we have $\mathbb{E}_{x\sim\mathcal{Z}}\|\mathbf{\Delta}_{\ell,1}S_1(x)\|^4+\|\mathbf{\Delta}_{\ell,0}x\|^4 \le \mathsf{poly}(\widetilde{\kappa}_\ell)$.
- From definition of ε we have $\mathbb{E}_{x\sim\mathcal{Z}}\left|G^\star(x)-\widetilde{F}(x)\right|^2=\varepsilon^2$.

Therefore, taking expectation we have

$$
\mathbb{E}_{x\sim\mathcal{Z}}\left|\left(G^\star(x)-\widetilde{F}(x)\right)^2-\left(G^\star(x)-\widetilde{F}(x)'\right)^2\right| \le \varepsilon\alpha_{\ell+1}\eta\,\mathsf{poly}(\widetilde{\kappa}_\ell,\overline{B}'_\ell)\ . \qquad \square
$$

C.5 Lipschitz Smoothness: Empirical Truncated Loss (Crude Bound)

Recall a function $f(x)$ over domain $\mathcal{X}$ is

- lip-Lipschitz continuous if $f(y)\le f(x)+\mathsf{lip}\cdot\|y-x\|_F$ for all $x,y\in\mathcal{X}$;
- lip-Lipschitz smooth if $f(y)\le f(x)+\langle\nabla f(x),y-x\rangle+\frac{\mathsf{lip}}{2}\cdot\|y-x\|_F^2$ for all $x,y\in\mathcal{X}$;
- lip-Lipschitz second-order smooth if $f(y)\le f(x)+\langle\nabla f(x),y-x\rangle+\frac{1}{2}(y-x)^\top\nabla f(x)(y-x)+\frac{\mathsf{lip}}{6}\cdot\|y-x\|_F^3$ for all $x,y\in\mathcal{X}$.

We have the following crude bound:

Claim C.5. *Consider the domain consisting of all $\mathbf{W},\mathbf{K}$ with*

$$
\forall j=2,3,\dots,L:\quad \|\mathbf{W}_j\|_2\le\widetilde{\kappa}_j,\ \|\mathbf{K}_j\|_2\le\widetilde{\kappa}_j
$$

for some $\widetilde{\kappa}_j\ge\overline{k}_j+L+\kappa$, we have for every $x\sim\mathcal{D}$,

- $|\widetilde{F}(x;\mathbf{W},\mathbf{K})|\le\mathsf{poly}(\widetilde{\kappa},B')\cdot\sum_\ell(\|\mathbf{W}_{\ell,0}x\|^2+\|\mathbf{W}_{\ell,1}S_1(x)\|^2)$.
- $\widetilde{F}(x;\mathbf{W},\mathbf{K})$ *is lip-Lipschitz continuous, lip-Lipschitz smooth, and lip-Lipschitz second-order smooth in $\mathbf{W},\mathbf{K}$ for $\mathsf{lip}=\prod_\ell(\widetilde{\kappa}_\ell,\overline{B}'_\ell)^{O(1)}\cdot\mathsf{poly}(G^\star(x),\|x\|)$*

Suppose the sampled set $\mathcal{Z}$ satisfies the event of Proposition C.2, then

- $\widetilde{\mathbf{Loss}}(\mathcal{Z};\mathbf{W},\mathbf{K})$ *is lip-Lipschitz continuous, lip-Lipschitz smooth, and lip-Lipschitz second-order smooth in $\mathbf{W},\mathbf{K}$ for $\mathsf{lip}=\prod_\ell(\widetilde{\kappa}_\ell,\overline{B}'_\ell)^{O(1)}\cdot\mathsf{poly}\left(B,(c_4(2^L))^{c_4(2^L)},\log^{c_4(2^L)}\frac{1}{\delta},d\right)$.*

We first state the following bound on chain of derivatives

Claim C.6 (chain derivatives)**.** *For every integer $K > 0$, every functions $f, g_1, g_2, \ldots, g_K : \mathbb{R} \to \mathbb{R}$, and every integer $p_0 > 0$, suppose there exists a value $R_0, R_1 > 1$ and an integer $s \geq 0$ such that*

$$\forall p \in \{0, 1, \cdots, p_0\}, i \in [K] \colon \left| \frac{d^p f(x)}{dx^p} \right| \leq R_0^p, \quad \left| \frac{d^p g_i(x)}{dx^p} \right| \leq R_1^p \ .$$

Then, the function $h(x, w) = f(\sum_{i \in [K]} w_i g_i(x))$ satisfies:

$$\forall p \in \{0, 1, \cdots, p_0\} \colon \left| \frac{\partial^p h(x, w)}{\partial x^p} \right| \leq (p R_0 \|w_i\|_1 R_1)^p$$

$$\forall p \in \{0, 1, \cdots, p_0\}, i \in [K] \colon \left| \frac{\partial^p h(x, w)}{\partial w_i^p} \right| \leq |R_0 g_i(x)|^p$$

Proof of Claim C.6. We first consider $\left| \frac{\partial^p h(x,w)}{\partial x^p} \right|$. Using Fa à di Bruno's formula, we have that

$$\frac{\partial^p h(x, w)}{\partial x^p} = \sum_{1 \cdot p_1 + 2 \cdot p_2 + \cdots + p \cdot p_p = p} \frac{p!}{p_1! p_2! \cdots p_p!} f^{(p_1 + \cdots p_p)} \left(\sum_{i \in [K]} w_i g_i(x) \right) \prod_{j=1}^{p} \left(\frac{\sum_{i \in [K]} w_i g_i^{(j)}(x)}{j!} \right)^{p_j}$$

Note that from our assumption

- $\prod_{j=1}^p \left| \left(\frac{\sum_{i \in [K]} w_i g_i^{(j)}(x)}{j!} \right)^{p_j} \right| \leq \prod_{j=1}^p (\|w\|_1 R_1)^{j p_j} = (\|w\|_1 R_1)^p.$
- $|f^{(p_1 + \cdots p_p)} \left(\sum_{i \in [K]} w_i g_i(x) \right)| \leq R_0^p$

Combining them, we have

$$\left| \frac{\partial^p h(x, w)}{\partial x^p} \right| \leq (p R_0 \|w_i\|_1 R_1)^p$$

On the other hand, consider each w_i, we also have:

$$\left| \frac{\partial^p h(x, w)}{\partial w_i^p} \right| = \left| f^{(p)} \left(\sum_{i \in [K]} w_i g_i(x) \right) (g_i(x))^p \right| \leq |R_0 g_i(x)|^p \qquad \square$$

Proof of Claim C.5. The first 4 inequalities is a direct corollary of Claim C.6.

Initially, we have a multivariate function but it suffices to check its directional first, second and third-order gradient. (For any function $g(y) \colon \mathbb{R}^m \to \mathbb{R}^n$, we can take $g(y + \alpha\delta)$ and consider $\frac{d^p g_j(y+\alpha\delta)}{d\alpha^p}$ for every coordinate j and every unit vector w.)

- In the base case, we have multivariate functions $f(\mathbf{K}_{\ell,0}) = \mathbf{K}_{\ell,0} x$ or $f(\mathbf{K}_{\ell,1}) = \mathbf{K}_{\ell,1} S_1(x)$. For each direction $\|\mathbf{\Delta}\|_F = 1$ we have $\left| \frac{d}{d\alpha^p} f(\mathbf{K}_{\ell,0} + \alpha \mathbf{\Delta}_{\ell,0}) \right| \leq \|x\|^p$ so we can take $R_1 = \|x\|$ (and for $f(\mathbf{K}_{\ell,1})$ we can take $R_1 = \|x\|^2$.)
- Whenever we compose with $\widetilde{\sigma}$ at layer ℓ, for instance calculating $h(w, y) = \widetilde{\sigma}(\sum_i w_i f_i(y))$ (when viewing all matrices as vectors), we only need to calculate $\frac{\partial^p}{\partial \alpha^p} h_j(w, y + \alpha\delta) = \frac{\partial^p}{\partial \alpha^p} \widetilde{\sigma}(\sum_i w_{j,i} f_i(y + \alpha\delta))$, so we can apply Claim C.6 and R_1 becomes $O(\overline{B}'_\ell \widetilde{\kappa}_\ell \overline{k}_\ell L) \cdot R_1$. We can do the same for the w variables, so overall for any unit (δ_x, δ_w) it satisfies $|\frac{\partial^p}{\partial \alpha^p} h_j(w + \alpha\delta_w, y + \alpha\delta_y)| \leq \big(O(\overline{B}'_\ell \widetilde{\kappa}_\ell (\overline{k}_\ell L)^2) \cdot R_1 \big)^p$.
- We also need to compose with the vanilla σ function three times:
  - once of the form $\sigma(f(\mathbf{K}_2, \ldots, \mathbf{K}_{\ell-1}))$ for calculating $\widetilde{F}_\ell(x)$,
  - once of the form $\sigma(\mathbf{W}_\ell f(\mathbf{K}_2, \ldots, \mathbf{K}_{\ell-1)})$ for calculating $\widetilde{F}_\ell(x)$, and

– once of the form $(f(\mathbf{W},\mathbf{K}) - G^\star(x))^2$ for the final squared loss.

In those calculations, although $g(x) = x^2$ does not have a bounded gradient (indeed, $\frac{d}{dx}g(x) = x$ can go to infinity when x is infinite), we know that the input x is always *bounded* by $\mathsf{poly}(\widetilde{\kappa}, \|x\|, B', G^\star(x))$. Therefore, we can also invoke Claim C.6.

Finally, we obtain the desired bounds on the first, second, and third order Lipschitzness property of $\widetilde{\mathbf{Loss}}(x;\mathbf{W},\mathbf{K})$.

For the bounds on $\widetilde{\mathbf{Loss}}(\mathcal{Z};\mathbf{W},\mathbf{K})$, we can use the absolute bounds on $\mathbf{Sum}(G^\star(x))$ and $\|x\|$ for all $x \in \mathcal{Z}$ (see Proposition C.2). □

C.6 Closeness: Empirical Truncated vs. Population Truncated Loss

Proposition C.7 (population $\leq$ empirical + ε_{s})**.** *Let P be the total number of parameters in $\{\mathbf{W}_\ell, \mathbf{K}_\ell\}_{\ell\in[L]}$. Then for every $\varepsilon_{\mathsf{s}}, \delta \geq 0$ and $\widetilde{\kappa} \geq k + L + \kappa$, as long as*

$$N = \Omega\left(\frac{P\log(d/\delta)}{\varepsilon_{\mathsf{s}}^2} \cdot \mathsf{poly}(\widetilde{\kappa}, B') \left(c_4(2^L)\log\frac{\widetilde{\kappa}B'}{\varepsilon_{\mathsf{s}}}\right)^{c_4(2^L)+O(1)}\right) \enspace ,$$

with probability at least $1-\delta$ over the choice of $\mathcal{Z}$, we have that for every $\{\mathbf{W}_\ell, \mathbf{K}_\ell\}_{\ell\in[L]}$ satisfying $\|\mathbf{W}_\ell\|_F, \|\mathbf{K}_\ell\|_F \leq \widetilde{\kappa}$, it holds:

$$\widetilde{\mathbf{Loss}}(\mathcal{D};\mathbf{W},\mathbf{K}) \leq \widetilde{\mathbf{Loss}}(\mathcal{Z};\mathbf{W},\mathbf{K}) + \varepsilon_{\mathsf{s}}$$

Proof of Proposition C.7. Observe that for every fixed $R_0 > 0$ and $R_1 > B' > 0$ (to be chosen later),

$$\mathbb{E}_{x\sim\mathcal{Z}}\left[\left(G^\star(x) - \widetilde{F}(x)\right)^2 \mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|\leq R_0, \|x\|\leq R_1}\right] \leq \mathbb{E}_{x\sim\mathcal{Z}}\left[\left(G^\star(x) - \widetilde{F}(x)\right)^2\right]$$

Moreover, each function $R(x) = \left(G^\star(x) - \widetilde{F}(x)\right)^2 \mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|\leq R_0, \|x\|\leq R_1}$ satisfies that

- boundedness: $|R(x)| \leq R_0^2$, and
- Lipschitz continuity: $R(x)$ is a $\mathsf{lip} \leq \mathsf{poly}(\widetilde{\kappa}, B', R_0, R_1, d)$-Lipschitz continuous in $(\mathbf{W},\mathbf{K})$ (by applying Claim C.5 and the fact $G^\star(x) \leq R_0 + \widetilde{F}(x) \leq \mathsf{poly}(\widetilde{\kappa}, B', R_0, R_1, d)$)

Therefore, we can take an epsilon-net on $(\mathbf{W},\mathbf{K})$ to conclude that as long as $N = \Omega\left(\frac{R_0^4 P\log(\widetilde{\kappa}B'R_1 d/(\delta\varepsilon_{\mathsf{s}}))}{\varepsilon_{\mathsf{s}}^2}\right)$, we have that w.p. at least $1-\delta$, for every $(\mathbf{W},\mathbf{K})$ within our bound (e.g. every $\|\mathbf{W}_\ell\|_2, \|\mathbf{K}_\ell\|_2 \leq \widetilde{\kappa}$), it holds:

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[\left(G^\star(x) - \widetilde{F}(x)\right)^2 \mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|\leq R_0, \|x\|\leq R_1}\right] \\
&\leq \mathbb{E}_{x\sim\mathcal{Z}}\left[\left(G^\star(x) - \widetilde{F}(x)\right)^2 \mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|\leq R_0, \|x\|\leq R_1}\right] + \varepsilon_{\mathsf{s}}/2 \leq \mathbb{E}_{x\sim\mathcal{Z}}\left[\left(G^\star(x) - \widetilde{F}(x)\right)^2\right] + \varepsilon_{\mathsf{s}}/2
\end{aligned}$$

As for the remaining terms, let us write

$$\begin{aligned}
&\left(G^\star(x) - \widetilde{F}(x)\right)^2 \mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|> R_0 \text{ or } \|x\|> R_1} \\
&\leq \left(G^\star(x) - \widetilde{F}(x)\right)^2 \mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|> R_0} + R_0^2 \cdot \mathbb{1}_{\|x\|> R_1} \\
&\leq 4\left(G^\star(x)\right)^2 \mathbb{1}_{|G^\star(x)|> R_0/2} + 4(\widetilde{F}(x))^2 \mathbb{1}_{|\widetilde{F}(x)|> R_0/2} + R_0^2 \cdot \mathbb{1}_{\|x\|> R_1}
\end{aligned}$$

- For the first term, recalling $\mathbb{E}_{x\sim\mathcal{D}}[G^\star(x) \leq B]$ so we can apply the hyper-contractivity Property 5.3 to show that, as long as $R_0 \geq \mathsf{poly}(\widetilde{\kappa})\cdot\big(c_4(2^L)\log\frac{\widetilde{\kappa}}{\varepsilon_\mathsf{s}}\big)^{c_4(2^L)}$ then it satisfies $\mathbb{E}_{x\sim\mathcal{D}}[4\left(G^\star(x)\right)^2 \mathbb{1}_{|G^\star(x)|>R_0/2}] \leq \varepsilon_\mathsf{s}/10$.
- For the second term, recall from Claim C.5 that $|\widetilde{F}(x)| \leq \mathsf{poly}(\widetilde{\kappa}, B')\cdot\sum_\ell(\|\mathbf{W}_{\ell,0}x\|^2+\|\mathbf{W}_{\ell,1}S_1(x)\|^2)$; therefore, we can write

$$\begin{aligned}
&4(\widetilde{F}(x))^2\mathbb{1}_{|\widetilde{F}(x)|>R_0/2} \\
&\leq \mathsf{poly}(\widetilde{\kappa}, B')\sum_\ell\left(\|\mathbf{W}_{\ell,0}x\|^2\mathbb{1}_{\|\mathbf{W}_{\ell,0}x\|^2>\frac{R_0}{\mathsf{poly}(\widetilde{\kappa},B')}} + \|\mathbf{W}_{\ell,1}S_1(x)\|^2\mathbb{1}_{\|\mathbf{W}_{\ell,1}S_1(x)\|^2>\frac{R_0}{\mathsf{poly}(\widetilde{\kappa},B')}}\right) .
\end{aligned}$$

  Applying the isotropic Property 5.1 and the hyper-contractivity (5.3) on $\|\mathbf{W}_{\ell,0}x\|^2$ and $\|\mathbf{W}_{\ell,1}S_1(x)\|^2$, we have as long as $R_0 \geq \mathsf{poly}(\widetilde{\kappa}, B')\cdot\big(\log\frac{\widetilde{\kappa}B'}{\varepsilon_\mathsf{s}}\big)^{\Omega(1)}$, then it satisfies

$$\mathbb{E}_{x\sim\mathcal{D}}[4(\widetilde{F}(x))^2\mathbb{1}_{|\widetilde{F}(x)|>R_0/2}] \leq \varepsilon_\mathsf{s}/10 \qquad \text{(for every } \mathbf{W},\mathbf{K} \text{ in the range)}$$

- For the third term, as long as $R_1 = d\log^{\Omega(1)}(R_0/\varepsilon_\mathsf{s})$ then we have $\mathbb{E}_{x\sim\mathcal{D}}[R_0^2\cdot\mathbb{1}_{\|x\|>R_1}] \leq \varepsilon_\mathsf{s}/10$.

Putting them together, we can choose $R_0 = \mathsf{poly}(\widetilde{\kappa}, B')\big(c_4(2^L)\log\frac{\widetilde{\kappa}B'}{\varepsilon_\mathsf{s}}\big)^{O(1)+c_4(2^L)}$ and we have

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\left(G^\star(x)-\widetilde{F}(x)\right)^2\mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|>R_0 \text{ or } \|x\|>R_1}\right] \leq \varepsilon_\mathsf{s}/2 .$$

This completes the proof that

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\left(G^\star(x)-\widetilde{F}(x)\right)^2\right] \leq \mathbb{E}_{x\sim\mathcal{Z}}\left[\left(G^\star(x)-\widetilde{F}(x)\right)^2\right] + \varepsilon_\mathsf{s} .$$

□

Proposition C.8 (empirical $\leq$ population $+\ \varepsilon_\mathsf{s}$)**.** *Let P be the total number of parameters in $\{\mathbf{W}_\ell, \mathbf{K}_\ell\}_{\ell\in[L]}$. Then for every $\varepsilon_\mathsf{s}, \delta \geq 0$ and $\widetilde{\kappa} \geq k + L + \kappa$, as long as*

$$N = \Omega\left(\frac{P\log d}{\varepsilon_\mathsf{s}^2}\cdot\mathsf{poly}(\widetilde{\kappa}, B')\left(c_4(2^L)\log\frac{\widetilde{\kappa}B'}{\delta\varepsilon_\mathsf{s}}\right)^{c_4(2^L)+O(1)}\right) ,$$

for any **fixed** *$\{\mathbf{W}_{\ell,0}, \mathbf{W}_{\ell,1}\}_{\ell\in[L]}$, with probability at least $1-\delta$ over the choice of $\mathcal{Z}$, we have that for* every *$\{\mathbf{W}_\ell, \mathbf{K}_\ell\}_{\ell\in[L]}$ satisfying (1) $\|\mathbf{W}_\ell\|_F, \|\mathbf{K}_\ell\|_F \leq \widetilde{\kappa}$ and (2) consistent with $\{\mathbf{W}_{\ell,0}, \mathbf{W}_{\ell,1}\}_{\ell\in[L]}$, it holds:*

$$\mathbb{E}_{x\sim\mathcal{Z}}[\widetilde{\mathbf{Loss}}(x;\mathbf{W},\mathbf{K})] = \mathbb{E}_{x\sim\mathcal{Z}}\left[\left(G^\star(x)-\widetilde{F}(x)\right)^2\right] \leq \mathbb{E}_{x\sim\mathcal{D}}\left[\left(G^\star(x)-\widetilde{F}(x)\right)^2\right] + \varepsilon_\mathsf{s} = \mathbb{E}_{x\sim\mathcal{D}}[\widetilde{\mathbf{Loss}}(x;\mathbf{W},\mathbf{K})] + \varepsilon_\mathsf{s}$$

Proof. We first reverse the argument of Proposition C.7 and have that as long as $N = \Omega\left(\frac{R_0^4 P\log(\widetilde{\kappa}B'R_1 d/(\delta\varepsilon_\mathsf{s}))}{\varepsilon_\mathsf{s}^2}\right)$, we have that w.p. at least $1-\delta/2$, for every $(\mathbf{W},\mathbf{K})$ within our bound (e.g. every $\|\mathbf{W}_\ell\|_2, \|\mathbf{K}_\ell\|_2 \leq \widetilde{\kappa}$), it holds:

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{Z}}\left[\left(G^\star(x)-\widetilde{F}(x)\right)^2\mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|\leq R_0, \|x\|\leq R_1}\right] \\
&\leq \mathbb{E}_{x\sim\mathcal{D}}\left[\left(G^\star(x)-\widetilde{F}(x)\right)^2\mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|\leq R_0, \|x\|\leq R_1}\right] + \varepsilon_\mathsf{s}/2 \leq \mathbb{E}_{x\sim\mathcal{D}}\left[\left(G^\star(x)-\widetilde{F}(x)\right)^2\right] + \varepsilon_\mathsf{s}/2
\end{aligned}$$

As for the remaining terms, we again write

$$\begin{aligned}
&\left(G^\star(x) - \widetilde{F}(x)\right)^2 \mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|>R_0 \text{ or } \|x\|>R_1} \\
&\leq 4\left(G^\star(x)\right)^2 \mathbb{1}_{|G^\star(x)|>R_0/2} + R_0^2 \cdot \mathbb{1}_{\|x\|>R_1} \\
&\quad + \mathsf{poly}(\widetilde{\kappa}, B') \sum_\ell \left(\|\mathbf{W}_{\ell,0} x\|^2 \mathbb{1}_{\|\mathbf{W}_{\ell,0}x\|^2 > \frac{R_0}{\mathsf{poly}(\widetilde{\kappa}, B')}} + \|\mathbf{W}_{\ell,1} S_1(x)\|^2 \mathbb{1}_{\|\mathbf{W}_{\ell,1} S_1(x)\|^2 > \frac{R_0}{\mathsf{poly}(\widetilde{\kappa}, B')}} \right) := RHS
\end{aligned}$$

For this right hand side RHS, we notice that it does not depend on $\mathbf{K}$. The identical proof of Proposition C.7 in fact proves that if $R_0 = \mathsf{poly}(\widetilde{\kappa}, B')\big(c_4(2^L) \log \frac{\widetilde{\kappa} B'}{\delta \varepsilon_{\mathsf{s}}}\big)^{O(1)+c_4(2^L)}$ then for every $\mathbf{W}$ with $\|\mathbf{K}_\ell\|_2 \leq \widetilde{\kappa}$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[RHS] \leq \delta \varepsilon_{\mathsf{s}}/4 \ .$$

This means, by Markov bound, for the given *fixed* $\mathbf{W}$, with probability at least $1-\delta/2$ over the randomness of $\mathcal{Z}$, it satisfies

$$\mathop{\mathbb{E}}_{x\sim\mathcal{Z}}[RHS] \leq \varepsilon_{\mathsf{s}}/2 \ .$$

This implies for every $\mathbf{K}$ in the given range,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{Z}}\left[\left(G^\star(x) - \widetilde{F}(x)\right)^2 \mathbb{1}_{|G^\star(x)-\widetilde{F}(x)|>R_0 \text{ or } \|x\|>R_1}\right] \leq \varepsilon_{\mathsf{s}}/2 \ . \qquad \square$$

D An Implicit Implication of Our Distribution Assumption

Let us define

$$\begin{aligned}
\widehat{S}^\star_0(x) &= x \\
\widehat{S}^\star_1(x) &= \sigma(x) \\
\widehat{S}^\star_2(x) &= \mathbf{W}^\star_{2,1}\widehat{S}^\star_1(x) = \mathbf{W}^\star_{2,1}\sigma(x) \\
\widehat{S}^\star_\ell(x) &= \mathbf{W}^\star_{\ell,\ell-1}\sigma\left(\widehat{S}^\star_{\ell-1}(x)\right) \quad \text{for } \ell = 2,\dots,L
\end{aligned}$$

so that $\widehat{S}^\star_\ell(x)$ is the top-degree (i.e. degree $2^{\ell-1}$) part of $S^\star_\ell(x)$.[24] We have the following implication:

Lemma D.1 (Implication of singular-value preserving)**.** *Let us define*

$$z^0 = z^0(x) = \widehat{S}^\star_0(x) = x \tag{D.1}$$

$$z^1 = z^1(x) = \widehat{S}^\star_1(x) = \sigma(x) \tag{D.2}$$

$$z^\ell = z^\ell(x) = \widehat{S}^\star_\ell(x) * \widehat{S}^\star_\ell(x) \tag{D.3}$$

Then, for every $\ell \geq \ell_1, \ell_2 \geq 0$ with $|\ell_1 - \ell_2| \neq 1$, for every matrix $\mathbf{M}$: and the associated homogeneous polynomial $g_{\mathbf{M}}(x) = (z^{\ell_1})^\top \mathbf{M} z^{\ell_2}$,

- *If $\ell_1 = \ell_2 = \ell = 0$ or 1, then $\mathcal{C}_x(g_{\mathbf{M}}) = \|\mathbf{M}\|_F^2$,*
- *If $\ell_1 = \ell_2 = \ell \geq 2$, then $\mathcal{C}_x(g_{\mathbf{M}}) \geq \frac{1}{(\kappa 2^\ell)^{O(2^\ell)}}\|\mathbf{Sym}(\mathbf{M})\|_F^2$, and*
- *If $\ell_1 - 2 \geq \ell_2 \geq 0$, then $\mathcal{C}_x(g_{\mathbf{M}}) \geq \frac{1}{(\kappa 2^\ell)^{O(2^\ell)}}\|\mathbf{M}\|_F^2$ for $\ell = \ell_1$.*

[24] Meaning that $\widehat{S}^\star_\ell(x)$ is a (vector) of homogenous polynomials of x with degree $2^{\ell-1}$, and its coefficients coincide with $S^\star_\ell(x)$ on those monomials.

D.1 Proof of Lemma D.1

Proof of Lemma D.1. We divide the proof into several cases.

Case A: When $\ell_1 = \ell_2 = \ell$. The situation for $\ell = 0$ or $\ell = 1$ is obvious, so below we consider $\ell \geq 2$. Let $h_\ell(z) = (z * z)\mathbf{M}(z * z) = \sum_{i\leq j, k\leq l} \mathbf{M}_{\{i,j\},\{k,l\}} a_{i,j} a_{k,l} z_i z_j z_k z_l$ be the degree-4 polynomial defined by $\mathbf{M}$. We have

$$\mathcal{C}_z(h_\ell) \geq \|\mathbf{Sym}(\mathbf{M})\|_F^2$$

For every for every $j = \ell - 1, \dots, 1$, we define $h_j(z) = h_{j+1}(\mathbf{W}^\star_{j+1,j}\sigma(z))$, it holds that

Let $\widetilde{h}(z) = h_{j+1}(\mathbf{W}^\star_{j+1,j} z)$ so that $h_j(z) = \widetilde{h}(\sigma(z))$. This means

$$\mathcal{C}(h_j) = \mathcal{C}(\widetilde{h}) \geq \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \mathcal{C}(h_{j+1})$$

and finally we have $(z^\ell)^\top \mathbf{M} z^\ell = h_1(x)$ and therefore

$$\mathcal{C}_x\left((z^\ell)^\top \mathbf{M} z^\ell\right) \geq \frac{1}{(\kappa 2^\ell)^{O(2^\ell)}} \|\mathbf{Sym}(\mathbf{M})\|_F^2$$

Case B: When $\ell_1 - 1 > \ell_2 \geq 2$. We define $h_{\ell_1}(z, y) = (z * z)^\top \mathbf{M}(y * y)$ which is a degree-4 homogenous polynomial in (z, y), and obviously $\mathcal{C}_{y,z}(h_{\ell_1}) \geq \|\mathbf{M}\|_F^2$. Let us define

$$\forall j = \ell_1 - 1, \dots, \ell_2 + 2: \qquad h_j(z, y) = h_{j+1}\left((\mathbf{W}^\star_{j+1,j}\sigma(z), y)\right)$$

By the same argument as before, we have

$$\mathcal{C}_{z,y}(h_j) \geq \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \mathcal{C}_{z,y}(h_{j+1})$$

Next, for $j = \ell_2$, we define

$$h_j(y) = h_{j+2}\left(\mathbf{W}^\star_{j+2,j+1}\sigma(\mathbf{W}^\star_{j+1,j}\sigma(y)), y\right)$$

To analyze this, we first define

$$h'(z, y) = h_{j+2}\left(\mathbf{W}^\star_{j+2,j+1} z, y\right) \quad \text{so that} \quad h_j(y) = h'\left(\sigma(\mathbf{W}^\star_{j+1,j}\sigma(y)), y\right)$$

Since $h'(z, y)$ is of degree 2 in the variables from y, we can write it as

$$h'(z, y) = \underbrace{\sum_p (y_p)^2 h''_{\{p,p\}}(z)}_{h''_\perp(z, \sigma(y)))} + \sum_{p<q} y_p y_q h''_{\{p,q\}}(z) \tag{D.4}$$

where the first summation contains only those quadratic terms in $(y_p)^2$ and the second contain cross terms $y_p y_q$. Note in particular if we write the first summation as $h''_\perp(z, \sigma(y)))$ for polynomial $h''_\perp(z, \gamma)$ and $\gamma = \sigma(y)$, then $h''_\perp$ is *linear* in γ. Clearly,

$$\mathcal{C}_{z,y}(h') = \mathcal{C}_{z,\gamma}(h''_\perp) + \sum_{p<q} \mathcal{C}_z(h''_{\{p,q\}}) \tag{D.5}$$

As a consequence, we can write

$$h_j(y) = \underbrace{h''_\perp(\sigma(\mathbf{W}^\star_{j+1,j}\sigma(y)), \sigma(y))}_{\widetilde{h}_\perp(y)} + y_p y_q \cdot \underbrace{h''_{\{p,q\}}(\sigma(\mathbf{W}^\star_{j+1,j}\sigma(y)))}_{\widetilde{h}_{\{p,q\}}(y)}$$

Clearly, since any polynomial in $\sigma(y)$ only contain even degrees of variables in y, so $\widetilde{h}_\perp(y)$ and each

$\widetilde{h}_{\{p,q\}}$ share no common monomial, we have

$$\mathcal{C}_y(h_j) = \mathcal{C}_y(\widetilde{h}_\perp) + \sum_{p<q} \mathcal{C}_y(\widetilde{h}_{\{p,q\}}) \tag{D.6}$$

- On one hand, we have $\widetilde{h}_{\{p,q\}}(y) = h''_{\{p,q\}}(\sigma(\mathbf{W}^\star_{j+1,j}\sigma(y)))$ and therefore by previous argument

$$\mathcal{C}_y(\widetilde{h}_{\{p,q\}}) \geq \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \mathcal{C}_z(h''_{\{p,q\}}) \tag{D.7}$$

- On the other hand, to analyze $\widetilde{h}_\perp(y)$, let us construct a square matrix $\mathbf{W} \in \mathbb{R}^{k_j \times k_j}$ with singular values between $[1/\kappa, \kappa]$ so that

$$\mathbf{W}^\star_{j+1,j}\mathbf{W} = (\mathbf{I}_{k_{j+1}\times k_{j+1}}, 0) \tag{D.8}$$

Define $h'''_\perp(z,\beta) = h''_\perp(z, \mathbf{W}\beta)$ which is linear in β, it holds:[25]

$$\begin{aligned}
\mathcal{C}_y\left(\widetilde{h}_\perp(y)\right) &= \mathcal{C}_y\left(h''_\perp(\sigma(\mathbf{W}^\star_{j+1,j}\sigma(y), \sigma(y)))\right) \\
&= \mathcal{C}_y\left(h''_\perp(\sigma(\mathbf{W}^\star_{j+1,j}y), y)\right) \\
&\geq \mathcal{C}_\beta\left(h''_\perp\left(\sigma(\mathbf{W}^\star_{j+1,j}\mathbf{W}\beta), \mathbf{W}\beta\right)\right) \cdot \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \\
&= \mathcal{C}_\beta\left(h''_\perp\left(\sigma((\mathbf{I},0)\beta), \mathbf{W}\beta\right)\right) \cdot \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \\
&= \mathcal{C}_\beta\left(h'''_\perp\left(\sigma((\mathbf{I},0)\beta), \beta\right)\right) \cdot \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \\
&\overset{①}{=} \mathcal{C}_{z,\beta}\left(h'''_\perp(z,\beta)\right) \cdot \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \\
&\geq \mathcal{C}_{z,\gamma}\left(h'''_\perp(z,\mathbf{W}^{-1}\gamma)\right) \cdot \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \\
&= \mathcal{C}_{z,\gamma}\left(h''_\perp(z,\gamma)\right) \cdot \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}}
\end{aligned} \tag{D.9}$$

Finally, plugging the lower bounds (D.7) and (D.9) into expansions (D.5) and (D.6), we conclude that

$$C_y(h_j) \geq \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \cdot C_{z,y}(h') \geq \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \cdot C_{z,y}(h_{j+2})$$

Continuing from here, we can define $h_j(y) = h_{j+1}(\mathbf{W}^\star_{j+1,j}\sigma(y))$ for every $j = \ell_2 - 1, \ell_2 - 2, \dots, 1$ and using the same analysis as Case A, we have

$$\mathcal{C}(h_j) = \mathcal{C}(\widetilde{h}) \geq \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \mathcal{C}(h_{j+1})$$

[25] Above, equality ① holds because $h'''_\perp(z,\beta)$ is a multi-variate polynomial which is linear in β, so it can be written as

$$h'''_\perp(z,\beta) = \sum_i \beta_i \cdot h'''_{\perp,i}(z)$$

for each $h'''_{\perp,i}(z)$ being a polynomial in z; next, since we plug in $z = \sigma((\mathbf{I},0)\beta)$ which only contains even-degree variables in β, we have

$$\mathcal{C}_\beta\left(h'''_\perp\left(\sigma((\mathbf{I},0)\beta), \beta\right)\right) = \sum_i \mathcal{C}_\beta\left(h'''_{\perp,i}\left(\sigma((\mathbf{I},0)\beta)\right)\right) = \sum_i \mathcal{C}_z\left(h'''_{\perp,i}(z)\right) = \mathcal{C}_{z,\gamma}\left(h'''_\perp(z,\gamma)\right)$$

and finally we have $(z^\ell)^\top \mathbf{M} z^\ell = h_1(x)$ and therefore

$$\mathcal{C}_x\left((z^\ell)^\top \mathbf{M} z^\ell\right) \geq \frac{1}{(\kappa 2^\ell)^{O(2^\ell)}}\|\mathbf{M}\|_F^2$$

$\square$

Case C: When $\ell_1 - 1 > \ell_2 = 1$. Similar to Case B, we can $h_{\ell_1}(z,y) = (z * z)^\top \mathbf{M}\sigma(y)$ which is a degree-4 homogenous polynomial in (z,y), and obviously $\mathcal{C}_{y,z}(h_{\ell_1}) \geq \|\mathbf{M}\|_F^2$. Let us define

$$\forall j = \ell_1 - 1, \dots, 3\colon \qquad h_j(z,y) = h_{j+1}\left((\mathbf{W}^\star_{j+1,j}\sigma(z), y)\right)$$
$$h_1(y) = h_3\left(\mathbf{W}^\star_{3,2}\sigma(\mathbf{W}^\star_{2,1}\sigma(y)), y\right)$$

The rest of the proof now becomes identical to Case B. (In fact, we no longer have cross terms in (D.4) so the proof only becomes simpler.)

Case D: When $\ell_1 - 1 > \ell_2 = 0$. We define $h_{\ell_1}(z,y) = (z * z)^\top \mathbf{M} y$ which is a degree-3 homogenous polynomial in (z,y), and obviously $\mathcal{C}_{y,z}(h_{\ell_1}) \geq \|\mathbf{M}\|_F^2$. Let us define

$$\forall j = \ell_1 - 1, \dots, 2\colon \qquad h_j(z,y) = h_{j+1}\left((\mathbf{W}^\star_{j+1,j}\sigma(z), y)\right)$$
$$h_1(y) = h_2\left(\mathbf{W}^\star_{2,1}\sigma(y), y\right)$$

By defining $h'(z,y) = h_2(\mathbf{W}^\star_{2,1} z, y)$ we have $h_1(y) = h'(\sigma(y), y)$. This time, we have $\mathcal{C}_y(h_1) = \mathcal{C}_{z,y}(h')$, but the same proof of Case B tells us $\mathcal{C}_{z,y}(h') \geq \frac{1}{(\kappa 2^\ell)^{O(2^{\ell-j})}} \cdot \|\mathbf{M}\|_F^2$.

E Critical Lemma For Implicit Hierarchical Learning

The implicit hierarchical learning only requires one Lemma, which can be stated as the following:

Lemma E.1. *There exists absolute constant $c_0 \geq 2$ so that the following holds. Let $\tau_\ell \geq \overline{k}_\ell + L + \kappa$ and $\Upsilon_\ell \geq 1$ be arbitrary parameters for each layer $\ell \leq L$. Define parameters*

$$D_\ell \stackrel{\text{def}}{=} \left(\tau_\ell \cdot \kappa^{2^\ell} \cdot (2^\ell)^{2^\ell} \cdot c_1(2^\ell) \cdot c_3(2^\ell)\right)^{c_0 \ell}$$
$$C_\ell \stackrel{\text{def}}{=} C_{\ell-1} \cdot 2\Upsilon_\ell^3 (D_\ell)^{17} \qquad \text{with } C_2 = 1$$

Suppose $\mathbf{Obj}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \leq \varepsilon^2$ for some $0 \leq \varepsilon \leq \frac{\alpha_L}{(D_L)^9 \Upsilon_L}$ and suppose the parameters satisfy

- $\frac{\alpha_{\ell+1}}{\alpha_\ell} \leq \frac{1}{C_{\ell+1}}$ *for every $\ell = 2, 3, \dots, L-1$*
- $\mathbb{E}_{x\sim\mathcal{D}}[\|S_\ell(x)\|^2] \leq \tau_\ell$ *for every $\ell = 2, 3, \dots, L-1$*
- $\lambda_{6,\ell} \geq \frac{\varepsilon^2}{\tau_\ell^2}$, $\lambda_{3,\ell} \geq \frac{\alpha_\ell^2}{D_\ell \cdot \Upsilon_\ell}$, $\lambda_{4,\ell} \geq \frac{\alpha_\ell^2}{(D_\ell)^7 \Upsilon_\ell^2}$, $\lambda_{5,\ell} \geq \frac{\alpha_\ell^2}{(D_\ell)^{13}\Upsilon_\ell^3}$ *for every $\ell = 2, 3, \dots, L$*

Then, there exist unitary matrices $\mathbf{U}_\ell$ such that for every $\ell = 2, 3, \dots, L$

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \|\mathbf{U}_\ell S^\star_\ell(x) - S_\ell(x)\|_2^2 \leq \left(\frac{\varepsilon}{\sqrt{\alpha_{\ell+1}\alpha_\ell}}\right)^2 C_L$$

Since we shall prove Corollary E.1 by induction, we have stated only one of the main conclusions in order for the induction to go through. Once the Theorem E.1 is proved, in fact we can strengthen it as follows.

Definition E.2. *For each $\ell \geq 2$, let $\mathbf{Q}_\ell$ be the unitary matrix defined from Lemma B.5 satisfying*

$$\mathbf{R}_\ell \mathbf{U}_\ell * \mathbf{R}_\ell \mathbf{U}_\ell = (\mathbf{R}_\ell * \mathbf{R}_\ell)\mathbf{Q}_\ell$$

We also let $\mathbf{Q}_0 = \mathbf{Q}_1 = \mathbf{I}_{d\times d}$*, and let*

$$\mathbf{Q}_{\ell\triangleleft} \stackrel{\text{def}}{=} \mathbf{diag}(\mathbf{Q}_j)_{\mathcal{J}_\ell} \quad \textit{and} \quad \vec{\mathbf{Q}}_\ell \stackrel{\text{def}}{=} \mathbf{diag}(\mathbf{Q}_j)_{j\in\mathcal{J}_\ell}$$

Corollary E.3. *Under the same setting as Theorem E.1, we actually have for all* $\ell = 2, 3, \dots, L$,

(a) $\left\| \mathbf{Q}_{\ell-1}^\top \overline{\mathbf{W}}_{\ell,\ell-1}^\top \overline{\mathbf{W}}_{\ell\triangleleft} \mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{\ell,\ell-1}^\top \overline{\mathbf{W}^\star}_{\ell\triangleleft} \right\|_F^2 \leq (D_\ell)^2 \left(\frac{\varepsilon}{\alpha_\ell}\right)^2 \cdot \frac{C_L}{C_\ell}$

(b) $\left\| \mathbf{Q}_{\ell-1}^\top \overline{\mathbf{K}}_{\ell,\ell-1}^\top \overline{\mathbf{K}}_{\ell\triangleleft} \mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{\ell,\ell-1}^\top \overline{\mathbf{W}^\star}_{\ell\triangleleft} \right\|_F^2 \leq \Upsilon_\ell (D_\ell)^4 \left(\frac{\varepsilon}{\alpha_\ell}\right)^2 \cdot \frac{C_L}{C_\ell}$

(c) $\left\| \vec{\mathbf{Q}}_\ell^\top \overline{\mathbf{K}}_\ell^\top \overline{\mathbf{K}}_\ell \vec{\mathbf{Q}}_\ell - \overline{\mathbf{W}^\star}_\ell^\top \overline{\mathbf{W}^\star}_\ell \right\|_F^2 \leq \Upsilon_\ell^2 (D_\ell)^{14} \left(\frac{\varepsilon}{\alpha_\ell}\right)^2 \cdot \frac{C_L}{C_\ell}$

(d) $\mathbb{E}_{x\sim\mathcal{D}} \|\mathbf{U}_\ell S_\ell^\star(x) - S_\ell(x)\|_2^2 \leq 2\Upsilon_\ell^2 (D_\ell)^{17} \left(\frac{\varepsilon}{\alpha_\ell}\right)^2 \cdot \frac{C_L}{C_\ell}$

(e) $\mathbb{E}_{x\sim\mathcal{D}}[\|S_\ell(x)\|^2] \leq 2B_\ell$.

Corollary E.4. *Suppose we only have* $\varepsilon \leq \frac{\alpha_L}{(D_L)^3\sqrt{\Upsilon_L}}$*, which is a weaker requirement comparing to Theorem E.1. Then, Theorem E.1 and Corollary E.3 still hold for the first* $L-1$ *layers but for* ε *replaced with* $\alpha_L \cdot \sqrt{D_L}$*. In addition, for* $\ell = L$*, we have*

(a) $\left\| \mathbf{Q}_{L-1}^\top \overline{\mathbf{W}}_{L,L-1}^\top \overline{\mathbf{W}}_{L\triangleleft} \mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L\triangleleft} \right\|_F^2 \leq 2(D_L)^2 \left(\frac{\varepsilon}{\alpha_L}\right)^2$

(b) $\left\| \mathbf{Q}_{L-1}^\top \overline{\mathbf{K}}_{L,L-1}^\top \overline{\mathbf{K}}_{L\triangleleft} \mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L\triangleleft} \right\|_F^2 \leq 2\Upsilon_L (D_L)^4 \left(\frac{\varepsilon}{\alpha_L}\right)^2$

(c) $\left\| \vec{\mathbf{Q}}_L^\top \overline{\mathbf{K}}_L^\top \overline{\mathbf{K}}_L \vec{\mathbf{Q}}_L - \overline{\mathbf{W}^\star}_L^\top \overline{\mathbf{W}^\star}_L \right\|_F^2 \leq 2\Upsilon_L^2 (D_L)^{14} \left(\frac{\varepsilon}{\alpha_L}\right)^2$

E.1 Base Case

The base case is $L = 2$. In this case, the loss function

$$\varepsilon^2 \geq \mathbf{Obj}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \geq \alpha_2^2 \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left(\|\mathbf{W}_{2,1} S_1(x)\|^2 - \|\mathbf{W}_{2,1}^\star S_1(x)\|^2\right)^2$$

Applying the degree-preservation Property 5.2, we have

$$\mathcal{C}_x\left(\|\mathbf{W}_{2,1}\widehat{S}_1(x)\|^2 - \|\mathbf{W}_{2,1}^\star \widehat{S}_1(x)\|^2\right) \leq O(1)\left(\frac{\varepsilon}{\alpha_2}\right)^2$$

where recall from Section D that $\widehat{S}_1(x) = \sigma(x)$ is the top-degree homogeneous part of $S_1(x)$, and $\mathcal{C}_x(f(x))$ is the sum of squares of f's monomial coefficients. Applying Lemma D.1, we know

$$\|\mathbf{W}_{2,1}^\top \mathbf{W}_{2,1} - (\mathbf{W}_{2,1}^\star)^\top \mathbf{W}_{2,1}^\star\|_F^2 \leq O(1)\left(\frac{\varepsilon}{\alpha_2}\right)^2$$

On the other hand, our regularizer $\lambda_{4,L}$ ensures that

$$\left\|\mathbf{W}_{2,1}^\top \mathbf{W}_{2,1} - \mathbf{K}_{2,1}^\top \mathbf{K}_{2,1}\right\|_F^2 \leq \frac{\varepsilon^2}{\lambda_{4,2}} \leq (D_L)^7 \Upsilon_L^2 \left(\frac{\varepsilon}{\alpha_2}\right)^2$$

Putting them together we have

$$\left\|(\mathbf{W^\star}_{2,1})^\top \mathbf{W^\star}_{2,1} - \mathbf{K}_{2,1}^\top \mathbf{K}_{2,1}\right\|_F^2 \leq \frac{\varepsilon^2}{\lambda_{4,2}} \leq (D_L)^7 \Upsilon_L^2 \left(\frac{\varepsilon}{\alpha_2}\right)^2$$

By putting it into SVD decomposition, it is easy to derive the existence of some unitary matrix $\mathbf{U}_2$ satisfying (for a proof see Claim I.10)

$$\|\mathbf{U}_2\mathbf{K}_{2,1} - \mathbf{W}^\star_{2,1}\|_F^2 \le (D_L)^8 \Upsilon_L^2 \left(\frac{\varepsilon}{\alpha_2}\right)^2$$

Right multiplying it to $S_1(x)$, we have (using the isotropic Property 5.1)

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{D}}\|\mathbf{U}_2 S_2(x) - S^\star_2(x)\|_F^2 &= \mathbb{E}_{x\sim\mathcal{D}}\|\mathbf{U}_2\mathbf{K}_{2,1}S_1(x) - \mathbf{W}^\star_{2,1}S_1(x)\|_F^2 \\ &\le O(1)\cdot (D_L)^8 \Upsilon_L^2 \left(\frac{\varepsilon}{\alpha_2}\right)^2 \ll \left(\frac{\varepsilon}{\sqrt{\alpha_3\alpha_2}}\right)^2\end{aligned}$$

E.2 Preparing to Prove Theorem E.1

Let us do the proof by induction with the number of layers L. Suppose this Lemma is true for every $L \le L_0$, then let us consider $L = L_0 + 1$ Define

$$G^\star_{\le L-1}(x) = \textstyle\sum_{\ell=2}^{L-1} \alpha_\ell \mathbf{Sum}(G^\star_\ell(x))$$
$$F_{\le L-1}(x) = \textstyle\sum_{\ell=2}^{L-1} \alpha_\ell \mathbf{Sum}(F_\ell(x))$$

We know that the objective of the first $L-1$ layers

$$\begin{aligned}\mathbf{Loss}_{L-1}(\mathcal{D}) + \mathbf{Reg}_{L-1} &= \mathbb{E}_{x\sim\mathcal{D}}\left(G^\star_{\le L-1}(x) - F_{\le L-1}(x)\right)^2 + \mathbf{Reg}_{L-1} \\ &\le 2\,\mathbb{E}_{x\sim\mathcal{D}}\left(G^\star(x) - F(x)\right)^2 + 2\alpha_L^2\,\mathbb{E}_{x\sim\mathcal{D}}\left(\mathbf{Sum}(F_L(x)) - \mathbf{Sum}(G^\star_L(x))\right)^2 + \mathbf{Reg}_L \\ &\le 2\alpha_L^2\,\mathbb{E}_{x\sim\mathcal{D}}\left(\mathbf{Sum}(F_L(x)) - \mathbf{Sum}(G^\star_L(x))\right)^2 + 2\mathbf{Loss}(\mathcal{D}) + \mathbf{Reg}\ . \qquad \text{(E.1)}\end{aligned}$$

By our assumption on the network $G^\star$, we know that for every $\ell \in [L]$,

$$\mathbb{E}_{x\sim\mathcal{D}}[\mathbf{Sum}(G^\star_\ell(x))] \le B_\ell \iff \mathbb{E}_{x\sim\mathcal{D}}[\|S^\star_\ell(x)\|^2] \le B_\ell$$

By hyper-contractivity assumption (5.4), we have that

$$\mathbb{E}_{x\sim\mathcal{D}}[(\mathbf{Sum}(G^\star_\ell(x))^2] \le c_3(2^\ell)\cdot B_\ell^2 \iff \mathbb{E}_{x\sim\mathcal{D}}[\|S^\star_\ell(x)\|^4] \le c_3(2^\ell)\cdot B_\ell^2 \qquad \text{(E.2)}$$

Using our assumption $\mathbb{E}_{x\sim\mathcal{D}}[\|S_\ell(x)\|^2] \le \tau_\ell$ and the hyper-contractivity Property 5.4 we also have

$$\mathbb{E}_{x\sim\mathcal{D}}[\mathbf{Sum}(F_\ell(x))] \le c_3(2^\ell)(k_\ell L\tau_\ell)^4 \quad \text{and} \quad \mathbb{E}_{x\sim\mathcal{D}}[\mathbf{Sum}(F_\ell(x))^2] \le c_3(2^\ell)(k_\ell L\tau_\ell)^8$$

Putting these into (E.1) we have

$$\mathbf{Obj}_{L-1} \le \alpha_L^2\cdot(k_L L\overline{B}_L\tau_L)^8 c_3(2^L) + 2\varepsilon^2 \qquad \text{(E.3)}$$

By induction hypothesis[26] for every L replaced with $L-1$, there exist unitary matrices $\mathbf{U}_\ell$ such that

$$\forall \ell = 2,3,\dots,L-1: \quad \mathbb{E}_{x\sim\mathcal{D}}\|\mathbf{U}_\ell S^\star_\ell(x) - S_\ell(x)\|_2^2 \le \delta_\ell^2 \stackrel{\text{def}}{=} \left(\frac{\alpha_L}{\sqrt{\alpha_\ell\alpha_{\ell+1}}}\right)^2 C_{L-1}\cdot(k_L L\overline{B}_L\tau_L)^8 c_3(2^L) \ll 1 \qquad \text{(E.4)}$$

Let $\widehat{S}_\ell(x), \widehat{S}^\star_\ell(x)$ be the degree $2^{\ell-1}$ homogeneous part of $S_\ell(x), S^\star_\ell(x)$ respectively, notice that $\|\mathbf{U}_\ell S^\star_\ell(x) - S_\ell(x)\|_2^2$ is a polynomial of maximum degree $2^{\ell-1}$, therefore, using the degree-preservation

[26] To be precise, using our assumption on $\frac{\alpha_L}{\alpha_{L-1}}$ one can verify that $O\big(\alpha_L^2\cdot(k_L L\overline{B}_L\tau_L)^8 c_3(2^\ell)\big) \le \frac{\alpha_{L-1}^2}{2(D_{L-1})^8\sqrt{\Upsilon_{L-1}^3}}$ so the assumption from the inductive case holds.

Property 5.2, we know that

$$\forall \ell = 2,3,\dots,L-1: \qquad \sum_{i\in[k_\ell]} \mathcal{C}_x\left(\left[\mathbf{U}_\ell \widehat{S}^\star_\ell(x) - \widehat{S}_\ell(x)\right]_i\right) \le c_1(2^\ell)\cdot \delta_\ell^2 \tag{E.5}$$

$$\forall \ell = 2,3,\dots,L: \qquad \sum_{i\in[k_\ell]} \mathcal{C}_x\left(\left[\widehat{S}^\star_\ell(x)\right]_i\right) \le c_1(2^\ell)\cdot B_\ell$$

We begin by proof by grouping the 2^L-degree polynomials $G^\star(x)$ and $F(x)$, into monomials of different degrees. Since

$$G^\star(x) = \sum_{\ell=2}^{L} \alpha_\ell \mathbf{Sum}(G^\star_\ell(x)) \text{ and } F(x) = \sum_{\ell=2}^{L} \alpha_\ell \mathbf{Sum}(F(x)),$$

it is clear that all the monomials with degree between $2^{L-1}+1$ and 2^L are only present in the terms $\mathbf{Sum}(G^\star_L(x))$ and $\mathbf{Sum}(F_L(x))$ respectively. Recall also (we assume L is even for the rest of the proof, and the odd case is analogous).

$$\mathbf{Sum}(G^\star_L(x)) = \left\|\sum_{\ell\in\mathcal{J}_L\setminus\{0,1\}} \mathbf{W}^\star_{L,\ell}\sigma(S^\star_\ell(x)) + \sum_{\ell\in\mathcal{J}_L\cap\{0,1\}} \mathbf{W}^\star_{L,\ell} S^\star_\ell(x)\right\|^2 \tag{E.6}$$

$$\mathbf{Sum}(F_L(x)) = \left\|\sum_{\ell\in\mathcal{J}_L\setminus\{0,1\}} \mathbf{W}_{L,\ell}\sigma(\mathbf{R}_\ell S_\ell(x)) + \sum_{\ell\in\mathcal{J}_L\cap\{0,1\}} \mathbf{W}_{L,\ell} S_\ell(x)\right\|^2$$

E.3 Degree 2^L

We first consider all the monomials from $G^\star(x)$ and $F(x)$ in degree $2^{L-1}+2^{L-1}=2^L$ (i.e., top degree). As argued above, they must come from the top degree of (E.6).

Let $\widehat{G^\star}_L, \widehat{F}_L\colon \mathbb{R}^d \to \mathbb{R}^{k_L}$ be the degree 2^L part of $G^\star_L(x), F_L(x)$ respectively. Using

$$\mathbb{E}_{x\sim\mathcal{D}} |F(x) - G^\star(x)|^2 \le \mathbf{Obj} \le \varepsilon^2$$

and the degree-preservation Property 5.2 again, we have

$$\mathcal{C}_x\left(\mathbf{Sum}(\widehat{F}_L(x)) - \mathbf{Sum}(\widehat{G^\star}_L(x))\right) \le c_1(2^L)\left(\frac{\varepsilon}{\alpha_L}\right)^2 \tag{E.7}$$

From (E.6), we know that

$$\mathbf{Sum}(\widehat{G^\star}_L(x)) = \left\|\mathbf{W}^\star_{L,L-1}\sigma\left(\widehat{S}^\star_{L-1}(x)\right)\right\|^2 = \left\|\overline{\mathbf{W}^\star}_{L,L-1}\left(\widehat{S}^\star_{L-1}(x) * \widehat{S}^\star_{L-1}(x)\right)\right\|^2$$

We also have

$$\mathbf{Sum}(\widehat{F}_L(x)) = \left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\widehat{S}_{L-1}(x)\right)\right\|^2 = \left\|\overline{\mathbf{W}}_{L,L-1}\left(\widehat{S}_{L-1}(x) * \widehat{S}_{L-1}(x)\right)\right\|^2$$

For analysis, we also define $\overline{\overline{\mathbf{W}}}_{L,L-1} = \mathbf{W}_{L,L-1}(\mathbf{R}_{L-1}\mathbf{U}_{L-1} * \mathbf{R}_{L-1}\mathbf{U}_{L-1}) \in \mathbb{R}^{k_L \times \binom{k_{L-1}+1}{2}}$ so that

$$\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\mathbf{U}_{L-1}\widehat{S}^\star_{L-1}(x)\right) = \overline{\overline{\mathbf{W}}}_{L,L-1}\left(\widehat{S}^\star_{L-1}(x) * \widehat{S}^\star_{L-1}(x)\right)$$

where $\overline{\overline{\mathbf{W}}}_{L,L-1} = \overline{\mathbf{W}}_{L,L-1}\mathbf{Q}_{L-1}$ for a unitary matrix $\mathbf{Q}_{L-1}$ by Lemma B.5.

Using $\sum_{i\in[k_\ell]} \mathcal{C}_x\left(\left[\mathbf{U}_\ell \widehat{S}^\star_\ell(x) - \widehat{S}_\ell(x)\right]_i\right) \le c_1(2^\ell)\cdot\delta_\ell^2$ from (E.5) and $\sum_{i\in[k_\ell]} \mathcal{C}_x\left(\left[\widehat{S}^\star_\ell(x)\right]_i\right) \le c_1(2^\ell) B_\ell$, it is not hard to derive that[27]

[27] Indeed, if we define $g(z) = \|\mathbf{W}_{L,L-1}\sigma(\mathbf{R}z)\|^2 = \|\overline{\mathbf{W}}_{L,L-1}(z*z)\|^2$ then we have $\mathcal{C}_z(g) \le O(1)\cdot\|\overline{\mathbf{W}}_{L,L-1}\|_F^2$ using Fact B.4, and therefore $\mathcal{C}_z(g) \le O(\tau_L^2 L^2)$ using $\|\mathbf{W}_{L,L-1}\|_F \le \tau_L$ and $\|\mathbf{R}_{L-1} * \mathbf{R}_{L-1}\|_2 \le O(L)$ from Lemma B.6.

$$\mathcal{C}_x\left(\left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\widehat{S}_{L-1}(x)\right)\right\|^2 - \left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\mathbf{U}_{L-1}\widehat{S}^\star_{L-1}(x)\right)\right\|^2\right) \leq \xi_1$$

$$\text{for some } \xi_1 \leq \tau_L^6 \cdot \mathsf{poly}(\overline{B}_L, 2^{2^L}, c_1(2^L))\delta^2_{L-1}. \quad \text{(E.8)}$$

Combining (E.7) and (E.8) with the fact that $\mathcal{C}_x(f_1+f_2) \leq 2\mathcal{C}_x(f_1) + 2\mathcal{C}_x(f_2)$, we have

$$\mathcal{C}_x\left(\left\|\overline{\mathbf{W}^\star}_{L,L-1}\left(\widehat{S}^\star_{L-1}(x) * \widehat{S}^\star_{L-1}(x)\right)\right\|^2 - \left\|\overline{\overline{\mathbf{W}}}_{L,L-1}\left(\widehat{S}^\star_{L-1}(x) * \widehat{S}^\star_{L-1}(x)\right)\right\|^2\right) = \xi_2$$

$$\text{for some } \xi_2 \leq \tau_L^6 \cdot \mathsf{poly}(\overline{B}_L, 2^{2^L}, c_1(2^L))\delta^2_{L-1} + 2c_1(2^L)\left(\frac{\varepsilon}{\alpha_L}\right)^2$$

Applying the singular value property Lemma D.1 to the above formula, we have

$$\left\|\mathbf{Sym}\left(\overline{\overline{\mathbf{W}}}^\top_{L,L-1}\overline{\overline{\mathbf{W}}}_{L,L-1}\right) - \mathbf{Sym}\left(\overline{\mathbf{W}^\star}^\top_{L,L-1}\overline{\mathbf{W}^\star}_{L,L-1}\right)\right\|_F \leq \mathsf{poly}_1\left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right) \quad \text{(E.9)}$$

for some sufficiently large polynomial

$$\mathsf{poly}_1 = \mathsf{poly}(\overline{B}_L, \kappa^{2^L}, (2^L)^{2^L}, c_1(2^L), c_3(2^L))$$

This implies

$$\begin{aligned}
\left\|\mathbf{W}^\star_{L,L-1}\sigma\left(S^\star_{L-1}(x)\right)\right\|^2 &= \left(S^\star_{L-1}(x) * S^\star_{L-1}(x)\right)^\top \overline{\mathbf{W}^\star}^\top_{L,L-1}\overline{\mathbf{W}^\star}_{L,L-1}\left(S^\star_{L-1}(x) * S^\star_{L-1}(x)\right) \\
&\overset{①}{=} \left(S^\star_{L-1}(x) * S^\star_{L-1}(x)\right)^\top \mathbf{Sym}\left(\overline{\mathbf{W}^\star}^\top_{L,L-1}\overline{\mathbf{W}^\star}_{L,L-1}\right)\left(S^\star_{L-1}(x) * S^\star_{L-1}(x)\right) \\
&\overset{②}{=} \left(S^\star_{L-1}(x) * S^\star_{L-1}(x)\right)^\top \mathbf{Sym}\left(\overline{\overline{\mathbf{W}}}^\top_{L,L-1}\overline{\overline{\mathbf{W}}}_{L,L-1}\right)\left(S^\star_{L-1}(x) * S^\star_{L-1}(x)\right) + \xi_3 \\
&\overset{③}{=} \left(S^\star_{L-1}(x) * S^\star_{L-1}(x)\right)^\top \overline{\overline{\mathbf{W}}}^\top_{L,L-1}\overline{\overline{\mathbf{W}}}_{L,L-1}\left(S^\star_{L-1}(x) * S^\star_{L-1}(x)\right) + \xi_3 \\
&= \left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\mathbf{U}_{L-1}S^\star_{L-1}(x)\right)\right\|^2 + \xi_3 \\
&= \left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}S_{L-1}(x)\right)\right\|^2 + \xi_4
\end{aligned} \quad \text{(E.10)}$$

Above, ① and ③ hold because of Fact B.4. ② holds for some error term ξ_3 with

$$\mathbb{E}[(\xi_3)^2] \leq (\mathsf{poly}_1)^2 \cdot \left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right)^2$$

because of (E.9) and $\mathbb{E}_{x\sim\mathcal{D}}[\|S^\star_\ell(x)\|^2] \leq B_\ell$ together with the hyper-contractivity Property 5.4. ④ holds for

$$\mathbb{E}[(\xi_4)^2] \leq (\mathsf{poly}_1)^3 \cdot \left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right)^2$$

because of $\mathbb{E}_{x\sim\mathcal{D}}\left\|\mathbf{U}_{L-1}S^\star_{L-1}(x) - S_{L-1}(x)\right\|^2 \leq c_1(2^{L-1}) \cdot \delta^2_{L-1}$ which implies[28]

$$\left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}S_{L-1}(x)\right)\right\|^2 - \left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\mathbf{U}_{L-1}S^\star_{L-1}(x)\right)\right\|^2 = \xi_4'$$

Next, we apply Lemma I.7 with $f^{(1)}(x) = \mathbf{U}_{L-1}\widehat{S}^\star_{L-1}(x)$ and $f^{(2)}(x) = \widehat{S}_{L-1}(x)$ to derive the bound

$$\mathcal{C}_x(g(f_1(x)) - g(f_2(x))) \leq k_L^4 \cdot 2^{O(2^L)} \cdot (c_1(2^L))^8 \cdot (\delta^8_{L-1} + \delta^2_{L-1}\overline{B}^3_L) \cdot \mathcal{C}_z(g) \ .$$

[28]Specifically, one can combine

- $\|\sigma(a) - \sigma(b)\| \leq \|a-b\| \cdot (\|a\| + 2\|a-b\|)$,
- $(\|\mathbf{W}_{L,L-1}a\|^2 - \|\mathbf{W}_{L,L-1}b\|^2)^2 \leq \|\mathbf{W}_{L,L-1}(a-b)\|^2 \cdot (2\|\mathbf{W}_{L,L-1}a\| + \|\mathbf{W}_{L,L-1}(a-b)\|)^2$,
- the spectral norm bound $\|\mathbf{W}_{L,L-1}\|_2 \leq \tau_L$, $\|\mathbf{R}_{L-1}\|_2 \leq O(\tau_L)$,

for some $\xi_4' \in \mathbb{R}$ with $\mathbb{E}_{x\sim\mathcal{D}}[(\xi_1)^2] \leq \tau_L^{12} \cdot \mathsf{poly}(\overline{B}_L, c_3(2^L))\delta_{L-1}^2$. (E.11)

E.4 Degree $2^{L-1}+2^{L-3}$ Or Lower

Let us without loss of generality assuming that $L-3 \in \mathcal{J}_L$, otherwise we move to lower degrees. We now describe the strategy for this weight matrix $\mathbf{W}_{L,L-3}$.

Let us consider all the monomials from $G^\star(x)$ and $F(x)$ in degree $2^{L-1}+2^{L-3}$. As argued above, they must come from equation (E.6).

As for the degree $2^{L-1}+2^{L-3}$ degree monomials in $G^\star(x)$ and $F(x)$, either they come from

$$\left\|\mathbf{W}^\star_{L,L-1}\sigma(S^\star_{L-1}(x))\right\|^2 \text{ and } \|\mathbf{W}_{L,L-1}\sigma(\mathbf{R}_{L-1}S_{L-1}(x))\|^2,$$

which as we have argued in (E.10), they are sufficiently close; or they come from

$$\sigma\left(\widehat{S}^\star_{L-3}(x)\right)^\top(\mathbf{W}^\star_{L,L-3})^\top\mathbf{W}^\star_{L,L-1}\sigma\left(\widehat{S}^\star_{L-1}(x)\right) \qquad \text{from } \mathbf{Sum}(G^\star_{L-1}(x))$$
$$\sigma\left(\mathbf{R}_{L-3}\widehat{S}_{L-3}(x)\right)^\top(\mathbf{W}_{L,L-3})^\top\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\widehat{S}_{L-1}(x)\right) \qquad \text{from } \mathbf{Sum}(F_{L-1}(x))$$

For this reason, suppose we compare the following two polynomials

$$G^\star(x) - \alpha_L\left\|\mathbf{W}^\star_{L,L-1}\sigma(S^\star_{L-1}(x))\right\|^2 \quad \text{vs} \quad F(x) - \alpha_L\|\mathbf{W}_{L,L-1}\sigma(\mathbf{R}_{L-1}S_{L-1}(x))\|^2 \ ,$$

they are both of degree at most $2^{L-1}+2^{L-3}$, and they differ by an error term

$$\xi_5 = \left(G^\star(x) - \alpha_L\left\|\mathbf{W}^\star_{L,L-1}\sigma(S^\star_{L-1}(x))\right\|^2\right) - \left(F(x) - \alpha_L\|\mathbf{W}_{L,L-1}\sigma(\mathbf{R}_{L-1}S_{L-1}(x))\|^2\right)$$

which satisfies (using $\mathbf{Obj} \leq \varepsilon^2$ together with (E.10))

$$\mathbb{E}_{x\sim\mathcal{D}}[(\xi_5)^2] \leq (\mathsf{poly}_1)^4 \cdot \left(\varepsilon + \tau_L^3\alpha_L\delta_{L-1}\right)^2$$

Using and the degree-preservation Property 5.2 again (for the top degree $2^{L-1}+2^{L-3}$), we have

$$\mathcal{C}_x\Bigg(\sigma\left(\widehat{S}^\star_{L-3}(x)\right)^\top\left(\mathbf{W}^\star_{L,L-3}\right)^\top\mathbf{W}^\star_{L,L-1}\sigma\left(\widehat{S}^\star_{L-1}(x)\right)$$
$$-\sigma\left(\mathbf{R}_{L-3}\widehat{S}_{L-3}(x)\right)^\top\left(\mathbf{W}_{L,L-3}\right)^\top\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\widehat{S}_{L-1}(x)\right)\Bigg) \leq \xi_6^2$$

for some error term ξ_6 with $[(\xi_6)^2] \leq (\mathsf{poly}_1)^5 \cdot \left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right)^2$. Using a similar argument as (E.8), we also have

$$\mathcal{C}_x\Bigg(\left(\mathbf{R}_{L-3}\widehat{S}_{L-3}(x)\right)^\top\left(\mathbf{W}_{L,L-3}\right)^\top\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\widehat{S}_{L-1}(x)\right)$$
$$-\sigma\left(\mathbf{R}_{L-3}\mathbf{U}_{L-3}\widehat{S}^\star_{L-3}(x)\right)^\top\left(\mathbf{W}_{L,L-3}\right)^\top\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\mathbf{U}_{L-3}\widehat{S}^\star_{L-1}(x)\right)\Bigg) \leq \xi_7$$

---

to derive that

$$\left(\left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}S_{L-1}(x)\right)\right\|^2 - \left\|\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}\mathbf{U}_{L-1}S^\star_{L-1}(x)\right)\right\|^2\right)^2$$
$$\leq O(\tau_L^{12})\cdot\left(\left\|S^\star_\ell(x)\right\|^6\left\|\mathbf{U}_\ell S^\star_\ell(x) - S_\ell(x)\right\|^2 + \left\|\mathbf{U}_\ell S^\star_\ell(x) - S_\ell(x)\right\|^8\right)$$

Using $\|a\|^6\|b\|^2 \leq O(\delta_{L-1}^2\|a\|^{12} + \frac{\|b\|^4}{\delta_{L-1}^2})$, as well as the aforementioned bounds

- $\mathbb{E}_{x\sim\mathcal{D}}\left\|S^\star_{L-1}(x)\right\|^2 \leq \overline{B}_L$ and $\mathbb{E}_{x\sim\mathcal{D}}\left\|\mathbf{U}_{L-1}S^\star_{L-1}(x) - S_{L-1}(x)\right\|^2 \leq \delta_{L-1}^2$

and the hyper-contractivity assumption (5.4), we can prove (E.11).

for $\xi_7 \le \tau_L^6 \cdot \mathsf{poly}(\overline{B}_L, 2^{2^L}, c_1(2^L))\delta_{L-1}^2$. If we define $\overline{\overline{\mathbf{W}}}_{L,L-3} = \overline{\mathbf{W}}_{L,L-3}\mathbf{Q}_{L-1}$ for the same unitary matrix $\mathbf{Q}_{L-1}$ as before, we have

$$\mathbf{W}_{L,L-3}\sigma\left(\mathbf{R}_{L-3}\mathbf{U}_{L-3}\widehat{S}^\star_{L-2}(x)\right) = \overline{\overline{\mathbf{W}}}_{L,L-3}\left(\widehat{S}^\star_{L-3}(x) * \widehat{S}^\star_{L-3}(x)\right) \quad .$$

Using this notation, the error bounds on ξ_6 and ξ_7 together imply

$$\mathcal{C}_x\Bigg(\left(\widehat{S}^\star_{L-3}(x) * \widehat{S}^\star_{L-3}(x)\right)^\top \overline{\mathbf{W}^\star}^\top_{L,L-3}\overline{\mathbf{W}^\star}_{L,L-1}\left(\widehat{S}^\star_{L-1}(x) * \widehat{S}^\star_{L-1}(x)\right)$$
$$- \left(\widehat{S}^\star_{L-3}(x) * \widehat{S}^\star_{L-3}(x)\right)^\top \overline{\overline{\mathbf{W}}}^\top_{L,L-3}\overline{\overline{\mathbf{W}}}_{L,L-1}\left(\widehat{S}^\star_{L-1}(x) * \widehat{S}^\star_{L-1}(x)\right)\Bigg)^2 \le \xi_8$$

for $\xi_8 \le (\mathsf{poly}_1)^6 \cdot \left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right)^2$. Applying the singular value property Lemma D.1 to the above formula, we have

$$\left\|\overline{\overline{\mathbf{W}}}^\top_{L,L-3}\overline{\overline{\mathbf{W}}}_{L,L-1} - \overline{\mathbf{W}^\star}^\top_{L,L-3}\overline{\mathbf{W}^\star}_{L,L-1}\right\|_F^2 \le (\mathsf{poly}_1)^7\left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right)^2 \quad .$$

Following a similar argument to (E.10), we can derive that This implies

$$\begin{aligned}&\left(\mathbf{W}^\star_{L,L-3}\sigma\left(S^\star_{L-3}(x)\right)\right)^\top\mathbf{W}^\star_{L,L-1}\sigma\left(S^\star_{L-1}(x)\right)\\ &= \left(\mathbf{W}_{L,L-3}\sigma\left(\mathbf{R}_{L-3}S_{L-3}(x)\right)\right)^\top\mathbf{W}_{L,L-1}\sigma\left(\mathbf{R}_{L-1}S_{L-1}(x)\right) + \xi_9\end{aligned}$$

for some $\mathbb{E}[(\xi_9)^2] \le (\mathsf{poly}_1)^8\left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right)^2$

E.5 Until Degree $2^{L-1}+1$

If we repeat the process in Section E.4 to analyze monomials of degrees $2^{L-1}+2^j$ until $2^{L-1}+1$ (for all $j \in \mathcal{J}_L$), eventually we can conclude that[29]

$$\left\|\overline{\overline{\mathbf{W}}}^\top_{L,L-1}\overline{\overline{\mathbf{W}}}_{L\lhd} - \overline{\mathbf{W}^\star}^\top_{L,L-1}\overline{\mathbf{W}^\star}_{L\lhd}\right\|_F \le (\mathsf{poly}_1)^{2L+3}\left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right)$$

which implies that for unitary matrix $\mathbf{Q}_{L\lhd} \stackrel{\text{def}}{=} \mathrm{diag}(\mathbf{Q}_\ell)_{\ell\in\mathcal{J}_L\setminus\{L-1\}}$, we have that

$$\left\|\mathbf{Q}^\top_{L-1}\overline{\mathbf{W}}^\top_{L,L-1}\overline{\mathbf{W}}_{L\lhd}\mathbf{Q}_{L\lhd} - \overline{\mathbf{W}^\star}^\top_{L,L-1}\overline{\mathbf{W}^\star}_{L\lhd}\right\|_F \le (\mathsf{poly}_1)^{2L+3}\left(\frac{\varepsilon}{\alpha_L} + \tau_L^3\delta_{L-1}\right)$$

Let us define

$$\mathsf{poly}_2 = (\mathsf{poly}_1)^{2L+3}\tau_L^3 \qquad \text{(we eventually choose } D_L = \mathsf{poly}_2\text{)}$$

so that

$$\left\|\mathbf{Q}^\top_{L-1}\overline{\mathbf{W}}^\top_{L,L-1}\overline{\mathbf{W}}_{L\lhd}\mathbf{Q}_{L\lhd} - \overline{\mathbf{W}^\star}^\top_{L,L-1}\overline{\mathbf{W}^\star}_{L\lhd}\right\|_F \le \mathsf{poly}_2\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \tag{E.12}$$

By the regularizer that

$$\left\|\mathbf{W}^\top_{L,L-1}\mathbf{W}_{L\lhd} - \mathbf{K}^\top_{L,L-1}\mathbf{K}_{L\lhd}\right\|_F^2 \le \frac{\varepsilon^2}{\lambda_{3,L}}$$

[29] Technically speaking, for $j \in \mathcal{J}_L \cap \{0,1\}$, one needs to modify Section E.4 a bit, because the 4-tensor becomes 3-tensor: $\left(\widehat{S}^\star_j(x)\right)^\top \overline{\mathbf{W}^\star}^\top_{L,j}\overline{\mathbf{W}^\star}_{L,L-1}\left(\widehat{S}^\star_{L-1}(x) * \widehat{S}^\star_{L-1}(x)\right)$.

Using $\overline{\mathbf{W}}_{L,j} = \mathbf{W}_{L,j}(\mathbf{R}_j * \mathbf{R}_j)$ and $\overline{\mathbf{K}}_{L,j} = \mathbf{K}_{L,j}(\mathbf{R}_j * \mathbf{R}_j)$, using the properties that $\mathbf{R}_j * \mathbf{R}_j$ is well-conditioned (see Lemma B.6), and using $\mathbf{Q}_{L-1}$ and $\mathbf{Q}_{L\triangleleft}$ are unitary (see Lemma B.5), we have

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{W}}_{L,L-1}^\top \overline{\mathbf{W}}_{L\triangleleft}\mathbf{Q}_{L\triangleleft} - \mathbf{Q}_{L-1}^\top \overline{\mathbf{K}}_{L,L-1}^\top \overline{\mathbf{K}}_{L\triangleleft}\mathbf{Q}_{L\triangleleft}\right\|_F^2 \le \frac{\varepsilon^2}{\lambda_{3,L}} \cdot \mathsf{poly}(\overline{k}_L, L) \tag{E.13}$$

By our choice of$\lambda_{3,L} \ge \frac{1}{\mathsf{poly}_2 \cdot \Upsilon_L}\alpha_L^2$ and (E.12), we have

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{K}}_{L,L-1}^\top \overline{\mathbf{K}}_{L\triangleleft}\mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L\triangleleft}\right\|_F \le \sqrt{\Upsilon_L}(\mathsf{poly}_2)^2\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \tag{E.14}$$

E.6 Deriving $\overline{\mathbf{K}}_L$ Close To $\overline{\mathbf{W}^\star}_L$

Since $\|\mathbf{K}_{L,\triangleleft}\|_F, \|\mathbf{K}_{L,L-1}\|_F \le \tau_L$, we have $\|\overline{\mathbf{K}}_{L,\triangleleft}\|_F, \|\overline{\mathbf{K}}_{L,L-1}\|_F \le O(\tau_L L)$ from Lemma B.6. Also, the singular values of $\overline{\mathbf{W}^\star}_{L\triangleleft}, \overline{\mathbf{W}^\star}_{L,L-1}$ are between $1/\kappa$ and $L\kappa$ (see Fact B.7). Therefore, applying Claim I.9 to (E.14), we know that there exists square matrix $\mathbf{P} \in \mathbb{R}^{k_L \times k_L}$ satisfying[30]

$$\left\|\overline{\mathbf{K}}_{L,L-1}\mathbf{Q}_{L-1} - \mathbf{P}\overline{\mathbf{W}^\star}_{L,L-1}\right\|_F \le \sqrt{\Upsilon_L}(\mathsf{poly}_2)^3\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right)$$

$$\left\|\overline{\mathbf{K}}_{L\triangleleft}\mathbf{Q}_{L\triangleleft} - (\mathbf{P}^\top)^{-1}\overline{\mathbf{W}^\star}_{L\triangleleft}\right\|_F \le \sqrt{\Upsilon_L}(\mathsf{poly}_2)^3\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right)$$

and all the singular values of $\mathbf{P}$ are between $\frac{1}{\mathsf{poly}(\tau_L)}$ and $\mathsf{poly}(\tau_L)$. This implies that

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{K}}_{L,L-1}^\top \overline{\mathbf{K}}_{L,L-1}\mathbf{Q}_{L-1} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \mathbf{P}^\top\mathbf{P}\overline{\mathbf{W}^\star}_{L,L-1}\right\|_F \le \sqrt{\Upsilon_L}(\mathsf{poly}_2)^4\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \tag{E.15}$$

$$\left\|\mathbf{Q}_{L\triangleleft}^\top \overline{\mathbf{K}}_{L\triangleleft}^\top \overline{\mathbf{K}}_{L\triangleleft}\mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{L\triangleleft}^\top (\mathbf{P}^\top\mathbf{P})^{-1}\overline{\mathbf{W}^\star}_{L\triangleleft}\right\|_F \le \sqrt{\Upsilon_L}(\mathsf{poly}_2)^4\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \tag{E.16}$$

Our regularizer $\lambda_{4,L}$ ensures that

$$\left\|\mathbf{W}_{L,L-1}^\top \mathbf{W}_{L,L-1} - \mathbf{K}_{L,L-1}^\top \mathbf{K}_{L,L-1}\right\|_F^2 \le \frac{\varepsilon^2}{\lambda_{4,L}}$$

Using $\overline{\mathbf{W}}_{L,j} = \mathbf{W}_{L,j}(\mathbf{R}_j * \mathbf{R}_j)$ and $\overline{\mathbf{K}}_{L,j} = \mathbf{K}_{L,j}(\mathbf{R}_j * \mathbf{R}_j)$, using the properties that $\mathbf{R}_j * \mathbf{R}_j$ is well-conditioned (see Lemma B.6), and using $\mathbf{Q}_{L-1}$ and $\mathbf{Q}_{L\triangleleft}$ are unitary (see Lemma B.5), we have

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{W}}_{L,L-1}^\top \overline{\mathbf{W}}_{L,L-1}\mathbf{Q}_{L-1} - \mathbf{Q}_{L-1}^\top \overline{\mathbf{K}}_{L,L-1}^\top \overline{\mathbf{K}}_{L,L-1}\mathbf{Q}_{L-1}\right\|_F^2 \le \frac{\varepsilon^2}{\lambda_{4,L}} \cdot \mathsf{poly}(\overline{k}_L, L)$$

By our choice$\lambda_{4,L} \ge \frac{1}{(\mathsf{poly}_2)^7\sqrt{\Upsilon_L^2}}\alpha_L^2$, this together with (E.15) implies

$$\begin{aligned}
&\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{W}}_{L,L-1}^\top \overline{\mathbf{W}}_{L,L-1}\mathbf{Q}_{L-1} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \mathbf{P}^\top\mathbf{P}\overline{\mathbf{W}^\star}_{L,L-1}\right\|_F \le 2\sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^4\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \\
\iff &\left\|\overline{\overline{\mathbf{W}}}_{L,L-1}^\top \overline{\overline{\mathbf{W}}}_{L,L-1} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \mathbf{P}^\top\mathbf{P}\overline{\mathbf{W}^\star}_{L,L-1}\right\|_F \le 2\sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^4\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \qquad \text{(E.17)}
\end{aligned}$$

Recall we have already concluded in (E.9) that

$$\left\|\mathbf{Sym}\left(\overline{\overline{\mathbf{W}}}_{L,L-1}^\top \overline{\overline{\mathbf{W}}}_{L,L-1}\right) - \mathbf{Sym}\left(\overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L,L-1}\right)\right\|_F \le \mathsf{poly}_2\left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right)$$

[30] We note here, to apply Claim I.9, one also needs to ensure $\varepsilon \le \frac{\alpha_L}{(\mathsf{poly}_2)^3\sqrt{\Upsilon_L}}$ and $\delta_{L-1} \le \frac{1}{(\mathsf{poly}_2)^3\sqrt{\Upsilon_L}}$; however, both of them are satisfied under the assumptions $\varepsilon \le \frac{\alpha_L}{(D_L)^9\Upsilon_L}$ and $\frac{\alpha_L}{\alpha_{L-1}} \le \frac{1}{4\Upsilon_L^3(D_L)^{16}C_{L-1}}$, and the definition of δ_{L-1} from (E.4).

so putting it into (E.17) we have

$$\left\|\mathbf{Sym}\left(\overline{\mathbf{W}^\star}_{L,L-1}^\top \mathbf{P}^\top \mathbf{P} \overline{\mathbf{W}^\star}_{L,L-1}\right) - \mathbf{Sym}\left(\overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L,L-1}\right)\right\|_F \leq 3\sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^4 \left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right)$$

Since $\overline{\mathbf{W}^\star}_{L,L-1} = \mathbf{W}^\star_{L,L-1}$, by Fact B.4, we know that for any matrix $\mathbf{P}$,

$$\mathbf{Sym}\left(\overline{\mathbf{W}^\star}_{L,L-1}^\top \mathbf{P}^\top \mathbf{P} \overline{\mathbf{W}^\star}_{L,L-1}\right) = \overline{\mathbf{W}^\star}_{L,L-1}^\top \mathbf{P}^\top \mathbf{P} \overline{\mathbf{W}^\star}_{L,L-1}$$

This implies

$$\left\|\overline{\mathbf{W}^\star}_{L,L-1}^\top \mathbf{P}^\top \mathbf{P} \overline{\mathbf{W}^\star}_{L,L-1} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L,L-1}\right\|_F \leq 4\sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^4 \left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \quad .$$

By expanding $\overline{\mathbf{W}^\star}_{L,L-1}$ into its SVD decomposition, one can derive from the above inequality that

$$\left\|\mathbf{P}^\top \mathbf{P} - \mathbf{I}\right\|_F \leq \sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^5 \left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \tag{E.18}$$

Putting this back to (E.15) and (E.16), we have

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{K}}_{L,L-1}^\top \overline{\mathbf{K}}_{L,L-1} \mathbf{Q}_{L-1} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L,L-1}\right\|_F \leq \sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^6 \left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right)$$

$$\left\|\mathbf{Q}_{L\lhd}^\top \overline{\mathbf{K}}_{L\lhd}^\top \overline{\mathbf{K}}_{L\lhd} \mathbf{Q}_{L\lhd} - \overline{\mathbf{W}^\star}_{L\lhd}^\top \overline{\mathbf{W}^\star}_{L\lhd}\right\|_F \leq \sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^6 \left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right)$$

Combining this with (E.14), we derive that (denoting by $\vec{\mathbf{Q}}_L \stackrel{\text{def}}{=} \operatorname{diag}(\mathbf{Q}_\ell)_{\ell \in \mathcal{J}_L}$)

$$\left\|\vec{\mathbf{Q}}_L^\top \overline{\mathbf{K}}_L^\top \overline{\mathbf{K}}_L \vec{\mathbf{Q}}_L - \overline{\mathbf{W}^\star}_L^\top \overline{\mathbf{W}^\star}_L\right\|_F \leq \sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^7 \left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \tag{E.19}$$

E.7 Deriving $S_L(x)$ Close To $S_L^\star(x)$, Construct $\mathbf{U}_L$

From (E.19) we can also apply Claim I.10 and derive the existence of some unitary $\mathbf{U}_L \in \mathbb{R}^{k_L \times k_L}$ so that[31]

$$\left\|\overline{\mathbf{K}}_L \vec{\mathbf{Q}}_L - \mathbf{U}_L \overline{\mathbf{W}^\star}_L\right\|_F \leq \sqrt{\Upsilon_L^2}(\mathsf{poly}_2)^8 \left(\frac{\varepsilon}{\alpha_L} + \delta_{L-1}\right) \quad . \tag{E.20}$$

Simultaneously right applying the two matrices in (E.20) by the vector (where the operator $\frown$ is for concatenating two vectors)

$$\left(S_j^\star(x) * S_j^\star(x)\right)_{j \in \mathcal{J}_L \setminus \{0,1\}} \frown \left(S_j^\star(x)\right)_{j \in \mathcal{J}_L \setminus \{0,1\}} \ ,$$

we have

$$\sum_{j \in \mathcal{J}_L \setminus \{0,1\}} \mathbf{K}_{L,j} \sigma\left(\mathbf{R}_j \mathbf{U}_j S_j^\star(x)\right) + \sum_{j \in \mathcal{J}_L \cap \{0,1\}} \mathbf{K}_{L,j} S_j^\star(x)$$
$$= \mathbf{U}_L \left(\sum_{j \in \mathcal{J}_L \setminus \{0,1\}} \mathbf{W}^\star_{L,j} \sigma\left(S_j^\star(x)\right) + \sum_{j \in \mathcal{J}_L \cap \{0,1\}} \mathbf{W}^\star_{L,j} S_j^\star(x)\right) + \xi_{10}$$

[31] We note here, to apply Claim I.10, one also needs to ensure $\varepsilon \leq \frac{\alpha_L}{(\mathsf{poly}_2)^8 \sqrt{\Upsilon_L^2}}$ and $\delta_{L-1} \leq \frac{1}{(\mathsf{poly}_2)^8 \sqrt{\Upsilon_L^2}}$; however, both of them are satisfied under the assumptions $\varepsilon \leq \frac{\alpha_L}{(D_L)^9 \Upsilon_L}$ and $\frac{\alpha_L}{\alpha_{L-1}} \leq \frac{1}{4\Upsilon_L^3 (D_L)^{16} C_{L-1}}$, and the definition of δ_{L-1} from (E.4).

for some error vector ξ_{10} with

$$\mathbb{E}_{x\sim\mathcal{D}}[\|\xi_{10}\|^2] \le \Upsilon_L^2 \cdot L\overline{B}_L^2(\mathsf{poly}_2)^{16}\left(\frac{\varepsilon}{\alpha_L}+\delta_{L-1}\right)^2 .$$

Combining it with $\mathbb{E}_{x\sim\mathcal{D}}\left\|\mathbf{U}_{L-1}S^\star_{L-1}(x)-S_{L-1}(x)\right\|_2^2 \le \delta_{L-1}^2$ (see (E.4)) we know

$$\begin{aligned}
S_L(x) &= \sum_{j\in\mathcal{J}_L\setminus\{0,1\}} \mathbf{K}_{L,j}\sigma\left(\mathbf{R}_jS_j(x)\right) + \sum_{j\in\mathcal{J}_L\cap\{0,1\}} \mathbf{K}_{L,j}S_j(x) \\
&= \mathbf{U}_L\left(\sum_{j\in\mathcal{J}_L\setminus\{0,1\}} \mathbf{W}^\star_{L,j}\sigma\left(S^\star_j(x)\right) + \sum_{j\in\mathcal{J}_L\cap\{0,1\}} \mathbf{W}^\star_{L,j}S^\star_j(x)\right) + \xi_{11} = \mathbf{U}_LS^\star_L(x)+\xi_{11}
\end{aligned}$$

for some error vector ξ_{11} with

$$\mathbb{E}_{x\sim\mathcal{D}}[\|\xi_{11}\|^2] = \mathbb{E}_{x\sim\mathcal{D}}\|\mathbf{U}_LS^\star_L(x)-S_L(x)\|_2^2 \le \Upsilon_L^2(\mathsf{poly}_2)^{17}\left(\frac{\varepsilon}{\alpha_L}+\delta_{L-1}\right)^2 . \tag{E.21}$$

E.8 Deriving $F_L(x)$ Close To $G^\star(x)$

By the regularizer $\lambda_{5,L}$, we have that

$$\left\|\mathbf{W}_L^\top\mathbf{W}_L - \mathbf{K}_L^\top\mathbf{K}_L\right\|_F \le \frac{\varepsilon^2}{\lambda_{5,L}} \tag{E.22}$$

Using $\overline{\mathbf{W}}_{L,j} = \mathbf{W}_{L,j}(\mathbf{R}_j * \mathbf{R}_j)$ and $\overline{\mathbf{K}}_{L,j} = \mathbf{K}_{L,j}(\mathbf{R}_j * \mathbf{R}_j)$, using the properties that $\mathbf{R}_j * \mathbf{R}_j$ is well-conditioned (see Lemma B.6), and using $\mathbf{Q}_{L-1}$ and $\mathbf{Q}_{L\triangleleft}$ are unitary (see Lemma B.5), we have

$$\left\|\vec{\mathbf{Q}}_L^\top\overline{\mathbf{W}}_L^\top\overline{\mathbf{W}}_L\vec{\mathbf{Q}}_L - \vec{\mathbf{Q}}_L^\top\overline{\mathbf{K}}_L^\top\overline{\mathbf{K}}_L\vec{\mathbf{Q}}_L\right\|_F^2 \le \frac{\varepsilon^2}{\lambda_{5,L}}\cdot\mathsf{poly}(\overline{k}_L, L)$$

By our choice of $\lambda_{5,L} \ge \frac{1}{(\mathsf{poly}_2)^{13}\Upsilon_L^3}\alpha_L^2$, together with (E.19), we have that

$$\left\|\vec{\mathbf{Q}}_L^\top\overline{\mathbf{W}}_L^\top\overline{\mathbf{W}}_L\vec{\mathbf{Q}}_L - \overline{\mathbf{W}^\star}_L^\top\overline{\mathbf{W}^\star}_L\right\|_F \le \sqrt{\Upsilon_L^3}(\mathsf{poly}_2)^7\left(\frac{\varepsilon}{\alpha_L}+\delta_{L-1}\right) .$$

Note from the definition of $\mathbf{Sum}(F_L(x))$ and $\mathbf{Sum}(G^\star_L(x))$ (see (E.6)) we have

$$\begin{aligned}
\mathbf{Sum}(G^\star_L(x)) &= \left\|\overline{\mathbf{W}^\star}_L(S^\star_{L-1}(x) * S^\star_{L-1}(x),\dots)\right\|^2 \\
\mathbf{Sum}(F_L(x)) &= \left\|\overline{\mathbf{W}}_L(S_{L-1}(x) * S_{L-1}(x),\dots)\right\|^2
\end{aligned}$$

so using a similar derivation as (E.10), we have

$$\mathbb{E}_{x\sim\mathcal{D}}\left(\mathbf{Sum}(F_L(x)) - \mathbf{Sum}(G^\star_L(x))\right)^2 \le \Upsilon_L^3(\mathsf{poly}_2)^{15}\left(\frac{\varepsilon}{\alpha_L}+\delta_{L-1}\right)^2 . \tag{E.23}$$

E.9 Recursion

We can now put (E.23) back to the bound of $\mathbf{Obj}_{L-1}$ (see (E.1)) and derive that

$$\begin{aligned}
\mathbf{Obj}_{L-1} &\le 2\alpha_L^2\mathbb{E}_{x\sim\mathcal{D}}\left(\mathbf{Sum}(F_L(x)) - \mathbf{Sum}(G^\star_L(x))\right)^2 + 2\mathbf{Obj} \\
&\le \Upsilon_L^3(\mathsf{poly}_2)^{16}\left(\delta_{L-1}^2\alpha_L^2+\varepsilon^2\right) .
\end{aligned} \tag{E.24}$$

Note this is a tighter upper bound on $\mathbf{Obj}_{L-1}$ comparing to the previously used one in (E.3). Therefore, we can apply the induction hypothesis again and replace (E.4) also with a tighter bound

$$\forall \ell = 2,3,\dots,L-1: \quad \mathbb{E}_{x\sim\mathcal{D}} \|\mathbf{U}_\ell S^\star_\ell(x) - S_\ell(x)\|_2^2 \leq \left(\frac{\varepsilon + \delta_{L-1}\alpha_L}{\sqrt{\alpha_\ell \alpha_{\ell+1}}}\right)^2 \Upsilon_L^3 (\mathsf{poly}_2)^{16} C_{L-1} \ . \tag{E.25}$$

In other words, we can replace our previous crude bound on δ_{L-1} (see (E.3)) with this tighter bound (E.25), and repeat. By our assumption, $\frac{\alpha_L}{\alpha_{L-1}} \leq \frac{1}{4\Upsilon_L^3 (D_L)^{16} C_{L-1}}$, this implies that the process ends when[32]

$$\delta_{L-1}^2 = \left(\frac{\varepsilon}{\sqrt{\alpha_{L-1}\alpha_L}}\right)^2 \cdot 2\Upsilon_L^3 (\mathsf{poly}_2)^{16} C_{L-1} \ . \tag{E.26}$$

Plugging this choice back to (E.25), we have for every $\ell = 2,3,\dots,L-1$

$$\mathbb{E}_{x\sim\mathcal{D}} \|\mathbf{U}_\ell S^\star_\ell(x) - S_\ell(x)\|_2^2 \leq \left(\frac{\varepsilon}{\sqrt{\alpha_\ell\alpha_{\ell+1}}}\right)^2 \cdot 2\Upsilon_L^3 (\mathsf{poly}_2)^{16} C_{L-1} \leq \left(\frac{\varepsilon}{\sqrt{\alpha_\ell\alpha_{\ell+1}}}\right)^2 C_L$$

As for the case of $\ell = L$, we derive from (E.21) that

$$\mathbb{E}_{x\sim\mathcal{D}} \|\mathbf{U}_L S^\star_L(x) - S_L(x)\|_2^2 \leq 2\Upsilon_L^2 (\mathsf{poly}_2)^{17} \left(\frac{\varepsilon}{\alpha_L}\right)^2 \leq \left(\frac{\varepsilon}{\sqrt{\alpha_L\alpha_{L+1}}}\right)^2 C_L$$

This completes the proof of Theorem E.1. ■

E.10 Proof of Corollary E.3

Proof of Corollary E.3. As for Corollary E.3, we first note that our final choice of δ_{L-1} (see (E.26)), when plugged into (E.12), (E.14), (E.19) and (E.21), respectively give us

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{W}}_{L,L-1}^\top \overline{\mathbf{W}}_{L\triangleleft} \mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L\triangleleft}\right\|_F^2 \leq 2(D_L)^2 \left(\frac{\varepsilon}{\alpha_L}\right)^2$$

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{K}}_{L,L-1}^\top \overline{\mathbf{K}}_{L\triangleleft} \mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L\triangleleft}\right\|_F^2 \leq 2\Upsilon_L (D_L)^4 \left(\frac{\varepsilon}{\alpha_L}\right)^2$$

$$\left\|\vec{\mathbf{Q}}_L^\top \overline{\mathbf{K}}_L^\top \overline{\mathbf{K}}_L \vec{\mathbf{Q}}_L - \overline{\mathbf{W}^\star}_L^\top \overline{\mathbf{W}^\star}_L\right\|_F^2 \leq 2\Upsilon_L^2 (D_L)^{14} \left(\frac{\varepsilon}{\alpha_L}\right)^2$$

$$\mathbb{E}_{x\sim\mathcal{D}} \|\mathbf{U}_L S^\star_L(x) - S_L(x)\|_2^2 \leq 2\Upsilon_L^2 (D_L)^{17} \left(\frac{\varepsilon}{\alpha_L}\right)^2$$

So far this has only given us bounds for the L-th layer. As for other layers $\ell = 2,3,\dots,L-1$, we note that our final choice of δ_{L-1} (see (E.26)), when plugged into the formula of $\mathbf{Obj}_{L-1}$ (see (E.24)), in fact gives

$$\mathbf{Obj}_{L-1} \leq 2\Upsilon_L^3 (D_L)^{16} \varepsilon^2 < \left(2\sqrt{\Upsilon_L^3} (D_L)^8 \varepsilon\right)^2 \ll \left(\frac{\alpha_{L-1}}{(D_{L-1})^9 \Upsilon_{L-1}}\right)^2 \ .$$

using our assumptions $\varepsilon \leq \frac{\alpha_L}{(D_L)^9 \Upsilon_L}$ and $\frac{\alpha_L}{\alpha_{L-1}} \leq \frac{1}{4\Upsilon_L^3 (D_L)^{16} C_{L-1}}$. Therefore, we can recurse to the case of $L-1$ with ε^2 replaced with $4\Upsilon_L^3 (D_L)^{16} \varepsilon^2$. Continuing in this fashion gives the desired bounds.

Finally, our assumption $\varepsilon \leq \frac{\alpha_L}{(D_L)^9 \Upsilon_L}$ implies $\mathbb{E}_{x\sim\mathcal{D}} \|\mathbf{U}_L S^\star_L(x) - S_L(x)\|_2^2 \leq 1$, and using gap

[32]To be precise, we also need to verify that this new $\delta_{L-1} \leq \frac{1}{(\mathsf{poly}_2)^8}$ as before, but this is ensured from our assumptions $\varepsilon \leq \frac{\alpha_L}{(D_L)^9 \Upsilon_L}$ and $\frac{\alpha_L}{\alpha_{L-1}} \leq \frac{1}{4\Upsilon_L^3 (D_L)^{16} C_{L-1}}$.

assumption it also holds for previous layers:

$$\forall \ell < L: \qquad \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \|\mathbf{U}_\ell S^\star_\ell(x) - S_\ell(x)\|_2^2 \leq 2\Upsilon_\ell^2 (D_\ell)^{17} \left(\frac{\varepsilon}{\alpha_\ell}\right)^2 \cdot \frac{C_L}{C_\ell} \leq 1$$

They also imply $\mathbb{E}_{x\sim\mathcal{D}} \|S_\ell(x)\|_2^2 \leq 2B_\ell$ using $\mathbb{E}_{x\sim\mathcal{D}} \|S^\star_\ell(x)\|_2^2 \leq B_\ell$. □

E.11 Proof of Corollary E.4

Proof of Corollary E.4. This time, we begin by recalling that from (E.3):

$$\mathbf{Obj}_{L-1} \leq \alpha_L^2 \cdot (k_L L \overline{B}_L \tau_L)^8 c_3(2^L) + 2\varepsilon^2 \leq \alpha_L^2 \cdot D_L$$

Therefore, we can use $\varepsilon^2 = \alpha_L^2 \cdot D_L$ and apply Theorem E.1 and Corollary E.3 for the case of $L-1$. This is why we choose $\varepsilon_0 = \alpha_L \cdot \sqrt{D_L}$ for $\ell < L$.

As for the case of $\ell = L$, we first note the $L-1$ case tells us

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left\|\mathbf{U}_{L-1} S^\star_{L-1}(x) - S_{L-1}(x)\right\|_2^2 \leq \delta_{L-1}^2 \stackrel{\text{def}}{=} 6\Upsilon_{L-1}^2 (D_{L-1})^{17} \left(\frac{\varepsilon}{\alpha_{L-1}}\right)^2 \ll \left(\frac{\varepsilon}{\alpha_L}\right)^2$$

Therefore, we can plug in this choice of δ_{L-1} into (E.12), (E.14) and (E.19) to derive

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{W}}_{L,L-1}^\top \overline{\mathbf{W}}_{L\triangleleft} \mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L\triangleleft}\right\|_F^2 \leq 2(D_L)^2 \left(\frac{\varepsilon}{\alpha_L}\right)^2$$

$$\left\|\mathbf{Q}_{L-1}^\top \overline{\mathbf{K}}_{L,L-1}^\top \overline{\mathbf{K}}_{L\triangleleft} \mathbf{Q}_{L\triangleleft} - \overline{\mathbf{W}^\star}_{L,L-1}^\top \overline{\mathbf{W}^\star}_{L\triangleleft}\right\|_F^2 \leq 2\Upsilon_L (D_L)^4 \left(\frac{\varepsilon}{\alpha_L}\right)^2$$

$$\left\|\vec{\mathbf{Q}}_L^\top \overline{\mathbf{K}}_L^\top \overline{\mathbf{K}}_L \vec{\mathbf{Q}}_L - \overline{\mathbf{W}^\star}_L^\top \overline{\mathbf{W}^\star}_L\right\|_F^2 \leq 2\Upsilon_L^2 (D_L)^{14} \left(\frac{\varepsilon}{\alpha_L}\right)^2$$

Note that the three equations (E.12), (E.14) and (E.19) have only required the weaker requirement $\varepsilon \leq \frac{\alpha_L}{(D_L)^3 \sqrt{\Upsilon_L}}$ on ε comparing to the full Theorem E.1 (the stronger requirement was $\varepsilon \leq \frac{\alpha_L}{(D_L)^9 \Upsilon_L}$, but it is required only starting from equation (E.20)). □

F Construction of Descent Direction

Let $\mathbf{U}_\ell$ be defined as in Theorem E.1. Let us construct $\mathbf{V}^\star_{\ell,j} \in \mathbb{R}^{k_\ell \times \binom{k_j+1}{2}}$ or $\mathbb{R}^{k_\ell \times d}$ that satisfies

$$\forall j > 2: \mathbf{V}^\star_{\ell,j} \sigma(\mathbf{R}_j \mathbf{U}_j z) = \mathbf{W}^\star_{\ell,j} \sigma(z), \quad \forall j' \in [2], \mathbf{V}^\star_{\ell,j'} = \mathbf{W}^\star_{\ell,j'} \tag{F.1}$$

and the singular values of $\mathbf{V}^\star_{\ell,j}$ are between $[\frac{1}{O(k_\ell^4 L^2 \kappa)}, O(L^2\kappa)]$. (This can be done by defining $\mathbf{V}^\star_{\ell,j} = \mathbf{W}^\star_{\ell,j}(\mathbf{I}*\mathbf{I})(\mathbf{R}_j\mathbf{U}_j * \mathbf{R}_j \mathbf{U}_j)^{-1} \in \mathbb{R}^{k_\ell \times \binom{k_j+1}{2}}$, and the singular value bounds are due to Fact B.7, Lemma B.5 and Lemma B.6.) Let us also introduce notations

$$\begin{aligned}
\mathbf{E}_\ell &\stackrel{\text{def}}{=} \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_\ell - \left(\mathbf{V}^\star_{\ell,\ell-1}\right)^\top \mathbf{V}^\star_\ell = (\mathbf{E}_{\ell,\ell-1}, \mathbf{E}_{\ell\triangleleft}) \\
\mathbf{E}_{\ell\triangleleft} &\stackrel{\text{def}}{=} \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell\triangleleft} - \left(\mathbf{V}^\star_{\ell,\ell-1}\right)^\top \mathbf{V}^\star_{\ell\triangleleft} \\
\mathbf{E}_{\ell,\ell-1} &\stackrel{\text{def}}{=} \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1} - \left(\mathbf{V}^\star_{\ell,\ell-1}\right)^\top \mathbf{V}^\star_{\ell,\ell-1} \\
\widehat{\mathbf{E}}_\ell &\stackrel{\text{def}}{=} \mathbf{K}_\ell^\top \mathbf{K}_\ell - (\mathbf{V}^\star_\ell)^\top \mathbf{V}^\star_\ell
\end{aligned}$$

Let us consider updates (for some $\eta_2 \geq \eta_1$):

$$\mathbf{W}_\ell \leftarrow \sqrt{1-\eta_1}\mathbf{W}_\ell + \sqrt{\eta_1}\mathbf{D}_\ell \mathbf{V}_\ell^{\star,w}$$
$$\mathbf{K}_{\ell\triangleleft} \leftarrow \left(1+\frac{\eta_1}{2}\right)\mathbf{K}_{\ell\triangleleft} - \eta_1 \mathbf{Q}_\ell \mathbf{K}_{\ell\triangleleft} - \eta_2 \mathbf{K}_{\ell,\ell-1}\mathbf{E}_{\ell\triangleleft}$$
$$\mathbf{K}_{\ell,\ell-1} \leftarrow \left(1-\frac{\eta_1}{2}\right)\mathbf{K}_{\ell,\ell-1} + \eta_1 \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1} - \eta_2 \mathbf{K}_{\ell\triangleleft}\mathbf{E}_{\ell\triangleleft}^\top$$

where $\mathbf{V}_\ell^{\star,w} \in \mathbb{R}^{m\times *}$ is defined as $(\mathbf{V}_\ell^{\star,w})^\top = \frac{\sqrt{k_\ell}}{\sqrt{m}}((\mathbf{V}_\ell^\star)^\top, \dots (\mathbf{V}_\ell^\star)^\top)$ which contains $\frac{m}{k_\ell}$ identical copies of $\mathbf{V}_\ell^\star$, and $\mathbf{D}_\ell \in \mathbb{R}^{m\times m}$ is a diagonal matrix with diagonals as random ± 1, and $\mathbf{Q}_\ell$ is a symmetric matrix given by

$$\mathbf{Q}_\ell = \frac{1}{2}\left(\mathbf{K}_{\ell,\ell-1}\mathbf{K}_{\ell,\ell-1}^\top\right)^{-1}\mathbf{K}_{\ell,\ell-1}\left(\mathbf{V}_{\ell,\ell-1}^\star\right)^\top \mathbf{V}_{\ell,\ell-1}^\star \mathbf{K}_{\ell,\ell-1}^\top \left(\mathbf{K}_{\ell,\ell-1}\mathbf{K}_{\ell,\ell-1}^\top\right)^{-1}$$

F.1 Simple Properties

Fact F.1. *Suppose we know $\|\mathbf{W}_\ell\|_F \leq \widetilde{\kappa}_\ell$. Then,*

$$(\mathbf{W}_\ell^{(\mathsf{new})})^\top(\mathbf{W}_\ell^{(\mathsf{new})}) = (1-\eta_1)(\mathbf{W}_\ell)^\top \mathbf{W}_\ell + \eta_1 (\mathbf{V}_\ell^\star)^\top \mathbf{V}_\ell^\star + \sqrt{\eta_1}\xi$$

for some error matrix ξ with

$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}[\xi] = \mathbf{0} \quad \textit{and} \quad \mathop{\mathbf{Pr}}_{\mathbf{D}_\ell}\left[\|\xi\|_F > \frac{\log\delta^{-1}\cdot \mathsf{poly}(\widetilde{\kappa}_\ell)}{\sqrt{m}}\right] \leq \delta \quad \textit{and} \quad \mathop{\mathbb{E}}_{\mathbf{D}_\ell}[\|\xi\|_F^2] \leq \frac{\mathsf{poly}(\widetilde{\kappa}_\ell)}{m}$$

Proof. Trivial from vector version of Hoeffding's inequality. $\square$

Claim F.2. *Suppose $\sigma_{\min}(\mathbf{K}_{\ell,\ell-1}), \sigma_{\min}(\mathbf{K}_{\ell\triangleleft}) \geq \frac{1}{2\widetilde{\kappa}}$ and $\|\mathbf{K}_\ell\|_2 \leq 2\widetilde{\kappa}$ for some $\widetilde{\kappa} \geq \kappa + k_\ell + L$, we have:*

$$\langle \mathbf{E}_{\ell\triangleleft}, \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1}\mathbf{E}_{\ell\triangleleft} + \mathbf{E}_{\ell\triangleleft}\mathbf{K}_{\ell\triangleleft}^\top \mathbf{K}_{\ell\triangleleft}\rangle \geq \frac{1}{\mathsf{poly}(\widetilde{\kappa})}\|\mathbf{E}_{\ell\triangleleft}\|_F^2$$

Proof of Claim F.2. We first note the left hand side

$$LHS = \|\mathbf{K}_{\ell,\ell-1}\mathbf{E}_{\ell\triangleleft}\|_F^2 + \|\mathbf{K}_{\ell\triangleleft}\mathbf{E}_{\ell\triangleleft}^\top\|_F^2$$

Without loss of generality (by left/right multiplying with a unitary matrix), let us write $\mathbf{K}_{\ell,\ell-1} = (\mathbf{K}_1, \mathbf{0})$ and $\mathbf{K}_{\ell\triangleleft} = (\mathbf{K}_2, \mathbf{0})$ for square matrices $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{k_\ell\times k_\ell}$. Accordingly, let us write $\mathbf{E}_{\ell\triangleleft} = \left(\begin{smallmatrix}\mathbf{E}_1 & \mathbf{E}_2\\ \mathbf{E}_3 & \mathbf{E}_4\end{smallmatrix}\right)$ for $\mathbf{E}_1 \in \mathbb{R}^{k_\ell\times k_\ell}$. We have

$$LHS = \|(\mathbf{K}_1\mathbf{E}_1, \mathbf{K}_1\mathbf{E}_2)\|_F^2 + \|(\mathbf{K}_2\mathbf{E}_1^\top, \mathbf{K}_2\mathbf{E}_3^\top)\|_F^2 \geq \frac{1}{\mathsf{poly}(\widetilde{\kappa})}(\|\mathbf{E}_1\|_F^2 + \|\mathbf{E}_2\|_F^2 + \|\mathbf{E}_3\|_F^2) \ .$$

Note also $\|\mathbf{E}_{\ell\triangleleft}\|_F \leq \mathsf{poly}(\widetilde{\kappa})$. Let us write $\mathbf{V}_{\ell,\ell-1}^\star = (\mathbf{V}_1, \mathbf{V}_2)$ and $\mathbf{V}_{\ell\triangleleft}^\star = (\mathbf{V}_3, \mathbf{V}_4)$ for square matrices $\mathbf{V}_1, \mathbf{V}_3 \in \mathbb{R}^{k_\ell\times k_\ell}$. Then we have

$$\mathbf{E}_{\ell\triangleleft} = \left(\begin{smallmatrix}\mathbf{E}_1 & \mathbf{E}_2\\ \mathbf{E}_3 & \mathbf{E}_4\end{smallmatrix}\right) = \left(\begin{smallmatrix}\mathbf{K}_1^\top\mathbf{K}_2 - \mathbf{V}_1^\top\mathbf{V}_3 & -\mathbf{V}_1^\top\mathbf{V}_4\\ -\mathbf{V}_2^\top\mathbf{V}_3 & -\mathbf{V}_2^\top\mathbf{V}_4\end{smallmatrix}\right) \tag{F.2}$$

Recall we have $\|\mathbf{V}_{\ell,\ell-1}^\star\|_2, \|\mathbf{V}_{\ell\triangleleft}^\star\|_2 \leq L^2\kappa$. Consider two cases.

In the first case, $\sigma_{\min}(\mathbf{V}_1) \leq \frac{1}{16L^2\kappa(\widetilde{\kappa})^2}$. Then, it satisfies $\|\mathbf{E}_1\|_F \geq \frac{1}{2}\|\mathbf{K}_1^\top\mathbf{K}_2\|_F \geq \frac{1}{8(\widetilde{\kappa})^2}$ so we are done. In the second case, $\sigma_{\min}(\mathbf{V}_1) \geq \frac{1}{16L^2\kappa(\widetilde{\kappa})^2}$. We have

$$\|\mathbf{E}_2\|_F = \|\mathbf{V}_1^\top\mathbf{V}_4\|_F \geq \sigma_{\min}(\mathbf{V}_1)\|\mathbf{V}_4\|_F \geq \frac{\sigma_{\min}(\mathbf{V}_1)}{\sigma_{\max}(\mathbf{V}_2)}\|\mathbf{V}_2^\top\mathbf{V}_4\|_F \geq \frac{1}{\mathsf{poly}(\widetilde{\kappa})}\|\mathbf{E}_4\|_F$$

so we are also done. $\square$

Claim F.3. *Suppose* $\sigma_{\min}(\mathbf{K}_{\ell,\ell-1}) \geq \frac{1}{\widetilde{\kappa}}$ *and* $\|\mathbf{K}_\ell\|_2 \leq \widetilde{\kappa}$ *for some* $\widetilde{\kappa} \geq \kappa + k_\ell + L$*, we have*

$$\left\|2\mathbf{K}_{\ell,\ell-1}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1} - \left(\mathbf{V}_{\ell,\ell-1}^\star\right)^\top \mathbf{V}_{\ell,\ell-1}^\star\right\|_F \leq \mathsf{poly}(\widetilde{\kappa})\,\|\mathbf{E}_{\ell\triangleleft}\|_F$$
$$\text{and} \quad \|2\mathbf{Q}_\ell - \mathbf{I}\|_F \leq (\widetilde{\kappa})^2 \|\mathbf{E}_{\ell,\ell-1}\|_F$$

Proof of Claim F.3. Without loss of generality (by applying a unitary transformation), let us write $\mathbf{K}_{\ell,\ell-1} = (\mathbf{K}, \mathbf{0})$ for square matrix $\mathbf{K} \in \mathbb{R}^{k_\ell \times k_\ell}$, and let us write $\mathbf{V}_{\ell,\ell-1}^\star = (\mathbf{V}_1, \mathbf{V}_2)$ for square matrix $\mathbf{V}_1 \in \mathbb{R}^{k_\ell \times k_\ell}$. From (F.2), we have

$$\|\mathbf{V}_2\|_F \leq \frac{\|\mathbf{E}_{\ell\triangleleft}\|_F}{\sigma_{\min}(\mathbf{V}_{\ell\triangleleft}^\star)} \leq \mathsf{poly}(k_\ell, \kappa, L) \cdot \|\mathbf{E}_{\ell\triangleleft}\|_F \ .$$

From the definition of $\mathbf{Q}_\ell$ we have

$$2\mathbf{Q}_\ell = (\mathbf{K}\mathbf{K}^\top)^{-1}(\mathbf{K}, \mathbf{0})(\mathbf{V}_1, \mathbf{V}_2)^\top(\mathbf{V}_1, \mathbf{V}_2)(\mathbf{K}, \mathbf{0})^\top(\mathbf{K}\mathbf{K}^\top)^{-1} = \mathbf{K}^{-\top}\mathbf{V}_1^\top\mathbf{V}_1\mathbf{K}^{-1} \tag{F.3}$$

It is easy to verify that

$$2\mathbf{K}_{\ell,\ell-1}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1} - \left(\mathbf{V}_{\ell,\ell-1}^\star\right)^\top \mathbf{V}_{\ell,\ell-1}^\star = \begin{pmatrix} \mathbf{V}_1^\top\mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} - \left(\mathbf{V}_{\ell,\ell-1}^\star\right)^\top \mathbf{V}_{\ell,\ell-1}^\star = \begin{pmatrix} \mathbf{0} & \mathbf{V}_1^\top\mathbf{V}_2 \\ \mathbf{V}_2^\top\mathbf{V}_1 & \mathbf{V}_2^\top\mathbf{V}_2 \end{pmatrix}$$

which shows that

$$\left\|2\mathbf{K}_{\ell,\ell-1}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1} - \left(\mathbf{V}_{\ell,\ell-1}^\star\right)^\top \mathbf{V}_{\ell,\ell-1}^\star\right\|_F \leq 2\|\mathbf{V}_1\|_F\|\mathbf{V}_2\|_F + \|\mathbf{V}_2\|_F^2 \leq \mathsf{poly}(\widetilde{\kappa}) \cdot \|\mathbf{E}_{\ell\triangleleft}\|_F \ .$$

Next, we consider $\|2\mathbf{Q}_\ell - \mathbf{I}\|_F^2$, since

$$\|\mathbf{K}^\top\mathbf{K} - \mathbf{V}_1^\top\mathbf{V}_1\|_F \leq \left\|\mathbf{K}_{\ell,\ell-1}^\top\mathbf{K}_{\ell,\ell-1} - \left(\mathbf{V}_{\ell,\ell-1}^\star\right)^\top \mathbf{V}_{\ell,\ell-1}^\star\right\|_F = \|\mathbf{E}_{\ell,\ell-1}\|_F \ ,$$

we immediately have

$$\|2\mathbf{Q}_\ell - \mathbf{I}\|_F \leq \frac{1}{\sigma_{\min}(\mathbf{K})^2}\|\mathbf{K}^\top\mathbf{K} - \mathbf{V}_1^\top\mathbf{V}_1\|_F \leq (\widetilde{\kappa})^2\|\mathbf{E}_{\ell,\ell-1}\|_F \ .$$

□

F.2 Frobenius Norm Updates

Consider the F-norm regularizers given by

$$\begin{aligned}
\mathbf{R}_{6,\ell} &= \|\mathbf{K}_\ell\|_F^2 = \mathbf{Tr}(\mathbf{K}_\ell^\top\mathbf{K}_\ell) \\
&= \mathbf{Tr}(\mathbf{K}_{\ell,\ell-1}^\top\mathbf{K}_{\ell,\ell-1}) + 2\mathbf{Tr}(\mathbf{K}_{\ell,\ell-1}^\top\mathbf{K}_{\ell\triangleleft}) + \mathbf{Tr}(\mathbf{K}_{\ell\triangleleft}^\top\mathbf{K}_{\ell\triangleleft}) \\
\mathbf{R}_{7,\ell} &= \|\mathbf{W}_\ell\|_F^2 = \mathbf{Tr}(\mathbf{W}_\ell^\top\mathbf{W}_\ell)
\end{aligned}$$

Lemma F.4. *Suppose for some parameter* $\widetilde{\kappa}_\ell \geq \kappa + L + k_\ell$ *it satisfies*

$$\sigma_{\min}(\mathbf{K}_{\ell,\ell-1}) \geq \frac{1}{2\widetilde{\kappa}_\ell} \text{ and } \|\mathbf{K}_\ell\|_2 \leq 2\widetilde{\kappa}_\ell,\ \eta_1, \eta_2 < \frac{1}{\mathsf{poly}(\widetilde{\kappa}_\ell)}, \text{ and } \|\mathbf{E}_{\ell\triangleleft}\|_F \leq \frac{1}{(2\widetilde{\kappa}_\ell)^2}$$

then

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left[\mathbf{R}_{7,\ell}^{(\mathsf{new})}\right] &\leq (1-\eta_1)\mathbf{R}_{7,\ell} + \eta_1 \cdot \mathsf{poly}(k_\ell, L, \kappa) \\
\mathbf{R}_{6,\ell}^{(\mathsf{new})} &\leq (1-\eta_1)\mathbf{R}_{6,\ell} + \eta_1 \cdot \mathsf{poly}(k_\ell, \kappa, L) + (\eta_1^2 + \eta_2\|\mathbf{E}_{\ell\triangleleft}\|_F) \cdot \mathsf{poly}(\widetilde{\kappa}_\ell)
\end{aligned}$$

Proof of Lemma F.4. Our updates satisfy

$$\begin{aligned}
\mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1} &\leftarrow (1-\eta_1)\mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1} + 2\eta_1 \mathbf{K}_{\ell,\ell-1}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1} + \xi_1 \\
\mathbf{K}_{\ell\triangleleft}^\top \mathbf{K}_{\ell\triangleleft} &\leftarrow (1+\eta_1)\mathbf{K}_{\ell\triangleleft}^\top \mathbf{K}_{\ell\triangleleft} - 2\eta_1 \mathbf{K}_{\ell\triangleleft}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell\triangleleft} + \xi_2 \\
\mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell\triangleleft} &\leftarrow \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell\triangleleft} + \xi_3 \\
\mathbf{W}_\ell^\top \mathbf{W}_\ell &\leftarrow (1-\eta_1)(\mathbf{W}_\ell)^\top \mathbf{W}_\ell + \eta_1 (\mathbf{V}_\ell^\star)^\top \mathbf{V}_\ell^\star + \sqrt{\eta_1}\xi_4
\end{aligned}$$

where error matrices $\|\xi_1\|_F, \|\xi_2\|_F, \|\xi_3\|_F \leq (\eta_1^2 + \eta_2\|\mathbf{E}_{\ell\triangleleft}\|_F)\cdot \mathsf{poly}(\widetilde{\kappa}_\ell)$ and $\mathbb{E}_{\mathbf{D}_\ell}[\xi_4] = 0$. The $\mathbf{R}_{7,\ell}$ part is now trivial and the $\mathbf{R}_{6,\ell}$ part is a direct corollary of Claim F.5. $\square$

Claim F.5. *The following is always true*

$$\mathbf{Tr}\left(-\mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1} + 2\mathbf{K}_{\ell,\ell-1}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1}\right) \leq -\|\mathbf{K}_{\ell,\ell-1}\|_F^2 + O\left(k_\ell^2 \kappa^2\right)$$

Furthermore, suppose $\sigma_{\min}(\mathbf{K}_{\ell,\ell-1}) \geq \frac{1}{2\widetilde{\kappa}_\ell}$ and $\|\mathbf{K}_\ell\|_2 \leq 2\widetilde{\kappa}_\ell$ for $\widetilde{\kappa}_\ell \geq \kappa + L + k_\ell$, we have that as long as $\|\mathbf{E}_{\ell\triangleleft}\|_F \leq \frac{1}{(2\widetilde{\kappa}_\ell)^2}$ then

$$\mathbf{Tr}\left(\mathbf{K}_{\ell\triangleleft}^\top \mathbf{K}_{\ell\triangleleft} - 2\mathbf{K}_{\ell\triangleleft}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell\triangleleft}\right) \leq -\|\mathbf{K}_{\ell\triangleleft}\|_F^2 + O((L^2\kappa)^2 k_\ell)$$

Proof of Claim F.5. For the first bound, it is a direct corollary of the bound $\|2\mathbf{K}_{\ell,\ell-1}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1}\|_F \leq \mathsf{poly}(\kappa, L)$ (which can be easily verified from formulation (F.3)).

As for the second bound, let us assume without loss of generality (by left/right multiplying with a unitary matrix) that $\mathbf{K}_{\ell,\ell-1} = (\mathbf{K}_1, \mathbf{0})$ and $\mathbf{K}_{\ell\triangleleft} = (\mathbf{K}_2, \mathbf{0})$ for square matrices $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{k_\ell \times k_\ell}$. Let us write $\mathbf{V}_{\ell,\ell-1}^\star = (\mathbf{V}_1, \mathbf{V}_2)$ and $\mathbf{V}_{\ell\triangleleft}^\star = (\mathbf{V}_3, \mathbf{V}_4)$ for square matrices $\mathbf{V}_1, \mathbf{V}_3 \in \mathbb{R}^{k_\ell \times k_\ell}$. Then we have,

$$\mathbf{E}_{\ell\triangleleft} = \begin{pmatrix} \mathbf{E}_1 & \mathbf{E}_2 \\ \mathbf{E}_3 & \mathbf{E}_4 \end{pmatrix} = \begin{pmatrix} \mathbf{K}_1^\top \mathbf{K}_2 - \mathbf{V}_1^\top \mathbf{V}_3 & -\mathbf{V}_1^\top \mathbf{V}_4 \\ -\mathbf{V}_2^\top \mathbf{V}_3 & -\mathbf{V}_2^\top \mathbf{V}_4 \end{pmatrix}$$

We have

$$\begin{aligned}
\|\mathbf{K}_1^\top \mathbf{K}_2 - \mathbf{V}_1^\top \mathbf{V}_3\|_F \leq \|\mathbf{E}_{\ell\triangleleft}\|_F &\implies \|\mathbf{K}_2 - \mathbf{K}_1^{-\top}\mathbf{V}_1^\top \mathbf{V}_3\|_F \leq 2\widetilde{\kappa}_\ell \cdot \|\mathbf{E}_{\ell\triangleleft}\|_F \ . \\
&\implies \left\|\mathbf{K}_2\mathbf{K}_2^\top - \mathbf{K}_1^{-\top}\mathbf{V}_1^\top \mathbf{V}_3 \mathbf{V}_3^\top \mathbf{V}_1 \mathbf{K}_1^{-1}\right\|_F \leq (2\widetilde{\kappa}_\ell)^2 \cdot \|\mathbf{E}_{\ell\triangleleft}\|_F
\end{aligned}$$

Translating this into the spectral dominance formula (recalling $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive semi-definite), we have

$$\begin{aligned}
\mathbf{K}_2\mathbf{K}_2^\top &\preceq \mathbf{K}_1^{-\top}\mathbf{V}_1^\top \mathbf{V}_3 \mathbf{V}_3^\top \mathbf{V}_1 \mathbf{K}_1^{-1} + (2\widetilde{\kappa}_\ell)^2 \cdot \|\mathbf{E}_{\ell\triangleleft}\|_F \cdot \mathbf{I} \\
&\preceq (L^2\kappa)^2 \cdot \mathbf{K}_1^{-\top}\mathbf{V}_1^\top \mathbf{V}_1 \mathbf{K}_1^{-1} + (2\widetilde{\kappa}_\ell)^2 \cdot \|\mathbf{E}_{\ell\triangleleft}\|_F \cdot \mathbf{I} \qquad \text{(using } \|\mathbf{V}_{\ell\triangleleft}^\star\|_2 \leq L^2\kappa\text{)}
\end{aligned}$$

On the other hand, from (F.3) one can verify that

$$2\mathbf{K}_{\ell\triangleleft}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell\triangleleft} = \mathbf{K}_2^\top \mathbf{K}_1^{-\top}\mathbf{V}_1^\top \mathbf{V}_1 \mathbf{K}_1^{-1}\mathbf{K}_2$$

Combining the two formula above, we have

$$\begin{aligned}
2\mathbf{K}_{\ell\triangleleft}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell\triangleleft} &\succeq \frac{1}{(L^2\kappa)^2}\mathbf{K}_2^\top \mathbf{K}_2 \mathbf{K}_2^\top \mathbf{K}_2 - (2\widetilde{\kappa}_\ell)^2 \|\mathbf{E}_{\ell\triangleleft}\|_F \cdot \mathbf{K}_2^\top \mathbf{K}_2 \\
&\succeq 2\mathbf{K}_2^\top \mathbf{K}_2 - O((L^2\kappa)^2)\cdot \mathbf{I} \qquad \text{(using } \mathbf{A}^2 \succeq 2\mathbf{A} - \mathbf{I} \text{ for symmetric } \mathbf{A}\text{)}
\end{aligned}$$

Taking trace on both sides finish the proof. $\square$

F.3 Regularizer Updates

Let us consider three regularizer

$$\mathbf{R}_{3,\ell} = \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell\triangleleft} - \mathbf{W}_{\ell,\ell-1}^\top \mathbf{W}_{\ell\triangleleft}$$
$$\mathbf{R}_{4,\ell} = \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1} - \mathbf{W}_{\ell,\ell-1}^\top \mathbf{W}_{\ell,\ell-1}$$
$$\mathbf{R}_{5,\ell} = \mathbf{K}_{\ell}^\top \mathbf{K}_{\ell} - \mathbf{W}_{\ell}^\top \mathbf{W}_{\ell}$$

Lemma F.6. *Suppose for some parameter $\widetilde{\kappa} \geq \kappa + L + k_\ell$ it satisfies*

$$\sigma_{\min}(\mathbf{K}_{\ell,\ell-1}) \geq \frac{1}{2\widetilde{\kappa}},\ \sigma_{\min}(\mathbf{K}_{\ell\triangleleft}) \geq \frac{1}{2\widetilde{\kappa}},\ \|\mathbf{K}_\ell\|_2, \|\mathbf{W}_\ell\|_2 \leq 2\widetilde{\kappa},\ \eta_2 < \frac{1}{\mathsf{poly}(\widetilde{\kappa})},\ \eta_1 \leq \frac{\eta_2}{\mathsf{poly}(\widetilde{\kappa})}$$

then, suppose $\mathbf{Obj}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \leq \varepsilon^2$ and suppose Corollary E.3 holds for $L \geq \ell$, then

$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell} \left\| \mathbf{R}_{3,\ell}^{(\mathsf{new})} \right\|_F^2 \leq (1 - 1.8\eta_1) \|\mathbf{R}_{3,\ell}\|_F^2 + \eta_1^3 \cdot \mathsf{poly}(\widetilde{\kappa}) + (\eta_2 \frac{\varepsilon^2}{\alpha_\ell^2}) \cdot (D_\ell)^4 \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell} \left\| \mathbf{R}_{4,\ell}^{(\mathsf{new})} \right\|_F^2 \leq (1 - 1.8\eta_1) \|\mathbf{R}_{4,\ell}\|_F^2 + \eta_2 \frac{\varepsilon^2}{\alpha_\ell^2} \Upsilon_\ell \cdot (D_\ell)^6 \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell} \left\| \mathbf{R}_{5,\ell}^{(\mathsf{new})} \right\|_F^2 \leq (1 - 1.8\eta_1) \|\mathbf{R}_{5,\ell}\|_F^2 + \eta_2 \frac{\varepsilon^2}{\alpha_\ell^2} \Upsilon_\ell^2 \cdot (D_\ell)^{16} \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$

Proof of Lemma F.6. Let us check how these matrices get updated.

$$\mathbf{R}_{3,\ell} \leftarrow (1-\eta_1)\mathbf{R}_{3,\ell} + \eta_1 \mathbf{E}_{\ell\triangleleft} - \eta_2 \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1} \mathbf{E}_{\ell\triangleleft} - \eta_2 \mathbf{E}_{\ell\triangleleft} \mathbf{K}_{\ell\triangleleft}^\top \mathbf{K}_{\ell\triangleleft} + \xi_3 + \zeta_3$$
$$\text{(using } \mathbf{E}_{\ell\triangleleft} = \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell\triangleleft} - (\mathbf{V^\star}_{\ell,\ell-1})^\top \mathbf{V^\star}_{\ell\triangleleft})$$
$$\mathbf{R}_{4,\ell} \leftarrow (1-\eta_1)\mathbf{R}_{4,\ell} + \eta_1 \left(2\mathbf{K}_{\ell,\ell-1}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1} - (\mathbf{V^\star}_{\ell,\ell-1})^\top \mathbf{V^\star}_{\ell,\ell-1}\right) + \xi_4 + \zeta_4$$
$$\mathbf{R}_{5,\ell} \leftarrow (1-\eta_1)\mathbf{R}_{5,\ell} + \eta_1 \left(\mathbf{K}_\ell^\top \mathbf{K}_\ell - (\mathbf{V}_\ell^\star)^\top \mathbf{V}_\ell^\star\right) - \eta_1 \mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell,\ell-1} + 2\eta_1 \mathbf{K}_{\ell,\ell-1}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell,\ell-1}$$
$$+ \eta_1 \mathbf{K}_{\ell\triangleleft}^\top \mathbf{K}_{\ell\triangleleft} - 2\eta_1 \mathbf{K}_{\ell\triangleleft}^\top \mathbf{Q}_\ell \mathbf{K}_{\ell\triangleleft} + \xi_5 + \zeta_5$$

where error matrices $\mathbb{E}_{\mathbf{D}_\ell}[\zeta_3] = 0, \mathbb{E}_{\mathbf{D}_\ell}[\zeta_4] = 0, \mathbb{E}_{\mathbf{D}_\ell}[\zeta_5] = 0$ and

$$\|\xi_3\|_F \leq (\eta_1^2 + \eta_2^2 \|\mathbf{E}_{\ell\triangleleft}\|_F^2) \cdot \mathsf{poly}(\widetilde{\kappa})$$
$$\|\xi_4\|_F, \|\xi_5\|_F \leq (\eta_1^2 + \eta_2 \|\mathbf{E}_{\ell\triangleleft}\|_F) \cdot \mathsf{poly}(\widetilde{\kappa})$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell} \|\zeta_3\|_F^2, \mathop{\mathbb{E}}_{\mathbf{D}_\ell} \|\zeta_4\|_F^2, \mathop{\mathbb{E}}_{\mathbf{D}_\ell} \|\zeta_5\|_F^2 \leq \frac{\eta_1}{m} \cdot \mathsf{poly}(\widetilde{\kappa})$$

The update on $\mathbf{R}_{3,\ell}$ now tells us (by applying Claim F.2)

$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell} \left\| \mathbf{R}_{3,\ell}^{(\mathsf{new})} \right\|_F^2 \leq (1 - 2\eta_1) \|\mathbf{R}_{3,\ell}\|_F^2 + 2\eta_1 \|\mathbf{R}_{3,\ell}\|_F \|\mathbf{E}_{\ell\triangleleft}\|_F - \frac{\eta_2}{\mathsf{poly}(\widetilde{\kappa})} \|\mathbf{E}_{\ell\triangleleft}\|_F^2$$
$$+ \eta_2 \mathsf{poly}(\widetilde{\kappa}) \left\| \mathbf{W}_\ell^\top \mathbf{W}_{\ell,\ell-1} - (\mathbf{V}_\ell^\star)^\top \mathbf{V}_{\ell,\ell-1}^\star \right\|_F \|\mathbf{E}_{\ell\triangleleft}\|_F$$
$$+ (\eta_1^2 \|\mathbf{R}_{3,\ell}\|_F + \eta_1^2 \|\mathbf{E}_{\ell\triangleleft}\|_F + \eta_2^2 \|\mathbf{E}_{\ell\triangleleft}\|_F^2 + \frac{\eta_1}{m}) \cdot \mathsf{poly}(\widetilde{\kappa})$$

As for $\mathbf{R}_{4,\ell}$ and $\mathbf{R}_{5,\ell}$, applying Claim F.3 and using the notation $\widehat{\mathbf{E}}_\ell = \mathbf{K}_\ell^\top \mathbf{K}_\ell - (\mathbf{V}_\ell^\star)^\top \mathbf{V}_\ell^\star$, we can further simplify them to

$$\mathbf{R}_{4,\ell} \leftarrow (1-\eta_1)\mathbf{R}_{4,\ell} + \xi_4' + \zeta_4 \qquad \text{for } \|\xi_4'\|_F \leq (\eta_1 \|\mathbf{E}_{\ell\triangleleft}\|_F + \eta_2 \|\mathbf{E}_{\ell\triangleleft}\|_F) \cdot \mathsf{poly}(\widetilde{\kappa})$$
$$\mathbf{R}_{5,\ell} \leftarrow (1-\eta_1)\mathbf{R}_{5,\ell} + \eta_1 \widehat{\mathbf{E}}_\ell + \xi_5' + \zeta_5 \qquad \text{for } \|\xi_5'\|_F \leq (\eta_1 \|\mathbf{E}_\ell\|_F + \eta_2 \|\mathbf{E}_{\ell\triangleleft}\|_F) \cdot \mathsf{poly}(\widetilde{\kappa})$$

As a result,

$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left\|\mathbf{R}_{4,\ell}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.9\eta_1)\|\mathbf{R}_{4,\ell}\|_F^2 + \|\mathbf{R}_{4,\ell}\|_F \cdot (\eta_1\|\mathbf{E}_{\ell\triangleleft}\|_F + \eta_2\|\mathbf{E}_{\ell\triangleleft}\|_F + \frac{\eta_1}{m})\cdot \mathsf{poly}(\widetilde{\kappa})$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left\|\mathbf{R}_{5,\ell}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.9\eta_1)\|\mathbf{R}_{5,\ell}\|_F^2 + \|\mathbf{R}_{5,\ell}\|_F \cdot (\eta_1\|\widehat{\mathbf{E}}_{\ell}\|_F + \eta_2\|\mathbf{E}_{\ell\triangleleft}\|_F + \frac{\eta_1}{m})\cdot \mathsf{poly}(\widetilde{\kappa})$$

Since $\mathbf{Obj} = \varepsilon^2$, by applying Corollary E.3, we have

$$\begin{aligned}
&\text{Corollary E.3a}: && \|\mathbf{W}_\ell^\top \mathbf{W}_{\ell,\ell-1} - (\mathbf{V}_\ell^\star)^\top \mathbf{V}_{\ell,\ell-1}^\star\|_F^2 \leq \left(\frac{\varepsilon}{\alpha_\ell}\right)^2 \cdot (D_\ell)^3 \cdot \frac{C_L}{C_\ell}\\
&\text{Corollary E.3b}: && \|\mathbf{E}_{\ell\triangleleft}\|_F^2 = \|\mathbf{K}_{\ell,\ell-1}^\top \mathbf{K}_{\ell\triangleleft} - \left(\mathbf{V}_{\ell,\ell-1}^\star\right)^\top \mathbf{V}_{\ell\triangleleft}^\star\|_F^2 \leq \left(\frac{\varepsilon}{\alpha_\ell}\right)^2 \cdot (D_\ell)^5 \Upsilon_\ell \cdot \frac{C_L}{C_\ell}\\
&\text{Corollary E.3c}: && \|\widehat{\mathbf{E}}_\ell\|_F^2 = \|\mathbf{K}_\ell^\top \mathbf{K}_\ell - (\mathbf{V}_\ell^\star)^\top \mathbf{V}_\ell^\star\|_F^2 \leq \left(\frac{\varepsilon}{\alpha_\ell}\right)^2 \cdot (D_\ell)^{15} \Upsilon_\ell^2 \cdot \frac{C_L}{C_\ell}
\end{aligned} \tag{F.4}$$

Plugging these into the bounds above, and using $\eta_2 \geq \eta_1 \cdot \mathsf{poly}(\widetilde{\kappa})$ and $\eta_2 \leq \frac{1}{\mathsf{poly}(\widetilde{\kappa})}$, and repeatedly using $2ab \leq a^2 + b^2$, we have

$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left\|\mathbf{R}_{3,\ell}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{3,\ell}\|_F^2 + \eta_1^3 \cdot \mathsf{poly}(\widetilde{\kappa}) + (\eta_2 \frac{\varepsilon^2}{\alpha_\ell^2}) \cdot (D_\ell)^4 \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left\|\mathbf{R}_{4,\ell}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{4,\ell}\|_F^2 + \eta_2 \frac{\varepsilon^2}{\alpha_\ell^2} \Upsilon_\ell \cdot (D_\ell)^6 \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left\|\mathbf{R}_{5,\ell}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{5,\ell}\|_F^2 + (\eta_1 \frac{\varepsilon^2}{\alpha_\ell^2} \Upsilon_\ell^2 + \eta_2 \frac{\varepsilon^2}{\alpha_\ell^2} \Upsilon_\ell) \cdot (D_\ell)^{16} \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$

□

Lemma F.7. *In the same setting as Lemma F.6, suppose the weaker Corollary E.4 holds for $L \geq \ell$ instead of Corollary E.3. Then, for every $\ell < L$,*

$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left\|\mathbf{R}_{3,\ell}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{3,\ell}\|_F^2 + \eta_1^3 \cdot \mathsf{poly}(\widetilde{\kappa}) + (\eta_2 \frac{\alpha_L^2 D_L}{\alpha_\ell^2}) \cdot (D_\ell)^4 \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left\|\mathbf{R}_{4,\ell}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{4,\ell}\|_F^2 + \eta_2 \frac{\alpha_L^2 D_L}{\alpha_\ell^2} \Upsilon_\ell \cdot (D_\ell)^6 \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_\ell}\left\|\mathbf{R}_{5,\ell}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{5,\ell}\|_F^2 + \eta_2 \frac{\alpha_L^2 D_L}{\alpha_\ell^2} \Upsilon_\ell^2 \cdot (D_\ell)^{16} \cdot \frac{C_L}{C_\ell} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_L}\left\|\mathbf{R}_{3,L}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{3,L}\|_F^2 + \eta_1^3 \cdot \mathsf{poly}(\widetilde{\kappa}) + (\eta_2 \frac{\varepsilon^2}{\alpha_L^2}) \cdot (D_L)^4 + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_L}\left\|\mathbf{R}_{4,L}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{4,L}\|_F^2 + \eta_2 \frac{\varepsilon^2}{\alpha_L^2} \Upsilon_L \cdot (D_L)^6 + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$
$$\mathop{\mathbb{E}}_{\mathbf{D}_L}\left\|\mathbf{R}_{5,L}^{(\mathsf{new})}\right\|_F^2 \leq (1-1.8\eta_1)\|\mathbf{R}_{5,L}\|_F^2 + \eta_2 \frac{\varepsilon^2}{\alpha_L^2} \Upsilon_L^2 \cdot (D_L)^{16} + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa})}{m}$$

Proof. Proof is identical to Lemma F.6 but replacing the use of Corollary E.3 with Corollary E.4.

□

F.4 Loss Function Update

For analysis purpose, let us denote by

$$\widetilde{\mathbf{Loss}}_{\leq \ell}(x;\mathbf{W},\mathbf{K}) \stackrel{\text{def}}{=} \Big(G^\star(x) - \sum_{j=2}^{\ell} \alpha_j \mathbf{Sum}(\widetilde{F}_j(x;\mathbf{W},\mathbf{K}))\Big)^2$$

$$\mathbf{Loss}_{\leq \ell}(x;\mathbf{W},\mathbf{K}) \stackrel{\text{def}}{=} \Big(G^\star(x) - \sum_{j=2}^{\ell} \alpha_j \mathbf{Sum}(F_j(x;\mathbf{W},\mathbf{K}))\Big)^2$$

$$\mathsf{OPT}_{\leq \ell} = \mathbb{E}_{x\sim\mathcal{D}}\left[\Big(G^\star(x) - \sum_{j=2}^{\ell} \alpha_j \mathbf{Sum}(G^\star_j(x))\Big)^2\right]$$

Lemma F.8. *Suppose the sampled set $\mathcal{Z}$ satisfies the event of Proposition C.2, Proposition C.8, Proposition C.7 (for $\varepsilon_{\mathsf{s}} \leq \varepsilon^2/100$). Suppose for some parameter $\widetilde{\kappa}_\ell \geq \kappa + L + \overline{k}_\ell$ and $\tau_\ell \geq \widetilde{\kappa}_\ell$ it satisfies*

$$\sigma_{\min}(\mathbf{K}_{\ell,\ell-1}) \geq \frac{1}{2\widetilde{\kappa}_\ell},\ \sigma_{\min}(\mathbf{K}_{\ell\triangleleft}) \geq \frac{1}{2\widetilde{\kappa}_\ell},\ \|\mathbf{K}_\ell\|_F, \|\mathbf{W}_\ell\|_F \leq \widetilde{\kappa}_\ell,\ \eta_2 < \frac{1}{\mathsf{poly}(\widetilde{\kappa})},\ \eta_1 \leq \frac{\eta_2}{\mathsf{poly}(\widetilde{\kappa})}$$

Suppose parameters are set to satisfy Definition A.4. Suppose the assumptions of Theorem E.1 hold for some $L = \ell - 1$, then for every constant $\gamma > 1$,

$$\begin{aligned}&\mathbb{E}_{\mathbf{D}}[\widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{W}^{(\mathsf{new})},\mathbf{K}^{(\mathsf{new})})]\\ &\leq (1-0.99\eta_1)\widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{W},\mathbf{K}) + \eta_1\Big(0.04\varepsilon^2 + \frac{\mathsf{poly}(\widetilde{\kappa},B')}{m} + (1+\frac{1}{\gamma})^2\mathsf{OPT}_{\leq \ell}\Big)\end{aligned}$$

Proof of Lemma F.8. Let us first focus on

$$\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{W},\mathbf{K})) = \|\mathbf{W}_j(\sigma(\mathbf{R}_{j-1}\widetilde{S}_{j-1}(x;\mathbf{K})),\dots)\|^2$$

and first consider only the movement of $\mathbf{W}$. Recall from Fact F.1 that

$$(\mathbf{W}_j^{(\mathsf{new})})^\top(\mathbf{W}_j^{(\mathsf{new})}) \leftarrow (1-\eta_1)(\mathbf{W}_j)^\top\mathbf{W}_j + \eta_1(\mathbf{V}^\star_j)^\top\mathbf{V}^\star_j + \sqrt{\eta_1}\xi_j$$

for some $\mathbb{E}_{\mathbf{D}}[\xi_j] = 0$ and $\mathbb{E}_{\mathbf{D}}[\|\xi_j\|_F^2] \leq \mathsf{poly}(\widetilde{\kappa}_j)/m$. Therefore,

$$\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K})) = (1-\eta_1)\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{W},\mathbf{K})) + \eta_1\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{V}^\star,\mathbf{K})) + \sqrt{\eta_1}\xi_{j,1} \quad \text{(F.5)}$$

for some $\xi_{j,1} = (\sigma(\mathbf{R}_{j-1}\widetilde{S}_{j-1}),\dots)^\top\xi(\sigma(\mathbf{R}_{j-1}\widetilde{S}_{j-1}),\dots)$ satisfying $\mathbb{E}[\xi_{j,1}] = 0$ and $|\xi_{j,1}| \leq (\mathsf{poly}(\widetilde{\kappa}_j,\overline{B}'_j) + \|x\|^2 + \|S_1(x)\|^2)\|\xi_j\|_F$. Therefore, for every x,

$$\begin{aligned}&\mathbb{E}_{\mathbf{D}}[\widetilde{\mathbf{Loss}}_{\leq \ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K})]\\
&= \mathbb{E}_{\mathbf{D}}\Big[\Big(G^\star(x) - (1-\eta_1)\sum_{j=2}^{\ell}\alpha_j\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{W},\mathbf{K})) - \eta_1\sum_{j=2}^{\ell}\alpha_j\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{V}^\star,\mathbf{K})) + \sum_{j=2}^{\ell}\alpha_j\sqrt{\eta_1}\xi_{j,1}\Big)^2\Big]\\
&\overset{①}{=} \Big(G^\star(x) - (1-\eta_1)\sum_{j=2}^{\ell}\alpha_j\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{W},\mathbf{K})) - \eta_1\sum_{j=2}^{\ell}\alpha_j\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{V}^\star,\mathbf{K}))\Big)^2 + \eta_1\mathbb{E}_{\mathbf{D}}\left[\sum_{j=2}^{\ell}\alpha_j^2\xi_{j,1}^2\right]\\
&\overset{②}{\leq} (1-\eta_1)\Big(G^\star(x) - \sum_{j=2}^{\ell}\alpha_j\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{W},\mathbf{K}))\Big)^2 + \eta_1\Big(G^\star(x) - \eta_1\sum_{j=2}^{\ell}\alpha_j\mathbf{Sum}(\widetilde{F}_j(x;\mathbf{V}^\star,\mathbf{K}))\Big)^2\\
&\quad + \eta_1\frac{\mathsf{poly}(\widetilde{\kappa},B')}{m}\end{aligned}$$

$$= (1-\eta_1)\widetilde{\mathbf{Loss}}_{\leq \ell}(x;\mathbf{W},\mathbf{K}) + \eta_1 \widetilde{\mathbf{Loss}}_{\leq \ell}(x;\mathbf{V}^\star,\mathbf{K}) + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{m}$$

Above, ① uses the fact that $\mathbb{E}_{\mathbf{D}}[\xi_{j,1}] = 0$ and the fact that $\xi_{j,1}$ and $\xi_{j,1}$ are independent for $j \neq j$; and ② uses $((1-\eta)a + \eta b)^2 \leq (1-\eta)a^2 + \eta b^2$, as well as the bound on $\mathbb{E}_{\mathbf{D}}[\|\xi_j\|_F^2]$ from Fact F.1.

Applying expectation with respect to $x \sim \mathcal{Z}$ on both sides, we have

$$\mathbb{E}_{\mathbf{D}}[\widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{W}^{(\mathsf{new})},\mathbf{K})] \leq (1-\eta_1)\widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{W},\mathbf{K}) + \eta_1 \widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{V}^\star,\mathbf{K}) + \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{m}$$

On the other hand, for the update in $\mathbf{K}_j$ in every $j < \ell$, we can apply $\|2\mathbf{Q}_j - \mathbf{I}\|_F \leq (\widetilde{\kappa}_j)^2 \|\mathbf{E}_{j,j-1}\|_F$ from Claim F.3 and apply the bounds in (F.4) to derive that (using our lower bound assumption on $\lambda_{3,j}, \lambda_{4,j}$ from Theorem E.1)

$$\|\mathbf{K}_j^{(\mathsf{new})} - \mathbf{K}_j\|_F \leq \eta_1 \|\mathbf{E}_j\|_F + \eta_2 \|\mathbf{E}_{j\triangleleft}\|_F \cdot \mathsf{poly}(\widetilde{\kappa}_j) \leq \frac{1}{\alpha_j}\big(\eta_1\varepsilon + \eta_2\varepsilon\big)\cdot (D_j)^8 \sqrt{\Upsilon_j^2} \cdot \frac{\sqrt{C_L}}{\sqrt{C_j}} \tag{F.6}$$

Putting this into Claim C.4 (for $L = \ell$), and using the gap assumption on $\frac{\alpha_{\ell+1}}{\alpha_\ell}$ from Definition A.4, we derive that

$$\begin{aligned}
&\widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{W}^{(\mathsf{new})},\mathbf{K}^{(\mathsf{new})}) \\
&\leq (1+0.01\eta_1)\widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{W}^{(\mathsf{new})},\mathbf{K}) + \eta_1 \frac{\varepsilon^2 \cdot \alpha_\ell^2}{\alpha_{\ell-1}^2}(D_{\ell-1})^{16}\Upsilon_{\ell-1}^2 \frac{C_L}{C_{\ell-1}} \\
&\leq (1+0.01\eta_1)\widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{W}^{(\mathsf{new})},\mathbf{K}) + \eta_1 \frac{\varepsilon^2}{100}
\end{aligned}$$

Finally, we calculate that

$$\begin{aligned}
\widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{Z};\mathbf{V}^\star,\mathbf{K}) &\overset{①}{\leq} \widetilde{\mathbf{Loss}}_{\leq \ell}(\mathcal{D};\mathbf{V}^\star,\mathbf{K}) + 0.01\varepsilon^2 \overset{②}{\leq} \big(1+\frac{1}{\gamma}\big)\mathbf{Loss}_{\leq \ell}(\mathcal{D};\mathbf{V}^\star,\mathbf{K}) + 0.02\varepsilon^2 \\
&\overset{③}{\leq} (1+\frac{1}{\gamma})^2 \mathsf{OPT}_{\leq \ell} + 0.03\varepsilon^2
\end{aligned} \tag{F.7}$$

where ① uses Proposition C.8 and $\gamma > 1$ is a constant, ② uses Claim C.1, and ③ uses Claim F.9 below. Combining all the inequalities we finish the proof. □

F.4.1 Auxiliary

Claim F.9. *Suppose parameters are set to satisfy Definition A.4, and the assumptions of Theorem E.1 hold for some $L = \ell - 1$. Then, for the $\mathbf{V}^\star = (\mathbf{V}_2^\star, \dots, \mathbf{V}_\ell^\star)$ that we constructed from (F.1), and suppose $\{\alpha_j\}_j$ satisfies the gap assumption from Definition A.4, it satisfies for every constant $\gamma > 1$,*

$$\mathbf{Loss}_{\leq \ell}(\mathcal{D};\mathbf{V}^\star,\mathbf{K}) \leq \frac{\varepsilon^2}{100} + (1+\frac{1}{\gamma})\mathsf{OPT}_{\leq \ell}$$

Proof. Recalling that

$$F(x;\mathbf{W},\mathbf{K}) = \sum_\ell \alpha_\ell \mathbf{Sum}(F_\ell(x)) = \sum_\ell \alpha_\ell \left\|\mathbf{W}_\ell(\sigma(\mathbf{R}_{\ell-1}S_{\ell-1}(x)),\dots)\right\|^2$$

Using the conclusion that for every $j < \ell$, $\mathbb{E}_{x\sim\mathcal{D}}\left\|\mathbf{U}_j S_j^\star(x) - S_j(x)\right\|_2^2 \leq \delta_j^2 \overset{\text{def}}{=} (D_j)^{18}\left(\frac{\varepsilon}{\alpha_j}\right)^2 \cdot \frac{C_L}{C_\ell}$ from Corollary E.3d, one can carefully verify that (using an analogous proof to (E.11)) for every $j \leq \ell$,

$$\left\|\mathbf{V}_j^\star(\sigma(\mathbf{R}_{j-1}\mathbf{U}_{j-1}S_{j-1}^\star(x)),\dots)\right\|^2 = \left\|\mathbf{V}_j^\star(\sigma(\mathbf{R}_{j-1}S_{j-1}(x)),\dots)\right\|^2 + \xi_j$$

for some

$$\mathbb{E}[(\xi_j)^2] \le \mathsf{poly}(\widetilde{\kappa}_j, B_j, c_3(2^j))\delta_{j-1}^2 \le D_j(D_{j-1})^{18}\left(\frac{\varepsilon}{\alpha_{j-1}}\right)^2 \cdot \frac{C_L}{C_j}$$

Since our definition of $\mathbf{V}^\star$ satisfies (F.1), we also have for every $j \le \ell$

$$\left\|\mathbf{V}_j^\star(\sigma(\mathbf{R}_{j-1}\mathbf{U}_{j-1}S_{j-1}^\star(x)), \dots)\right\|^2 = \mathbf{Sum}(G_j^\star(x))$$

Putting them together, and using the gap assumption on $\frac{\alpha_j}{\alpha_{j-1}}$ from Definition A.4,

$$\mathbb{E}_{x\sim\mathcal{D}}(\sum_{j=2}^{\ell}\alpha_j\mathbf{Sum}(F_j(x;\mathbf{V}^\star,\mathbf{K})) - \alpha_j\mathbf{Sum}(G_j^\star(x)))^2 \le L\sum_{j=2}^{\ell}\alpha_j^2 D_j(D_{j-1})^{19}\left(\frac{\varepsilon}{\alpha_{j-1}}\right)^2 \cdot \frac{C_L}{C_j} \le \frac{\varepsilon^2}{100(1+\gamma)} \enspace .$$

Finally, using Young's inequality that

$$\begin{aligned}\mathbf{Loss}_{\le\ell}(x;\mathbf{V}^\star,\mathbf{K}) \le (1+\frac{1}{\gamma})&\left(\sum_{\ell=2}^{\ell}\alpha_\ell\mathbf{Sum}(G_\ell^\star(x)) - G^\star(x)\right)^2 \\ &+ (1+\gamma)\left(\sum_{\ell=2}^{L}\alpha_\ell\mathbf{Sum}(F_\ell(x;\mathbf{V}^\star,\mathbf{K})) - \alpha_\ell\mathbf{Sum}(G_\ell^\star(x))\right)^2\end{aligned}$$

we finish the proof. □

F.5 Objective Decrease Direction: Stage $\ell^\triangle$

Theorem F.10. *Suppose we are in stage $\ell^\triangle$, meaning that $\lambda_{3,j} = \lambda_{4,j} = \lambda_{5,j} = 0$ for $j \ge \ell$ and the trainable parameters are $\mathbf{W}_1, \dots, \mathbf{W}_\ell, \mathbf{K}_1, \dots, \mathbf{K}_{\ell-1}$. Suppose it satisfies*

$$\varepsilon^2 \stackrel{\text{def}}{=} \widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K}) \le \left(\frac{\alpha_{\ell-1}}{(D_{\ell-1})^9\Upsilon_{\ell-1}}\right)^2 \quad \text{and} \quad \left\{\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|_2^2] \le \tau_j\right\}_{j<\ell}$$

Suppose the sampled set $\mathcal{Z}$ satisfies the event of Proposition C.2, Proposition C.8, Proposition C.7 (for $\varepsilon_{\mathsf{s}} \le \varepsilon^2/100$). Suppose parameters are set to satisfy Definition A.4. Then, for every $\eta_2 < \frac{1}{\mathsf{poly}(\widetilde{\kappa})}$ and $\eta_1 \le \frac{\eta_2}{\mathsf{poly}(\widetilde{\kappa})}$,

$$\mathbb{E}_{\mathbf{D}}\widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W}^{(\mathsf{new})},\mathbf{K}^{(\mathsf{new})}) \le (1-0.7\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K}) + 2\eta_1\alpha_{\ell+1}^2$$

And also we have $\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|^2] \le 2B_j$ for every $j < \ell$.

Proof of Theorem F.10. We first verify the prerequisites of many of the lemmas we need to invoke.

Prerequisite 1. Using $\lambda_{6,\ell} \ge \frac{\varepsilon^2}{(\widetilde{\kappa}_\ell)^2}$ and $\widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K}) \le \varepsilon^2$, we have

$$\|\mathbf{K}_\ell\|_F, \|\mathbf{W}_\ell\|_F \le \widetilde{\kappa}_\ell$$

which is a prerequisite for Lemma F.4, Lemma F.6, Lemma F.8 that we need to invoke.

Prerequisite 2. Applying Proposition C.7, we have

$$\widetilde{\mathbf{Loss}}(\mathcal{Z};\mathbf{W},\mathbf{K}) \le \varepsilon^2 \xRightarrow{\text{Proposition C.7}} \widetilde{\mathbf{Loss}}(\mathcal{D};\mathbf{W},\mathbf{K}) \le 2\varepsilon^2 \tag{F.8}$$

Since $\mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|^2] \le \tau_j$ for all $j < \ell$, we can apply Claim C.1 and get

$$\widetilde{\mathbf{Loss}}(\mathcal{D};\mathbf{W},\mathbf{K}) \le 2\varepsilon^2 \xRightarrow{\text{Claim C.1 and choice } B'} \mathbf{Loss}(\mathcal{D};\mathbf{W},\mathbf{K}) \le 3\varepsilon^2 \tag{F.9}$$

Next, consider a dummy loss function against only the first $\ell-1$ layers

$$\mathbf{Loss}_{dummy}(\mathcal{D};\mathbf{W},\mathbf{K}) \stackrel{\text{def}}{=} \sum_{x\sim\mathcal{D}} \Big[\Big(\sum_{j=2}^{\ell-1} \alpha_j \mathbf{Sum}(F_j(x)) - \alpha_j \mathbf{Sum}(G^\star_j(x))\Big)^2\Big] \le 1.1\mathbf{Loss}(\mathcal{D};\mathbf{W},\mathbf{K}) + O(\alpha_\ell^2) \le 4\varepsilon^2$$

so in the remainder of the proof we can safely apply Theorem E.1 and Corollary E.3 for $L=\ell-1$. Note that this is also a prerequisite for Lemma F.8 with ℓ layers that we want to invoke. As a side note, we can use Corollary E.3d to derive

$$\forall j<\ell: \qquad \mathbb{E}_{x\sim\mathcal{D}}[\|S_j(x)\|^2] \le 2B_j \ .$$

Prerequisite 3. Corollary E.3b tells us for every $j<\ell$,

$$\begin{aligned}\Big\|\mathbf{Q}_{j-1}^\top \overline{\mathbf{K}}_{j,j-1}^\top \overline{\mathbf{K}}_{j\triangleleft}\mathbf{Q}_{j\triangleleft} - \overline{\mathbf{W}^\star}_{j,j-1}^\top \overline{\mathbf{W}^\star}_{j\triangleleft}\Big\|_F^2 &\le \Upsilon_j (D_j)^4 \left(\frac{\varepsilon}{\alpha_j}\right)^2 \frac{C_\ell}{C_j} \\ &\stackrel{①}{\le} \frac{\Upsilon_j(D_j)^4}{\Upsilon_{\ell-1}^2(D_{\ell-1})^{18}}\left(\frac{\alpha_{\ell-1}}{\alpha_j}\right)^2 \frac{C_\ell}{C_j} \stackrel{②}{\le} \frac{1}{(D_j)^{14}}\end{aligned} \tag{F.10}$$

Above, inequality ① uses the assumption $\varepsilon \le \frac{\alpha_{\ell-1}}{(D_{\ell-1})^9 \Upsilon_{\ell-1}}$. Inequality ② holds when $j=\ell-1$ by using $\frac{1}{\Upsilon_{\ell-1}}\frac{C_\ell}{C_{\ell-1}} \ll 1$ from our sufficiently large choice of $\Upsilon_{\ell+1}$, and ineuqliaty ② holds when $j<\ell-1$ using the gap assumption on $\frac{\alpha_j}{\alpha_{j-1}}$ when $j<\ell-1$.

Note that the left hand side of (F.10) is identical to (since $\overline{\mathbf{K}}_{j,i}\mathbf{Q}_i = \mathbf{K}_{j,i}(\mathbf{R}_i\mathbf{U}_i * \mathbf{R}_i\mathbf{U}_i)$)

$$\Big\|\mathbf{A}\mathbf{K}_{j,j-1}^\top \mathbf{K}_{j\triangleleft}\mathbf{B} - \mathbf{C}(\mathbf{W}^\star_{j,j-1})^\top \mathbf{W}^\star_{j\triangleleft}\mathbf{D}\Big\|_F^2$$

for some well-conditioned sqaure matrices $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}$ with singular values between $[\frac{1}{\mathsf{poly}(\overline{k}_j,L)}, O(\mathsf{poly}(\overline{k}_j,L))]$ (see Lemma B.6 and Lemma B.5). Therefore, combining the facts that (1) $\mathbf{K}_{j,j-1}^\top\mathbf{K}_{j\triangleleft}$ and $(\mathbf{W}^\star_{j,j-1})^\top\mathbf{W}^\star_{j\triangleleft}$ are both of rank exactly k_j, (2) $\|\mathbf{K}_j\| \le \widetilde{\kappa}_j$, (3) minimal singular value $\sigma_{\min}(\mathbf{W}^\star_{j,i}) \ge 1/\kappa$, we must have

$$\sigma_{\min}(\mathbf{K}_{j,j-1}) \ge \frac{1}{\widetilde{\kappa}_j \cdot \mathsf{poly}(\overline{k}_j,\kappa,L)} \quad \text{and} \quad \sigma_{\min}(\mathbf{K}_{j\triangleleft}) \ge \frac{1}{\widetilde{\kappa}_j \cdot \mathsf{poly}(\overline{k}_j,\kappa,L)}$$

as otherwise this will contract to (F.10). This lower bound on the minimum singular value is a prerequisite for Lemma F.4, Lemma F.6 that we need to invoke.

Prerequisite 4. Using Corollary E.3b, we also have for every $j<\ell$ (see the calculation in (F.4))

$$\begin{aligned}\|\mathbf{E}_{j\triangleleft}\|_F^2 = \|\mathbf{K}_{j,j-1}^\top\mathbf{K}_{j\triangleleft} - \left(\mathbf{V}^\star_{j,j-1}\right)^\top \mathbf{V}^\star_{j\triangleleft}\|_F^2 &\le \left(\frac{\varepsilon}{\alpha_j}\right)^2 \Upsilon_j \cdot (D_j)^5 \cdot \frac{C_\ell}{C_j} \\ &\le \left(\frac{\alpha_{\ell-1}}{\alpha_j}\right)^2 \cdot \frac{\Upsilon_j (D_j)^5}{\Upsilon_{\ell-1}(D_{\ell-1})^{18}} \cdot \frac{C_\ell}{C_j} \le \frac{1}{(D_j)^{13}}\end{aligned}$$

which is a prerequisite for Lemma F.4 that we need to invoke.

Main Proof Begins. Now we are fully prepared and can begin the proof. In the language of this section, our objective

$$\begin{aligned}\widetilde{\mathbf{Obj}}(\mathcal{Z};\mathbf{W},\mathbf{K}) = \widetilde{\mathbf{Loss}}(\mathcal{Z};\mathbf{W},\mathbf{K}) &+ \sum_{j<\ell}\Big(\lambda_{3,j}\|\mathbf{R}_{3,j}\|_F^2 + \lambda_{4,j}\|\mathbf{R}_{4,j}\|_F^2 + \lambda_{5,j}\|\mathbf{R}_{5,j}\|_F^2 + \lambda_{6,j}\mathbf{R}_{6,j}\Big) \\ &+ \sum_{j\le\ell}\lambda_{6,j}\left(\mathbf{R}_{7,j}\right)\end{aligned}$$

We can apply Lemma F.4 to bound the decrease of $\mathbf{R}_{6,j}$ for $j < \ell$ and $\mathbf{R}_{7,j}$ for $j \leq \ell$, apply Lemma F.6 to bound the decrease of $\mathbf{R}_{3,j}, \mathbf{R}_{4,j}, \mathbf{R}_{5,j}$ for $j < \ell$, and apply Lemma F.8 to bound the decrease of $\widetilde{\mathbf{Loss}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})$ (with the choice $\mathsf{OPT}_{\leq \ell} \leq 2\alpha_{\ell+1}^2$). By combining all the lemmas, we have (using $\eta_2 = \eta_1/\mathsf{poly}(\widetilde{\kappa})$ and sufficiently small choice of η_1)

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{D}} \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}) \\
&\overset{①}{\leq} (1 - 0.9\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + \eta_1(\varepsilon_{sample} + \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{m}) + \eta_1 \sum_{j \leq \ell} \lambda_{6,j} \mathsf{poly}(k_j, L, \kappa) + 2\eta_1 \alpha_{\ell+1}^2 \\
&\qquad + \eta_1 \sum_{j<\ell} \left(\frac{1}{\Upsilon_j} + \frac{\Upsilon_j}{\Upsilon_j^2} + \frac{\Upsilon_j^2}{\Upsilon_j^3} \right) \varepsilon^2 (D_j)^4 \frac{C_\ell}{C_j} \\
&\overset{②}{\leq} (1 - 0.8\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + \eta_1(\varepsilon_{sample} + \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{m}) + \eta_1 \sum_{j \leq \ell} \lambda_{6,j} \mathsf{poly}(k_j, L, \kappa) + 2\eta_1 \alpha_{\ell+1}^2 \\
&\overset{③}{\leq} (1 - 0.7\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + 2\eta_1 \alpha_{\ell+1}^2
\end{aligned}
$$

Above, inequality ① uses our parameter choices that $\lambda_{3,j} = \frac{\alpha_j^2}{(D_j)\Upsilon_j}$, $\lambda_{4,j} = \frac{\alpha_j^2}{(D_j)^7 \Upsilon_j^2}$, and $\lambda_{5,j} = \frac{\alpha_j^2}{\Upsilon_j^3 (D_j)^{13}}$. Inequality ② uses our choices of Υ_j (see Definition A.4). Inequality ③ uses $m \geq \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{\varepsilon^2}$ from Definition A.4, $\varepsilon_{\mathsf{s}} \leq 0.01\varepsilon^2$, and $\lambda_{6,j} = \frac{\varepsilon^2}{\widetilde{\kappa}_j^2} \leq \frac{\varepsilon^2}{\mathsf{poly}(k_j, L, \kappa)}$ from Definition A.4. □

F.6 Objective Decrease Direction: Stage $\ell^\triangledown$

Theorem F.11. *Suppose we are in stage $\ell^\triangledown$, meaning that $\lambda_{3,j} = \lambda_{4,j} = \lambda_{5,j} = 0$ for $j > \ell$ and the trainable parameters are $\mathbf{W}_1, \ldots, \mathbf{W}_\ell, \mathbf{K}_1, \ldots, \mathbf{K}_\ell$. Suppose it satisfies*

$$\left(\frac{\alpha_\ell}{(D_\ell)^9 \Upsilon_\ell} \right)^2 \leq \varepsilon^2 \overset{\text{def}}{=} \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) \leq \left(\frac{\alpha_\ell}{(D_\ell)^3 \sqrt{\Upsilon_\ell}} \right)^2 \quad \text{and} \quad \left\{ \mathbb{E}_{x \sim \mathcal{D}} [\|S_j(x)\|_2^2] \leq \tau \right\}_{j<\ell}$$

Suppose the sampled set $\mathcal{Z}$ satisfies the event of Proposition C.2, Proposition C.8, Proposition C.7 (for $\varepsilon_{\mathsf{s}} \leq \varepsilon^2/100$). Suppose parameters are set to satisfy Definition A.4. Then, for every $\eta_2 < \frac{1}{\mathsf{poly}(\widetilde{\kappa})}$ and $\eta_1 \leq \frac{\eta_2}{\mathsf{poly}(\widetilde{\kappa})}$,

$$\mathbb{E}_{\mathbf{D}} \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}) \leq (1 - 0.7\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + 2\eta_1 \alpha_{\ell+1}^2$$

And also we have $\mathbb{E}_{x \sim \mathcal{D}}[\|S_j(x)\|^2] \leq 2B_j$ for every $j < \ell$. Furthermore, if $\varepsilon^2 \leq \left(\frac{\alpha_\ell}{(D_\ell)^9 \Upsilon_\ell} \right)^2$ then we also have $\mathbb{E}_{x \sim \mathcal{D}}[\|S_\ell(x)\|^2] \leq 2B_\ell$.

Proof of Theorem F.11. The proof is analogous to Theorem F.10 but with several changes.

Prerequisite 1. For analogous reasons, we have

$$\|\mathbf{K}_\ell\|_F, \|\mathbf{W}_\ell\|_F \leq \widetilde{\kappa}_\ell$$

which is a prerequisite for Lemma F.4, Lemma F.7, Lemma F.8 that we need to invoke.

Prerequisite 2. This time, we have $\varepsilon^2 \leq \frac{\alpha_\ell}{(D_\ell)^3 \sqrt{\Upsilon_\ell}}$. This means the weaker assumption of Corollary E.4 has been satisfied for $L = \ell$, and as a result Theorem E.1 and Corollary E.3 hold with $L = \ell - 1$. This is a prerequisite for Lemma F.8 with ℓ layers that we want to invoke. Note in particular, Corollary E.3d implies

$$\forall j < \ell: \qquad \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\|S_j(x)\|^2] \leq 2B_j \ .$$

Note also, if $\varepsilon^2 \leq \left(\frac{\alpha_\ell}{(D_\ell)^9 \Upsilon_\ell}\right)^2$, then Corollary E.3 holds with $L = \ell$, so we can invoke Corollary E.3e to derive the above bound for $j = \ell$.

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\|S_\ell(x)\|^2] \leq 2B_\ell$$

Prerequisite 3. Again using Corollary E.3b for $L = \ell - 1$, we can derive for all $j < \ell$

$$\sigma_{\min}(\mathbf{K}_{j,j-1}) \geq \frac{1}{\widetilde{\kappa}_j \cdot \mathsf{poly}(k_j, \kappa, L)} \quad \text{and} \quad \sigma_{\min}(\mathbf{K}_{j\lhd}) \geq \frac{1}{\widetilde{\kappa}_j \cdot \mathsf{poly}(k_j, \kappa, L)}$$

This time, one can also use Corollary E.4b with $L = \ell$ to derive that the above holds also for $j = \ell$. This is a prerequisite for Lemma F.4, Lemma F.7 that we need to invoke.

Prerequisite 4. Using Corollary E.3b, we also have for every $j < \ell$ (see the calculation in (F.4))

$$\|\mathbf{E}_{j\lhd}\|_F^2 = \|\mathbf{K}_{j,j-1}^\top \mathbf{K}_{j\lhd} - \left(\mathbf{V}^\star_{j,j-1}\right)^\top \mathbf{V}^\star_{j\lhd}\|_F^2 \leq \frac{1}{(D_j)^{13}}$$

This time, one can also use Corollary E.4b with $L = \ell$ to derive that the above holds also for $j = \ell$.

Main Proof Begins. Now we are fully prepared and can begin the proof. In the language of this section, our objective

$$\begin{aligned}\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) = \widetilde{\mathbf{Loss}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) &+ \sum_{j<\ell} \left(\lambda_{3,j} \|\mathbf{R}_{3,j}\|_F^2 + \lambda_{4,j} \|\mathbf{R}_{4,j}\|_F^2 + \lambda_{5,j} \|\mathbf{R}_{5,j}\|_F^2 + \lambda_{6,j}\mathbf{R}_{6,j}\right) \\ &+ \sum_{j\leq\ell} \lambda_{6,j} (\mathbf{R}_{7,j})\end{aligned}$$

We can apply Lemma F.4 to bound the decrease of $\mathbf{R}_{6,j}, \mathbf{R}_{7,j}$ for $j \leq \ell$, apply Lemma F.7 to bound the decrease of $\mathbf{R}_{3,j}, \mathbf{R}_{4,j}, \mathbf{R}_{5,j}$ for $j \leq \ell$, and apply Lemma F.8 to bound the decrease of $\widetilde{\mathbf{Loss}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})$ (with the choice $\mathsf{OPT}_{\leq \ell} \leq 2\alpha_{\ell+1}^2$). By combining all the lemmas, we have (using $\eta_2 = \eta_1/\mathsf{poly}(\widetilde{\kappa})$ and sufficiently small choice of η_1)

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{D}} \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}) \\
&\overset{①}{\leq} (1 - 0.9\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + \eta_1(\varepsilon_{sample} + \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{m}) + \eta_1 \sum_{j\leq\ell} \lambda_{6,j} \mathsf{poly}(k_j, L, \kappa) + 2\eta_1\alpha_{\ell+1}^2 \\
&\qquad + \eta_1 \left(\frac{1}{\Upsilon_\ell} + \frac{\Upsilon_\ell}{\Upsilon_\ell^2} + \frac{\Upsilon_\ell^2}{\Upsilon_\ell^3}\right) \varepsilon^2 (D_\ell)^4 + \eta_1 \sum_{j<\ell} \left(\frac{1}{\Upsilon_j} + \frac{\Upsilon_j}{\Upsilon_j^2} + \frac{\Upsilon_j^2}{\Upsilon_j^3}\right) (\alpha_\ell)^2 D_\ell (D_j)^4 \frac{C_\ell}{C_j} \\
&\overset{②}{\leq} (1 - 0.9\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + \eta_1(\varepsilon_{sample} + \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{m}) + \eta_1 \sum_{j\leq\ell} \lambda_{6,j} \mathsf{poly}(k_j, L, \kappa) + 2\eta_1\alpha_{\ell+1}^2 \\
&\qquad + \eta_1 \left(\frac{1}{\Upsilon_\ell} + \frac{\Upsilon_\ell}{\Upsilon_\ell^2} + \frac{\Upsilon_\ell^2}{\Upsilon_\ell^3}\right) \varepsilon^2 (D_\ell)^4 + \eta_1 \sum_{j<\ell} \left(\frac{1}{\Upsilon_j} + \frac{\Upsilon_j}{\Upsilon_j^2} + \frac{\Upsilon_j^2}{\Upsilon_j^3}\right) \varepsilon^2 (D_\ell)^{19} \Upsilon_\ell^2 (D_j)^4 \frac{C_\ell}{C_j} \\
&\overset{③}{\leq} (1 - 0.8\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + \eta_1(\varepsilon_{sample} + \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{m}) + \eta_1 \sum_{j\leq\ell} \lambda_{6,j} \mathsf{poly}(k_j, L, \kappa) + 2\eta_1\alpha_{\ell+1}^2 \\
&\overset{④}{\leq} (1 - 0.7\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + 2\eta_1\alpha_{\ell+1}^2
\end{aligned}$$

Above, inequality ① uses our parameter choices that $\lambda_{3,j} = \frac{\alpha_j^2}{(D_j)\Upsilon_j}$, $\lambda_{4,j} = \frac{\alpha_j^2}{(D_j)^7\Upsilon_j}$, and $\lambda_{5,j} = \frac{\alpha_j^2}{(D_j)^{13}}$. Inequality ② uses our assumption that $\varepsilon \geq \frac{\alpha_\ell}{(D_\ell)^9\Upsilon_\ell}$. Inequality ③ uses our choices of Υ_j (see Definition A.4). Inequality ④ uses $m \geq \frac{\mathsf{poly}(\widetilde{\kappa},B')}{\varepsilon^2}$ from Definition A.4, $\varepsilon_{\mathsf{s}} \leq 0.01\varepsilon^2$, and $\lambda_{6,j} = \frac{\varepsilon^2}{\widetilde{\kappa}_j^2} \leq \frac{\varepsilon^2}{\mathsf{poly}(k_j,L,\kappa)}$ from Definition A.4. □

G Extension to Classification

Let us assume without loss of generality that $\mathbf{Var}[G^\star(x)] = \frac{1}{C\cdot c_3(2^L)}$ for some sufficiently large constant $C > 1$. We have the following proposition that relates the ℓ_2 and cross entropy losses. (Proof see Appendix G.2.)

Proposition G.1. *For every function $F(x)$ and $\varepsilon \geq 0$, we have*

1. *If $F(x)$ is a polynomial of degree 2^L and $\mathbb{E}_{(x_0,x)\sim\mathcal{D}} \mathsf{CE}\left(Y(x_0,x), v(x_0+F(x))\right) \leq \varepsilon$ for some there $v \geq 0$, then*
$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left(F(x) - G^\star(x)\right)^2 = O(c_3(2^L)^2\varepsilon^2)$$
2. *If $\mathbb{E}_{x\sim\mathcal{D}} \left(F(x) - G^\star(x)\right)^2 \leq \varepsilon^2$ and $v \geq 0$, then*
$$\mathop{\mathbb{E}}_{(x_0,x)\sim\mathcal{D}} \mathsf{CE}\left(Y(x_0,x), v(x_0+F(x))\right) \leq O\left(v\varepsilon^2 + \frac{\log^2 v}{v}\right)$$

At a high level, when setting $v = \frac{1}{\varepsilon}$, Proposition G.1 implies, up to small factors such as $c_3(2^L)$ and $\log(1/\varepsilon)$, it satisfies
$$\ell_2\text{-loss} = \varepsilon^2 \Longleftrightarrow \text{cross-entropy loss} = \varepsilon$$
Therefore, applying SGD on the ℓ_2 loss (like we do in this paper) should behave very similarly to applying SGD on the cross-entropy loss.

Of course, to turn this into an actual rigorous proof, there are subtleties. Most notably, we cannot naively convert back and forth between cross-entropy and ℓ_2 losses for *every SGD step*, since doing so we losing a multiplicative factor per step, killing the objective decrease we obtain. Also, one has to deal with truncated activation vs. quadratic activation. In the next subsection, we sketch perhaps the simplest possible way to prove our classification theorem by reducing its proof to that of our ℓ_2 regression theorem.

G.1 Detail Sketch: Reduce the Proof to Regression

Let us use the same parameters in Definition A.4 with minor modifications:

- additionally require one $\log(1/\varepsilon)$ factor in the gap assumption $\frac{\alpha_{\ell+1}}{\alpha_\ell}$,[33]
- additionally require one $1/\varepsilon$ factor in the over-parameterization m, and
- additionally require one $\mathsf{poly}(d)$ factor in the sample complexity N.

[33] We need this log factor because there is a logarithmic factor loss when translating between cross-entropy and the ℓ_2 loss (see Lemma G.1). This log factor prevents us from working with extremely small $\varepsilon > 0$, and therefore we have required $\varepsilon > \frac{1}{d^{100\log d}}$ in the statement of Theorem 3.3.

Recall from Theorem F.10 and Theorem F.11 that the main technical statement for the convergence in the regression case was to construct some $\mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}$ satisfying

$$\mathbb{E}_{\mathbf{D}} \widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}) \leq (1 - 0.7\eta_1)\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + 2\eta_1 \alpha_{\ell+1}^2 \ .$$

We show that the same construction $\mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}$ also satisfies, denoting by $\varepsilon = \widetilde{\mathbf{Obj}}^{\mathsf{xE}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})$,

$$\mathbb{E}_{\mathbf{D}} \widetilde{\mathbf{Obj}}^{\mathsf{xE}}(\mathcal{Z}; \mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}) \leq (1 - 0.7\eta_1)\widetilde{\mathbf{Obj}}^{\mathsf{xE}}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + \eta_1 \cdot O(\frac{\log^2(1/\varepsilon)}{\varepsilon}) \cdot \alpha_{\ell+1}^2 \ . \tag{G.1}$$

This means the objective can sufficiently decrease at least until $\varepsilon \approx \alpha_{\ell+1} \cdot \log \frac{1}{\alpha_{\ell+1}}$ (or to arbitrarily small when $\ell = L$). The rest of the proof will simplify follow from here.

Quick Observation. Let us assume without loss of generality that $v = \frac{\log(1/\varepsilon)}{100\varepsilon}$ always holds.[34] Using an analogous argument to Proposition C.7 and Claim C.1, we also have

$$\widetilde{\mathbf{Obj}}^{\mathsf{xE}}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \leq 2\varepsilon \quad \text{and} \quad \mathbf{Obj}^{\mathsf{xE}}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \leq 3\varepsilon \ .$$

Applying Lemma G.1, we immediately know $\mathbf{Obj}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \leq O(c_3(2^L)^2\varepsilon^2)$ for the original ℓ_2 objective. Therefore, up to a small factor $c_3(2^L)^2$, the old inequality $\mathbf{Obj}(\mathcal{D}; \mathbf{W}, \mathbf{K}) \leq \varepsilon^2$ remains true. This ensures that we can still apply many of the technical lemmas (especially the critical Lemma E.1 and the regularizer update Lemma F.6).

Going back to (G.1). In order to show sufficient objective value decrease in (G.1), in principle one needs to look at loss function decrease as well as regularizer decrease. This is what we did in the proofs of Theorem F.10 and Theorem F.11 for the regression case.

Now for classification, the regularizer decrease *remains the same as before* since we are using the same regularizer. The only technical lemma that requires non-trivial changes is Lemma F.8 which talks about loss function decrease from $\mathbf{W}, \mathbf{K}$ to $\mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}$. As before, let us write for notational simplicity

$$\widetilde{F}_{\leq \ell}(x; \mathbf{W}, \mathbf{K}) \stackrel{\text{def}}{=} \sum_{j=2}^{\ell} \alpha_j \mathbf{Sum}(\widetilde{F}_j(x; \mathbf{W}, \mathbf{K}))$$

$$\widetilde{\mathbf{Loss}}^{\mathsf{xE}}_{\leq \ell}(x_0, x; \mathbf{W}, \mathbf{K}) \stackrel{\text{def}}{=} \mathsf{CE}(Y(x_0, x), v(x_0 + \widetilde{F}_{\leq \ell}(x; \mathbf{W}, \mathbf{K})))$$

One can show that the following holds (proved in Appendix G.1.1):

Lemma G.2 (classification variant of Lemma F.8)**.**

$$\begin{aligned}
&\mathbb{E}_{\mathbf{D}} \widetilde{\mathbf{Loss}}^{\mathsf{xE}}_{\leq \ell}(\mathcal{Z}; \mathbf{W}^{(\mathsf{new})}, \mathbf{K}^{(\mathsf{new})}) \\
&\leq (1 - \eta_1)\widetilde{\mathbf{Loss}}^{\mathsf{xE}}_{\leq \ell}(\mathcal{Z}; \mathbf{W}, \mathbf{K}) + \eta_1 \left(\frac{O(\log^2(1/\varepsilon))}{\varepsilon} \mathsf{OPT}_{\leq \ell} + 0.1\varepsilon + \frac{v^2 \cdot \mathsf{poly}(\widetilde{\kappa}, B')}{m} \right)
\end{aligned}$$

Combining this with the regularizer decrease lemmas, we arrive at (G.1).

[34] This can be done by setting $v = \frac{\log(1/\varepsilon_0)}{100\varepsilon_0}$ where ε_0 is the current target error in Algorithm 1. Since ε and ε_0 are up to a factor of at most 2, the equation $v = \frac{\log(1/\varepsilon)}{100\varepsilon}$ holds up to a constant factor. Also, whenever ε_0 shrinks by a factor of 2 in Algorithm 1, we also increase v accordingly. This is okay, since it increases the objective value $\widetilde{\mathbf{Obj}}(\mathcal{Z}; \mathbf{W}, \mathbf{K})$ by more than a constant factor.

G.1.1 Proof of Lemma G.2

Sketched proof of Lemma G.2. Let us rewrite

$$\widetilde{F}_{\le \ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K}^{(\mathsf{new})}) = (1-\eta_1)\widetilde{F}_{\le \ell}(x;\mathbf{W},\mathbf{K}) + \eta_1 H(x) + Q(x) \tag{G.2}$$

$$\text{for } H(x) \stackrel{\text{def}}{=} \frac{\widetilde{F}_{\le \ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K}^{(\mathsf{new})}) - \widetilde{F}_{\le \ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K})}{\eta_1} + \widetilde{F}_{\le \ell}(x;\mathbf{V}^\star,\mathbf{K})$$

$$\text{for } Q(x) \stackrel{\text{def}}{=} \widetilde{F}_{\le \ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K}) - \eta_1 \widetilde{F}_{\le \ell}(x;\mathbf{V}^\star,\mathbf{K}) - (1-\eta_1)\widetilde{F}_{\le \ell}(x;\mathbf{W},\mathbf{K})$$

We make two observations from here.

- First, we can calculate the ℓ_2 loss of the auxilary function $H(x)$. The original proof of Lemma F.8 can be modified to show the following (proof in Appendix G.1.2)

  Claim G.3. $\mathbb{E}_{x\sim\mathcal{D}}(G^\star(x) - H(x))^2 \le 0.00001\frac{\varepsilon^2}{\log^2(1/\varepsilon)} + 6\mathsf{OPT}_{\le \ell}$.

  Using Lemma G.1, and our choice of $v = \frac{100\log^2(1/\varepsilon)}{\varepsilon}$, we can connect this back to the cross entropy loss:

  $$\mathop{\mathbb{E}}_{(x_0,x)\sim\mathcal{D}} \mathsf{CE}\left(Y(x_0,x), v(x_0 + H(x))\right) \le \frac{O(\log^2(1/\varepsilon))}{\varepsilon}\mathsf{OPT}_{\le \ell} + 0.09\varepsilon$$

  Through a similar treatment to Proposition C.8 we can also translate this to the training set

  $$\mathop{\mathbb{E}}_{(x_0,x)\sim\mathcal{Z}} \mathsf{CE}\left(Y(x_0,x), v(x_0 + H(x))\right) \le \frac{O(\log^2(1/\varepsilon))}{\varepsilon}\mathsf{OPT}_{\le \ell} + 0.1\varepsilon \tag{G.3}$$

- Second, recall from (F.5) in the original proof of Lemma F.8 that we have

  $$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{D}}[(Q(x))^2] &= \mathop{\mathbb{E}}_{\mathbf{D}}\Big(\widetilde{F}_{\le \ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K}) - \eta_1 \widetilde{F}_{\le \ell}(x;\mathbf{V}^\star,\mathbf{K}) - (1-\eta_1)\widetilde{F}_{\le \ell}(x;\mathbf{W},\mathbf{K})\Big)^2 \\ &= \mathop{\mathbb{E}}_{\mathbf{D}}\Big(\sum_{j=2}^{\ell}\alpha_j \xi_{j,1}\Big)^2 \le \eta_1 \frac{\mathsf{poly}(\widetilde{\kappa}, B')}{m} \quad .\end{aligned} \tag{G.4}$$

  as well as $\mathbb{E}_{\mathbf{D}}[Q(x)] = 0$.

We are now ready to go back to (G.2), and apply convexity and the Lipscthiz smoothness of the cross-entropy loss function to derive:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{D}} \widetilde{\mathbf{Loss}}^{\mathsf{xE}}_{\le \ell}(\mathcal{Z};\mathbf{W}^{(\mathsf{new})},\mathbf{K}^{(\mathsf{new})}) &\le (1-\eta_1)\widetilde{\mathbf{Loss}}^{\mathsf{xE}}_{\le \ell}(\mathcal{Z};\mathbf{W},\mathbf{K}) + \eta_1 \mathop{\mathbb{E}}_{(x_0,x)\sim\mathcal{Z}}[\mathsf{CE}(Y(x_0,x), v(x_0 + H(x)))] \\ &\quad + v^2 \cdot \mathop{\mathbb{E}}_{\mathbf{D}}[(Q(x))^2]\end{aligned}$$

Plugging (G.3) and (G.4) into the above formula, we finish the proof. □

G.1.2 Proof of Claim G.3

Proof of Claim G.3. Let us write

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{Z}}(G^\star(x) - H(x))^2 &\le \frac{2}{(\eta_1)^2}\mathop{\mathbb{E}}_{x\sim\mathcal{Z}}\Big(\widetilde{F}_{\le \ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K}^{(\mathsf{new})}) - \widetilde{F}_{\le \ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K})\Big)^2 \\ &\quad + 2\mathop{\mathbb{E}}_{x\sim\mathcal{Z}}\Big(G^\star(x) - \widetilde{F}_{\le \ell}(x;\mathbf{V}^\star,\mathbf{K})\Big)^2\end{aligned}$$

- For the first term, the same analysis of Claim C.4 gives

$$\mathbb{E}_{x\sim\mathcal{Z}}\left(\widetilde{F}_{\leq\ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K}^{(\mathsf{new})})-\widetilde{F}_{\leq\ell}(x;\mathbf{W}^{(\mathsf{new})},\mathbf{K})\right)^2$$
$$\leq \alpha_\ell^2\mathsf{poly}(\widetilde{\kappa}_{\ell-1},\overline{B}'_{\ell-1})\|\mathbf{K}^{(\mathsf{new})}-\mathbf{K}\|_F^2\leq(\eta_1)^2\frac{\varepsilon^2}{1000000\log^2(1/\varepsilon)}$$

where the last inequality has used the upper bound on $\|\mathbf{K}_j^{(\mathsf{new})}-\mathbf{K}_j\|_F$ for $j<\ell$— see (F.6) in the original proof of Lemma F.8 — as well as the gap assumption on $\frac{\alpha_\ell}{\alpha_{\ell-1}}$ (with an additional $\log(1/\varepsilon)$ factor).

- For the second term, the original proof of Lemma F.8 — specifically (F.7) — already gives

$$\mathbb{E}_{x\sim\mathcal{Z}}\left(G^\star(x)-\widetilde{F}_{\leq\ell}(x;\mathbf{V}^\star,\mathbf{K})\right)^2=\widetilde{\mathbf{Loss}}_{\leq\ell}(\mathcal{Z};\mathbf{V}^\star,\mathbf{K})\leq(1+\frac{1}{\gamma})^2\mathsf{OPT}_{\leq\ell}+\frac{\varepsilon^2}{1000000\log^2(1/\varepsilon)}$$

where the additional $\log(1/\varepsilon)$ factor comes from the gap assumption on $\frac{\alpha_\ell}{\alpha_{\ell-1}}$.

Putting them together, and applying a similar treatment to Proposition C.7 to go from the training set $\mathcal{Z}$ to the population $\mathcal{D}$, we have the desired bound. □

G.2 Proof of Proposition G.1

Proof of Proposition G.1.

1. Suppose by way of contradiction that

$$\mathbb{E}_{x\sim\mathcal{D}}(F(x)-G^\star(x))^2=\Omega\left(c_3(2^L)^2\varepsilon^2\right)$$

Let us recall a simple probability fact. Given any random variable $X\geq 0$, it satisfies[35]

$$\mathbf{Pr}[X>\frac{1}{2}\sqrt{\mathbb{E}[X^2]}]\geq\frac{9}{16}\frac{(\mathbb{E}[X^2])^2}{\mathbb{E}[X^4]}$$

Let us plug in $X=|F(x)-G^\star(x)|$, so by the hyper-contractivity Property 5.4, with probability at least $\Omega\left(\frac{1}{c_3(2^L)}\right)$ over $x\sim\mathcal{D}$,

$$|F(x)-G^\star(x)|=\Omega(c_3(2^L)\varepsilon)$$

Also by the hyper-contractivity Property 5.4 and Markov's inequality, with probability at least $1-O\left(\frac{1}{c_3(2^L)}\right)$,

$$G^\star(x)\leq\mathbb{E}[G^\star(x)]+O(c_3(2^L))\cdot\sqrt{\mathbf{Var}[G^\star(x)]}\leq\mathbb{E}[G^\star(x)]+1$$

When the above two events over x both take place— this happens with probability $\Omega(\frac{1}{c_3(2^L)})$— we further have with probability at least $\Omega(c_3(2^L)\varepsilon)$ over x_0, it satisfies $\mathsf{sgn}(x_0+F(x))\neq\mathsf{sgn}(x_0+G^\star(x))=Y(x_0,x)$. This implies $\mathbb{E}_{(x_0,x)\sim\mathcal{D}}\mathsf{CE}\left(Y(x_0,x),v(x_0+F(x))\right)>\varepsilon$ using the definition of cross entropy, giving a contradiction.

[35] The proof is rather simple. Denote by $\mathbb{E}[X^2]=a^2$ and let $\mathcal{E}=\{X\geq\frac{1}{2}a\}$ and $p=\mathbf{Pr}\left[X\geq\frac{1}{2}a\right]$. Then, we have

$$a^2=\mathbb{E}[X^2]\leq\frac{1}{4}(1-p)a^2+p\,\mathbb{E}[X^2\mid\mathcal{E}]\leq\frac{1}{4}a^2+p\sqrt{\mathbb{E}[X^4\mid\mathcal{E}]}=\frac{1}{4}a^2+\sqrt{p}\sqrt{p\,\mathbb{E}[X^4\mid\mathcal{E}]}\leq\frac{1}{4}a^2+\sqrt{p}\sqrt{\mathbb{E}[X^4]}$$

2. By the Lipschitz continuity of the cross-entropy loss, we have that

$$\begin{aligned}\mathsf{CE}\left(Y(x_0,x), v(x_0+F(x))\right) &\leq \mathsf{CE}\left(Y(x_0,x), v(x_0+G^\star(x))\right) + O\left(v|G^\star(x)-F(x)|\right)\\ &\leq O\left(1+v|G^\star(x)-F(x)|\right)\end{aligned}$$

Now, for a fixed x, we know that if $x_0 \geq -G^\star(x) + |G^\star(x)-F(x)| + 10\frac{\log v}{v}$ or $x_0 \leq -G^\star(x) - |G^\star(x)-F(x)| - 10\frac{\log v}{v}$, then $\mathsf{CE}\left(Y(x_0,x), v(x_0+F(x))\right) \leq \frac{1}{v}$. This implies

$$\begin{aligned}&\mathop{\mathbb{E}}_{x_0} \mathsf{CE}\left(Y(x_0,x), v(x_0+F(x))\right)\\ &\leq \frac{1}{v} + \mathop{\mathbf{Pr}}_{x_0}\left[x_0 \in -G^\star(x) \pm \left(|G^\star(x)-F(x)| + 10\frac{\log v}{v}\right)\right] \times O\left(1+v|G^\star(x)-F(x)|\right)\\ &\leq \frac{1}{v} + \left(|G^\star(x)-F(x)| + 10\frac{\log v}{v}\right) \times O\left(1+v|G^\star(x)-F(x)|\right)\\ &\leq \frac{1}{v} + O\left(\log v \times |G^\star(x)-F(x)| + v|G^\star(x)-F(x)|^2 + \frac{\log v}{v}\right)\end{aligned}$$

Taking expectation over x we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{(x_0,x)\sim\mathcal{D}} \mathsf{CE}\left(Y(x_0,x), v(x_0+F(x))\right)\\ &\leq \frac{1}{v} + O\left(\log v \mathop{\mathbb{E}}_{x\sim\mathcal{D}}|G^\star(x)-F(x)| + v\mathop{\mathbb{E}}_{x\sim\mathcal{D}}|G^\star(x)-F(x)|^2 + \frac{\log v}{v}\right) \leq O(v\varepsilon^2 + \frac{\log^2 v}{v}) \ .\end{aligned}$$

□

H Lower Bounds for Kernels, Feature Mappings and Two-Layer Networks

H.1 Lower Bound: Kernel Methods and Feature Mappings

This subsection is a direct corollary of [2] with simple modifications.

We consider the following L-layer target network as a separating hard instance for any kernel method. Let us choose $k=1$ with each $\mathbf{W}^\star_{\ell,0}, \mathbf{W}^\star_{\ell,1} \in \mathbb{R}^d$ sampled i.i.d. uniformly at random from $\mathcal{S}_{2^{L-1}}$, and other $\mathbf{W}^\star_{\ell,j} = 1$. Here, the set $\mathcal{S}_p$ is given by:

$$\mathcal{S}_p = \left\{\forall w \in \mathbb{R}^d \mid \|w\|_0 = p, w_i \in \left\{0, \frac{1}{\sqrt{p}}\right\}\right\} \ .$$

We assume input x follows from the d-dimensional standard Gaussian distribution.

Recall Theorem 3.1 says that, for every d and $L = o(\log\log d)$, under appropriate gap assumptions for $\alpha_1, \dots, \alpha_L$, for every $\varepsilon > 0$, the neural network defined in our paper requires only $\mathsf{poly}(d/\varepsilon)$ time and samples to learn this target function $G^\star(x)$ up to accuracy ε.

In contrast, we show the following theorem of the sample complexity lower bound for kernel methods:

Theorem H.1 (kernel lower bound)**.** *For every $d > 1$, every $L \leq \frac{\log\log d}{100}$, every $\alpha_L < 0.1$, every (Mercer) kernels $K: \mathbb{R}^{d\times d} \to \mathbb{R}$, and $N \leq \frac{1}{1000}\binom{d}{2^{L-1}}$, for every N i.i.d. samples $x^{(1)}, \dots, x^{(N)} \sim \mathcal{N}(0,1)$, the following holds for at least 99% of the target functions $G^\star(x)$ in the aforementioned class (over the choice in $\mathcal{S}_p$). For all kernel regression functions*

$$\mathfrak{K}(x) = \textstyle\sum_{n\in[N]} K(x, x^{(n)}) \cdot v_n$$

where weights $v_i \in \mathbb{R}$ can depend on $\alpha_1, \cdots, \alpha_L, x^{(1)}, \ldots, x^{(N)}, K$ and the training labels $\{y^{(1)}, \cdots, y^{(N)}\}$, it must suffer population risk

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\mathbf{I}_{d\times d})} (G^\star(x) - \mathfrak{K}(x))^2 = \Omega(\alpha_L^2 \log^{-2^{L+2}}(d)) \enspace .$$

Remark H.2. Let us compare this to our positive result in Theorem 3.1 for $L = o(\log\log d)$. Recall from Section 3 that α_L can be as large as for instance $d^{-0.001}$ in order for Theorem 3.1 to hold. When this holds, neural network achieves for instance $1/d^{100}$ error with $\mathsf{poly}(d)$ samples and time complexity. In contrast, Theorem H.1 says, unless there are more than $\frac{1}{1000}\binom{d}{2^{L-1}} = d^{\omega(1)}$ samples, no kernel method can achieve a regression error of even $1/d^{0.01}$.

Sketch proof of Theorem H.1. The proof is almost a direct application of [2], and the main difference is that we have Gaussian input distribution here (in order to match the upper bound), and in [2] the input distribution is uniform over $\{-1,1\}^d$. We sketch the main ideas below.

First, randomly sample $|x_i|$ for each coordinate of x, then we have that $x_i = |x_i|\tau_i$ where each τ_i i.i.d. uniformly on $\{-1,1\}$. The target function $G^\star(x)$ can be re-written as $G^\star(x) = \widetilde{G^\star}(\tau)$ for $\tau = (\tau_i)_{i\in[d]} \in \{-1,1\}^d$, where $\widetilde{G^\star}(\tau)$ is a degree $p = 2^{L-1}$ polynomial over τ, of the form:

$$\widetilde{G^\star}(\tau) = \alpha_L \langle w, \tau\rangle^p + \widehat{G^\star}(\tau)$$

where (for $a \circ b$ being the coordinate product of two vectors)

$$w = \mathbf{W}^\star_{2,0} \circ |x| \quad \text{and} \quad \deg(\widehat{G^\star}(\tau)) \leq p-1$$

For every function f, let us write the Fourier Boolean decomposition of f:

$$f(\tau) = \sum_{\mathcal{S}\subset[d]} \lambda_{\mathcal{S}} \prod_{j\in\mathcal{S}} \tau_j$$

and for any fixed w, write the decomposition of $\widetilde{G^\star}(\tau)$:

$$\widetilde{G^\star}(\tau) = \sum_{\mathcal{S}\subset[d]} \lambda'_{\mathcal{S}} \prod_{j\in\mathcal{S}} \tau_j$$

Let us denote the set of p non-zero coordinates of $\mathbf{W}^\star_{2,0}$ as $\mathcal{S}_w$. Using basic Fourier analysis of boolean variables, we must have that conditioning on the ≥ 0.999 probability event that $\prod_{i\in\mathcal{S}_w}|x_i| \geq \left(\log^{0.9} d\right)^{-2^L}$, it satisfies

$$|\lambda'_{\mathcal{S}_w}| = \left(\frac{1}{\sqrt{p}}\right)^p \alpha_L \prod_{i\in\mathcal{S}_w}|x_i| \geq \left(\frac{1}{\sqrt{p}}\right)^p \alpha_L \left(\log^{0.9} d\right)^{-2^L} \geq \alpha_L \log^{-2^L}(d) \enspace .$$

Moreover, since $\deg(\widehat{G^\star}(\tau)) \leq p-1$, we must have $\lambda'_{\mathcal{S}} = 0$ for any other $\mathcal{S} \neq \mathcal{S}_w$ with $|\mathcal{S}| = p$. This implies that for any function $f(\tau)$ with

$$f(\tau) = \sum_{\mathcal{S}\subset[d]} \lambda_{\mathcal{S}} \prod_{j\in\mathcal{S}} \tau_j \quad \text{and} \quad \mathop{\mathbb{E}}_{\tau}\left(f(\tau) - \widetilde{G^\star}(\tau)\right)^2 = O(\alpha_L^2 \log^{-2^{L+2}}(d)) \enspace ,$$

it must satisfy

$$\lambda^2_{\mathcal{S}_w} = \Omega(\alpha_L^2 \log^{-2^{L+1}}(d)) > \sum_{\mathcal{S}\subseteq[d], |\mathcal{S}|=p, \mathcal{S}\neq\mathcal{S}_w} \lambda^2_{\mathcal{S}} = O(\alpha_L^2 \log^{-2^{L+2}}(d))$$

Finally, using $\mathbb{E}_{x\sim\mathcal{N}(0,\mathbf{I})}\left(G^\star(x) - \mathfrak{K}(x)\right)^2 = \mathbb{E}_{|x|}\mathbb{E}_\tau\left(\mathfrak{K}(|x|\circ\tau) - \widetilde{G^\star}(\tau)\right)^2$, we have with probability at least 0.999 over the choice of $|x|$, it holds that

$$\mathbb{E}_{\tau}\left(\mathfrak{K}(|x|\circ\tau)-\widetilde{G^\star}(\tau)\right)^2=O(\alpha_L^2\log^{-2^{L+2}}(d)) \ .$$

From here, we can select $f(\tau)=\mathfrak{K}(|x|\circ\tau)$. The rest of the proof is a direct application of [2, Lemma E.2] (as the input τ is now uniform over the Boolean cube $\{-1,1\}^d$). (The precise argument also uses the observation that if for >0.999 fraction of w, event $\mathcal{E}_w(x)$ holds for >0.999 fraction of x, then there is an x such that $\mathcal{E}_w(x)$ holds for >0.997 fraction of w.) □

For similar reason, we also have the number of features lower bound for linear regression over feature mappings:

Theorem H.3 (feature mapping lower bound). *For every $d>1$, every $L\leq\frac{\log\log d}{100}$, every $d\geq 0$, every $\alpha_L\leq 0.1$, every $D\leq\frac{1}{1000}\binom{d}{2^{L-1}}$, and every feature mapping $\phi\colon\mathbb{R}^d\to\mathbb{R}^D$, the following holds for at least 99% of the target functions $G^\star(x)$ in the aforementioned class (over the choice in $\mathcal{S}_p$). For all linear regression functions*

$$\mathfrak{F}(x)=w^\top\phi(x),$$

where weights $w\in\mathbb{R}^D$ can depend on $\alpha_1,\cdots,\alpha_L$ and ϕ, it must suffer population risk

$$\mathbb{E}_{x\sim\mathcal{N}(0,\mathbf{I})}\|G^\star(x)-\mathfrak{F}(x)\|_2^2=\Omega\left(\alpha_L^2\log^{-2^{L+2}}(d)\right) \ .$$

Remark H.4. In the same setting as Remark H.2, we see that neural network achieves for instance $1/d^{100}$ regression error with $\mathsf{poly}(d)$ time complexity, but to achieve even just $1/d^{0.01}$ error, Theorem H.3 says that any linear regression over feature mappings must use at least $D=d^{\omega(1)}$ features. This usually needs $\Omega(D)=d^{\omega(1)}$ time complexity.[36]

H.2 Lower Bound: Certain Two-Layer Polynomial Neural Networks

We also give a preliminary result separating our positive result (for L-layer quadratic DenseNet) from two-layer neural networks with polynomial activations (of degree 2^L). The lower bound relies on the following technical lemma which holds for some absolute constant $C>1$:

Lemma H.5. *For $1\leq d_1\leq d$, consider inputs (x,y) where $x\in\mathbb{R}^{d_1}$ follows from $\mathcal{N}(0,\mathbf{I}_{d_1\times d_1})$ and $y\in\mathbb{R}^{d-d_1}$ follows from an arbitrary distribution independent of x. We have that for every $p\geq 1$,*

- *for every function $f(x,y)=\left(\frac{\|x\|_4^4}{d_1}\right)^p+g(x,y)$ where $g(x,y)$ is a polynomial and its degree over x is at most $4p-1$, and*
- *for every function $h(x,y)=\sum_{i=1}^r a_i\widetilde{\sigma}_i(\langle w_i,(x,x^2,y)+b_i\rangle)$ with $r=\frac{1}{C}(d_1/p)^p$ and each $\widetilde{\sigma}_i$ is an arbitrary polynomial of maximum degree $2p$,*

it must satisfy $\mathbb{E}_{x,y}(h(x,y)-f(x,y))^2\geq\frac{1}{p^{C\cdot p}}$.

Before we prove Lemma H.5 in Section H.2.1, let us quickly point out how it gives our lower bound theorem. We can for instance consider target functions with $k_2=d$, $k_3=\cdots=k_L=1$, $\mathbf{W}^\star_{2,1}=\mathbf{I}_{d\times d}$ and $\mathbf{W}^\star_{\ell,0},\mathbf{W}^\star_{\ell,1},\mathbf{W}^\star_{\ell,2}=\left(\frac{1}{\sqrt{d}},\cdots,\frac{1}{\sqrt{d}}\right)$, and other $\mathbf{W}^\star_{\ell,j}=1$ for $j>2$.

[36] One might argue that feature mapping can be implemented to run faster than $O(D)$ time. However, those algorithms are very complicated and may require a lot of work to design. It can be unfair to compare to them for a "silly" reason. One can for instance cheat by defining an infinitely-large feature mapping where each feature corresponds to a different neural network; then, one can train a neural network and just set the weight of the feature mapping corresponding to the final network to be 1. Therefore, we would tend to assume that a linear regression over feature mapping requires at least $\Omega(D)$ running time to implement, where D is the total number of features.

For such target functions, when $L = o(\log\log d)$, our positive result Theorem 3.1 shows that the (hierarchical) DenseNet learner considered in our paper only need $\mathsf{poly}(d/\varepsilon)$ time and sample complexity to learn it to an arbitrary $\varepsilon > 0$ error (where the degree of the $\mathsf{poly}(d/\varepsilon)$ does not depend on L).

On the other hand, since the aforementioned target $G^\star(x)$ can be written in the form $\alpha_L\left(\frac{\|x\|_4^4}{d_1}\right)^{2^{L-2}} + g(x)$ for some $g(x)$ of degree at most $2^L - 1$, Lemma H.5 directly implies the following:

Theorem H.6. *For any two-layer neural network of form $h(x) = \sum_{i=1}^r a_i\widetilde{\sigma}_i(\langle w_i, (x, S_1(x)) + b_i\rangle)$, with $r \le d^{2^{o(L)}}$ and each $\widetilde{\sigma}_i$ is any polynomial of maximum degree 2^{L-1}, we have that*

$$\mathbb{E}_{x\sim\mathcal{N}(0,\mathbf{I})}(h(x) - G^\star(x))^2 \ge \frac{\alpha_L^2}{2^{2^{O(L)}}} \enspace .$$

(Since $\widetilde{\sigma}_i$ is degree 2^{L-1} over $S_1(x)$, the final degree of $h(x)$ is 2^L in x; this is the same as our L-layer DenseNet in the positive result.)

To compare this with the upper bound, let us recall again (see Section 3) that when $L = o(\log\log d)$, parameter α_L can be as large as for instance $d^{-0.001}$ in order for Theorem 3.1 to hold. When this holds, neural network achieves for instance $1/d^{100}$ error with $\mathsf{poly}(d)$ samples and time complexity. In contrast, Theorem H.1 says, unless there are more than $d^{2^{\Omega(L)}} = d^{\omega(1)}$ neurons, the two-layer polynomial network cannot achieve regression error of even $1/d^{0.01}$. To conclude, the hierarchical neural network can learn this function class more efficiently.

Finally, we also remark here after some simple modifications to Lemma H.5, we can also obtain the following theorem when $k_2 = k_3 = \cdots = k_L = 1$, $\mathbf{W}^\star_{\ell,1}, \mathbf{W}^\star_{\ell,0} = \left(\frac{1}{\sqrt{d}}, \cdots, \frac{1}{\sqrt{d}}\right)$ and other $\mathbf{W}^\star_{\ell,j} = 1$.

Theorem H.7. *For every function of form $h(x) = \sum_{i=1}^r a_i\widetilde{\sigma}'_i(\langle w_i, x + b_i\rangle)$ with $r \le d^{2^{o(L)}}$ and each $\widetilde{\sigma}'_i$ is any polynomial of maximum degree 2^L, we have*

$$\mathbb{E}_{x\sim\mathcal{N}(0,\mathbf{I})}(h(x) - G^\star(x))^2 \ge \frac{\alpha_L^2}{2^{2^{O(L)}}} \enspace .$$

H.2.1 Proof of Lemma H.5

Proof of Lemma H.5. Suppose by way of contradiction that for some sufficiently large constant $C > 1$,

$$\mathbb{E}_{x,y}(h(x,y) - f(x,y))^2 \le \frac{1}{p^{C\cdot p}}$$

This implies that

$$\mathbb{E}_x\left(\mathbb{E}_y h(x,y) - \mathbb{E}_y f(x,y)\right)^2 \le \frac{1}{p^{C\cdot p}} \tag{H.1}$$

We break x into p parts: $x = (x^{(1)}, x^{(2)}, \cdots, x^{(p)})$ where each $x^{(j)} \in \mathbb{R}^{d_1/p}$. We also decompose w_i into $(w_i^{(1)}, w_i^{(2)}, \cdots, w_i^{(p)}, w'_i)$ accordingly. We can write

$$\left(\frac{\|x\|_4^4}{d_1}\right)^p = \left(\frac{\sum_{j\in[p]}\|x^{(j)}\|_4^4}{d_1}\right)^p \tag{H.2}$$

Since $\widetilde{\sigma}_i$ is of degree at most $2p$, we can write for some coefficients $a_{i,q}$:

$$\mathop{\mathbb{E}}_{y} a_i \widetilde{\sigma}_i(\langle w_i, (x, x^2, y) + b_i \rangle) = \sum_{q \in [2p]} a_{i,q} \left(\sum_{j \in [p]} \langle x^{(j)}, w_i^{(j)} \rangle + \langle \left(x^{(j)}\right)^2, w_i^{(j)} \rangle \right)^q \tag{H.3}$$

Let us now go back to (H.1). We know that $\mathbb{E}_y f(x,y)$ and $\mathbb{E}_y h(x,y)$ are both polynomials over $x \in \mathbb{R}^{d_1}$ with maximum degree $4p$.

- The only $4p$-degree monomials of $\mathbb{E}_y f(x,y)$ come from (H.2) which is $\frac{1}{(d_1)^p}\left(\sum_{j \in [p]} \|x^{(j)}\|_4^4\right)^p$. Among them, the only ones with homogeneous degree 4 for each $x^{(j)}$ is $\frac{1}{(d_1)^p} \prod_{j \in [p]} \|x^{(j)}\|_4^4$.
- The only $4p$-degree monomials of $\mathbb{E}_y h(x,y)$ come from (H.3) which is $a_{i,2p} \left(\sum_{j \in [p]} \langle \left(x^{(j)}\right)^2, w_i^{(j)} \rangle \right)^{2p}$. Among them, the only ones with homogeneous degree 4 for each $x^{(j)}$ can be written as $\frac{a_i'}{(d_1)^p} \prod_{j \in [p]} \left(\langle \left(x^{(j)}\right)^2, w_i^{(j)} \rangle \right)^2$.

Applying the degree-preserving Property 5.2 for Gaussian polynomials:

$$\mathcal{C}_x \left(\sum_i a_i' \prod_{j \in [p]} \langle \left(x^{(j)}\right)^2, w_i^{(j)} \rangle^2 - \prod_{j \in [p]} \|x^{(j)}\|_4^4 \right) \leq \frac{(d_1)^{2p}}{p^{(C-10)p}} \; .$$

Let us denote $\prod_{j \in [p]} \langle \left(x^{(j)}\right)^2, w_i^{(j)} \rangle = \langle \widetilde{w}_i, \widetilde{x} \rangle$ where $\widetilde{x}, \widetilde{w}_i \in \mathbb{R}^{(d_1/p)^p}$ are given as:

$$\widetilde{x} = \left(\prod_{j \in [p]} \left(x_{i_j}^{(j)} \right)^2 \right)_{i_1, \cdots, i_p \in [d_1/p]} \qquad \text{and} \quad \widetilde{w}_i = \left(\prod_{j \in [p]} [w_i^{(j)}]_{i_j} \right)_{i_1, \cdots, i_p \in [d_1/p]}$$

Under this notation, we have

$$\prod_{j \in [p]} \|x^{(j)}\|_4^4 = \|\widetilde{x}\|_2^2, \quad \sum_i a_i' \prod_{j \in [p]} \langle \left(x^{(j)}\right)^2, w_i^{(j)} \rangle^2 = \widetilde{x}^\top \sum_i a_i' \widetilde{w}_i (\widetilde{w}_i)^\top \widetilde{x}^\top$$

This implies that for $\mathbf{M} = \sum_i a_i' \widetilde{w}_i (\widetilde{w}_i)^\top \in \mathbb{R}^{(d_1/p)^p \times (d_1/p)^p}$, we have

$$\mathcal{C}_x \left(\widetilde{x}^\top (\mathbf{M} - \mathbf{I}) \widetilde{x}^\top \right) = \frac{(d_1)^{2p}}{p^{(C-10)p}}$$

By the special structure of $\mathbf{M}$ where $\mathbf{M}_{(i_1,i_1'),(i_2,i_2'),\cdots,(i_j,i_j')} = \mathbf{M}_{\{i_1,i_1'\},\{i_2,i_2'\},\cdots,\{i_j,i_j'\}}$ does not depend on the order of (i_j, i_j') (since each $\widetilde{w}_i (\widetilde{w}_i)^\top$ has this property), we further know that

$$\|\mathbf{I} - \mathbf{M}\|_F^2 = \frac{(d_1)^{2p}}{p^{(C-10)p}} \ll (d_1/p)^p \times (d_1/p)^p$$

This implies that the rank r of $\mathbf{M}$ must satisfy $r = \Omega((d_1/p)^p)$ using [2, Lemma E.2]. □

I Mathematical Preliminaries

I.1 Concentration of Gaussian Polynomials

Lemma I.1. *Suppose $f \colon \mathbb{R}^m \to \mathbb{R}$ is a degree q homogenous polynomial, and let $\mathcal{C}(f)$ be the sum of squares of all the monomial coefficients of f. Suppose $g \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian, then for every $\varepsilon \in (0, \frac{1}{10})$,*

$$\mathop{\mathbf{Pr}}_{g \sim \mathcal{N}(0,\mathbf{I})} \left[|f(g)| \leq \varepsilon \sqrt{\mathcal{C}(f)} \right] \leq O(q) \cdot \varepsilon^{1/q}$$

Proof. Recall from the anti-concentration of Gaussian polynomial (see Lemma I.2a)

$$\Pr_{g\sim\mathcal{N}(0,\mathbf{I})}\left[|f(g)-t|\le\varepsilon\sqrt{\mathbf{Var}[f(g)]}\right]\le O(q)\cdot\varepsilon^{1/q}$$

Next, one can verify when f is degree-q homogenous for $q\ge 1$, we have $\mathbf{Var}[f(g)]\ge\mathcal{C}(f)$. This can be seen as follows, first, we write $\mathbf{Var}[f(g)]=\mathbb{E}[(f(g)-\mathbb{E}\,f(g))^2]$. Next, we rewrite the polynomial $f(g)-\mathbb{E}\,f(g)$ in the Hermite basis of g. For instance, $g_1^5g_2^2$ is replaced with $(H_5(g_1)+\cdots)(H_2(g_2)+\cdots)$ where $H_k(x)$ is the (probabilists') k-th order Hermite polynomial and the "$\cdots$" hides lower-order terms. This transformation does not affect the coefficients of the highest degree monomials. (For instance, the coefficient in front of $H_5(g_1)H_2(g_2)$ is the same as the coefficient in front of $g_1^5g_2^2$. By the orthogonality of Hermite polynomials with respect to the Gaussian distribution, we immediately have $\mathbb{E}[(f(g)-\mathbb{E}\,f(g))^2]\ge\mathcal{C}(f)$. $\square$

Lemma I.2. *Let $f\colon\mathbb{R}^m\to\mathbb{R}$ be a degree q polynomial.*

(a) Anti-concentration (see e.g. [44, Eq. (1)]): for every $t\in\mathbb{R}$ and $\varepsilon\in(0,1)$,

$$\Pr_{g\sim\mathcal{N}(0,\mathbf{I})}\left[|f(g)-t|\le\varepsilon\sqrt{\mathbf{Var}[f(g)]}\right]\le O(q)\cdot\varepsilon^{1/q}$$

(b) Hypercontractivity concentration (see e.g. [48, Thm 1.9]): there exists constant $R>0$ so that

$$\Pr_{g\sim\mathcal{N}(0,\mathbf{I})}\left[|f(g)-\mathbb{E}[f(g)]|\ge\lambda\right]\le e^2\cdot e^{-\left(\frac{\lambda^2}{R\cdot\mathbf{Var}[f(g)]}\right)^{1/q}}$$

I.2 Random Initialization

Lemma B.6. *Let $\mathbf{R}_\ell\in\mathbb{R}^{\binom{k_\ell+1}{2}\times k_\ell}$ be a random matrix such that each entry is i.i.d. from $\mathcal{N}\left(0,\frac{1}{k_\ell^2}\right)$, then with probability at least $1-p$, $\mathbf{R}_\ell*\mathbf{R}_\ell$ has singular values between $[\frac{1}{O(k_\ell^4p^2)},O(1+\frac{1}{k_\ell^2}\log\frac{k_\ell}{p})]$, and $\|\mathbf{R}_\ell\|_2\le O(1+\frac{\sqrt{\log(1/p)}}{k_\ell})$.*

As a result, with probability at least 0.99, it satisfies for all $\ell=2,3,\dots,L$, the square matrices $\mathbf{R}_\ell*\mathbf{R}_\ell$ have singular values between $[\frac{1}{O(k_\ell^4L^2)},O(1+\frac{\log(Lk_\ell)}{k_\ell})]$ and $\|\mathbf{R}_\ell\|_2\le O(1+\frac{\sqrt{\log L}}{k_\ell})$.

Proof. Let us drop the subscript ℓ for simplicity, and denote by $m=\binom{k+1}{2}$. Consider any unit vector $u\in\mathbb{R}^m$. Define $v^{(i)}$ to (any) unit vector orthogonal to all the rows of $\mathbf{R}$ except its i-th row. We have

$$|u^\top(\mathbf{R}*\mathbf{R})v^{(i)}|=|u_i(\mathbf{R}_{i,:}*\mathbf{R}_{i,:})v^{(i)}|=|u_i|\left|\sum_{p\le q}a_{p,q}\mathbf{R}_{i,p}\mathbf{R}_{i,q}v^{(i)}_{p,q}\right|$$

Now, we have that $v^{(i)}$ is independent of the randomness of $\mathbf{R}_{i,:}$, and therefore, by anti-concentration of Gaussian homogenous polynomials (see Lemma I.1),

$$\Pr_{\mathbf{R}_{i,:}}\left[\left|\sum_{p\le q}a_{p,q}\mathbf{R}_{i,p}\mathbf{R}_{i,q}v^{(i)}_{p,q}\right|\le\varepsilon\|v^{(i)}\|\cdot\frac{1}{k}\right]\le O(\varepsilon^{1/2})\ .$$

Therefore, given any fixed i, with probability at least $1-O(\varepsilon^{1/2})$, it satisfies that for *every* unit vector u,

$$|u^\top(\mathbf{R}*\mathbf{R})v^{(i)}|\ge\frac{\varepsilon}{k}|u_i|\ .$$

By union bound, with probability at least $1-O(k\varepsilon^{1/2})$, the above holds for all i and all unit vectors u. Since $\max_i |u_i| \geq \frac{1}{k}$ for any unit vector $u \in \mathbb{R}^{\binom{k+1}{2}}$, we conclude that $\sigma_{\min}(\mathbf{R} * \mathbf{R}) \geq \frac{\varepsilon}{k^2}$ with probability at least $1-O(k\varepsilon^{1/2})$.

As for the upper bound, we can do a crude calculation by using $\|\mathbf{R} * \mathbf{R}|_2 \leq \|\mathbf{R} * \mathbf{R}\|_F$.

$$\|\mathbf{R} * \mathbf{R}\|_F^2 = \sum_{i,p\leq q} a_{p,q}^2 \mathbf{R}_{i,p}^2 \mathbf{R}_{i,q}^2 = \sum_i \left(\sum_{p\in[k]} \mathbf{R}_{i,p}^2 \right)^2 .$$

By concentration of chi-square distribution (and union bound), we know that with probability at least $1-p$, the above summation is at most $O(k^2) \cdot (\frac{1}{k} + \frac{\log(k/p)}{k^2})^2$.

Finally, the bound on $\|\mathbf{R}\|_2$ can be derived from any asymptotic bound for the maximum singular value of Gaussian random matrix: $\mathbf{Pr}[\|k\mathbf{R}\|_2 > tk] \leq e^{-\Omega(t^2k^2)}$ for every $t \geq \Omega(1)$. $\square$

I.3 Property on Symmetric Tensor

Lemma B.5. *If $\mathbf{U} \in \mathbb{R}^{p\times p}$ is unitary and $\mathbf{R} \in \mathbb{R}^{s\times p}$ for $s \geq \binom{p+1}{2}$, then there exists some* unitary *matrix $\mathbf{Q} \in \mathbb{R}^{\binom{p+1}{2}\times\binom{p+1}{2}}$ so that $\mathbf{RU} * \mathbf{RU} = (\mathbf{R} * \mathbf{R})\,\mathbf{Q}$.*

Proof of Lemma B.5. For an arbitrary vector $w \in \mathbb{R}^s$, let us denote by $w^\top(\mathbf{R} * \mathbf{R}) = (b_{i,j})_{1\leq i\leq j\leq p}$. Let $g \in \mathcal{N}(0, \mathbf{I}_{p\times p})$ be a Gaussian random vector so we have:

$$w^\top \sigma(\mathbf{R}g) = \sum_{i\in[s]} w_i (\mathbf{R}_i g)^2 = \sum_{i\in[s]} w_i \langle \mathbf{R}_i * \mathbf{R}_i, g * g\rangle = \sum_{i\in[p]} b_{i,i} g_i^2 + \sqrt{2} \sum_{1\leq i<j\leq p} b_{i,j} g_i g_j \enspace .$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left(w^\top \sigma(\mathbf{R}g)\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i\in[p]} b_{i,i} g_i^2 + \sqrt{2}\sum_{1\leq i<j\leq p} b_{i,j} g_i g_j\right)^2\right] \\
&= 2\sum_{1\leq i<j\leq p} b_{i,j}^2 + 2\sum_{1\leq i<j\leq p} b_{i,i} b_{j,j} + 3\sum_{i\in[p]} b_{i,i}^2 \\
&= 2\sum_{1\leq i<j\leq p} b_{i,j}^2 + \left(\sum_{i\in[p]} b_{i,i}\right)^2 + 2\sum_{i\in[p]} b_{i,i}^2 \enspace .
\end{aligned}$$

On the other hand, we have $\mathbb{E}\left[w^\top \sigma(\mathbf{R}g)\right] = \sum_{i\in[p]} b_{i,i}$. Therefore, we have

$$\mathbf{Var}\left[w^\top \sigma(\mathbf{R}g)\right] = 2\|w^\top(\mathbf{R} * \mathbf{R})\|_2^2 \enspace .$$

Note that $\mathbf{Var}[w^\top \sigma(\mathbf{R}g)] = \mathbf{Var}[w^\top \sigma(\mathbf{RU}g)]$ for a unitary matrix $\mathbf{U}$, therefore we conclude that

$$\|w^\top(\mathbf{RU} * \mathbf{RU})\|_2^2 = \|w^\top(\mathbf{R} * \mathbf{R})\|_2^2$$

for any vector w. Which implies that there exists some *unitary* matrix $\mathbf{Q} \in \mathbb{R}^{\binom{p+1}{2}\times\binom{p+1}{2}}$ so that $\mathbf{RU} * \mathbf{RU} = (\mathbf{R} * \mathbf{R})\,\mathbf{Q}$. $\square$

I.4 Properties On Homogeneous Polynomials

Given any degree-q homogenous polynomial $f(x) = \sum_{I\in\mathbb{N}^n \colon \|I\|_1 = q} a_I \prod_{j\in[n]} x_j^{I_j}$, recall we have defined

$$\mathcal{C}_x(f) \stackrel{\text{def}}{=} \sum_{I\in\mathbb{N}^n \colon \|I\|_1 = q} a_I^2$$

When it is clear from the context, we also denote $\mathcal{C}(f) = \mathcal{C}_x(f)$.

Definition I.3. *Given $f\colon \mathbb{R}^n \to \mathbb{R}$ and vector $y \in \mathbb{R}^n$, define the directional derivative*

$$(\Delta_y f)(x) \stackrel{\text{def}}{=} f(x+y) - f(x)$$

and given vectors $y^{(1)}, \dots, y^{(q)} \in \mathbb{R}^n$, define $\Delta_{y^{(1)},\dots,y^{(q)}} f = \Delta_{y^{(1)}} \Delta_{y^{(2)}} \dots \Delta_{y^{(q)}}$.

Lemma I.4. *Suppose $f\colon \mathbb{R}^n \to \mathbb{R}$ is a degree-q homogeneous polynomial. Then, the finite-differentiate polynomial*

$$\widehat{f}(y^{(1)}, \dots, y^{(q)}) = \Delta_{y^{(1)},\dots,y^{(q)}} f(x)$$

is also degree-q homogenous over $n \times q$ variables, and satisfies

- $\mathcal{C}(f) \cdot q! \le \mathcal{C}(\widehat{f}) \le \mathcal{C}(f) \cdot (q!)^2$.
- $\mathbb{E}_{y^{(1)},\dots,y^{(q)} \sim \mathcal{N}(0,\mathbf{I}_{n\times n})}[\big(\widehat{f}(y^{(1)}, \dots, y^{(q)})\big)^2] = \mathcal{C}(\widehat{f})$

Proof. Suppose $f(x) = \sum_{I \in \mathbb{N}^n \colon \|I\|_1 = q} a_I \prod_{j\in[n]} x_j^{I_j}$. Then, we have (see [44, Claim 3.2])

$$\widehat{f}(y^{(1)}, \dots, y^{(q)}) = \sum_{J \in [n]^q} \widehat{a}_J \prod_{j \in [q]} y_{J_j}^{(j)}$$

where $\widehat{a}_J = a_{I(J)} \cdot \prod_{k=1}^n (I_k(J))!$ and $I_k(J) = |\{j \in [q] \colon J_j = k\}|$.

On the other hand, for every $I^* \in \mathbb{N}^q$ with $\|I^*\|_1 = q$, there are $\frac{q!}{\prod_{k=1}^n (I_k^*)!}$ different choices of $J \in [n]^q$ that maps $I(J) = I^*$. Therefore, we have

$$\mathcal{C}(\widehat{f}) = \sum_{J \in [n]^q} \widehat{a}_J^2 = \sum_{J \in [n]^q} a_{I(J)}^2 \cdot \Big(\prod_{k=1}^n (I_k(J))!\Big)^2 = \sum_{I \in \mathbb{N}^n \colon \|I\|_1 = q} a_I^2 \cdot \Big(\prod_{k=1}^n (I_k)!\Big)^2 \cdot \frac{q!}{\prod_{k=1}^n (I_k)!}$$

As a result,

$$\sum_{I \in \mathbb{N}^n \colon \|I\|_1 = q} a_I^2 \cdot (q!) \le \mathcal{C}(\widehat{f}) \le \sum_{I \in \mathbb{N}^n \colon \|I\|_1 = q} a_I^2 \cdot (q!)^2$$

As for the second bullet, it is simple to verify. □

Lemma I.5. *Suppose $f\colon \mathbb{R}^n \to \mathbb{R}$ is a degree-q homogeneous polynomial.*

- *If $g(x) = f(\mathbf{U}x)$ for $\mathbf{U} \in \mathbb{R}^{n\times m}$ being row orthonormal (with $n \le m$), then $\mathcal{C}(g) \ge \frac{\mathcal{C}(f)}{q!}$.*
- *If $g(x) = f(\mathbf{W}x)$ for $\mathbf{W} \in \mathbb{R}^{n\times m}$ with $n \le m$ and $\sigma_{\min}(\mathbf{W}) \ge \frac{1}{\kappa}$, then $\mathcal{C}(g) \ge \frac{\mathcal{C}(f)}{(q!)^2 \kappa^q}$.*

Proof.

- For every $y^{(1)}, \dots, y^{(q)} \in \mathbb{R}^m$,

  $$\widehat{g}(y^{(1)}, \dots, y^{(q)}) = \Delta_{y^{(1)},\dots,y^{(q)}} g(x) = \Delta_{\mathbf{U}y^{(1)},\dots,\mathbf{U}y^{(q)}} f(\mathbf{U}x) = \widehat{f}(\mathbf{U}y^{(1)}, \dots, \mathbf{U}y^{(q)})$$

  Since Gaussian is invariant under orthonormal transformation, we have

  $$\mathcal{C}(\widehat{f}) = \mathop{\mathbb{E}}_{y^{(1)},\dots,y^{(q)} \sim \mathcal{N}(0,\mathbf{I}_{n\times n})}[\big(\widehat{f}(y^{(1)}, \dots, y^{(q)})\big)^2] = \mathop{\mathbb{E}}_{y^{(1)},\dots,y^{(q)} \sim \mathcal{N}(0,\mathbf{I}_{m\times m})}[\big(\widehat{f}(\mathbf{U}y^{(1)}, \dots, \mathbf{U}y^{(q)})\big)^2] = \mathcal{C}(\widehat{g})$$

- Suppose $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$ is its SVD decomposition. Define $f_1(x) = f(\mathbf{U}x)$, $f_2(x) = f_1(\mathbf{\Sigma}x)$, so that $g(x) = f_2(\mathbf{V}x)$. We have $\mathcal{C}(g) \ge \frac{1}{q!}\mathcal{C}(f_2) \ge \frac{1}{q!\kappa^q}\mathcal{C}(f_1) \ge \frac{1}{(q!)^2\kappa^q}\mathcal{C}(f)$.

□

Lemma I.6. *Suppose $f, g\colon \mathbb{R}^n \to \mathbb{R}$ are two homogeneous polynomials of degree p and q respectively, and denote by $h(x) = f(x)g(x)$. Then $\mathcal{C}_x(h) \le \binom{p+q}{p} \mathcal{C}_x(f) \mathcal{C}_x(g)$.*

Proof. Let us write

$$f(x) = \sum_{I \in \mathbb{N}^k \colon \|I\|_1 = p} a_I \prod_{j \in [k]} x_j^{I_j} \quad \text{and} \quad g(x) = \sum_{J \in \mathbb{N}^k \colon \|J\|_1 = q} b_J \prod_{j \in [k]} x_j^{J_j} \ .$$

On one hand, we obviously have $\sum_{I \in \mathbb{N}^k \colon \|I\|_1 = p} \sum_{J \in \mathbb{N}^k \colon \|J\|_1 = q} a_I^2 b_J^2 = \mathcal{C}(f)\mathcal{C}(g)$. On the other hand, when multiplied together, each monomial in the multiplication $f(x)g(x)$ comes from at most $\binom{p+q}{p}$ pairs of (I, J). If we denote this set as S, then

$$\Big(\sum_{(I,J) \in S} a_I b_J\Big)^2 \le \binom{p+q}{p} \sum_{(I,J) \in S} a_I^2 b_J^2 \ .$$

Putting the two together finishes the proof. $\square$

Lemma I.7. *Suppose $f^{(1)}, f^{(2)} \colon \mathbb{R}^n \to \mathbb{R}^k$ are degree-p homogeneous polynomials and $g \colon \mathbb{R}^k \to \mathbb{R}$ is degree q homogenous. Denote by $h(x) = g(f^{(1)}(x)) - g(f^{(2)}(x))$. Then,*

$$\mathcal{C}_x(h) \le k^q q^2 \cdot 2^{q-1} \cdot \binom{qp}{p, p, \dots, p} \cdot \mathcal{C}(g) \cdot (\max_i \mathcal{C}(f_i^{(1)} - f_i^{(2)})) \cdot (\max_i \mathcal{C}(f_i^{(1)}) + \max_i \mathcal{C}(f_i^{(1)} - f_i^{(2)}))^{q-1} \ .$$

Proof. Let us write

$$g(y) = \sum_{I \in \mathbb{N}^k \colon \|I\|_1 = q} a_I \prod_{j \in [k]} y_j^{I_j} \ .$$

For each monomial above, we need to bound $\mathcal{C}_x(h_I(x))$ for each

$$h_I(x) \stackrel{\text{def}}{=} \prod_{j \in [k]} (f_j^{(1)}(x))^{I_j} - \prod_{j \in [k]} (f_j^{(2)}(x))^{I_j} = \prod_{j \in S} f_j^{(1)}(x) - \prod_{j \in S} f_j^{(2)}(x)$$

where $S \subset [k]$ is a multiset that contains exactly I_j copies of j. Using the identity that $a_1a_2a_3a_4 - b_1b_2b_3b_4 = (a_1 - b_1)a_2a_3a_4 + b_1(a_2 - b_2)a_3a_4 + b_1b_2(a_3 - b_3)a_4 + b_1b_2b_3(a_4 - b_4)$, as well as applying Lemma I.6, one can derive that

$$\begin{aligned}
\mathcal{C}_x(h_I) &\le q^2 \cdot \binom{qp}{p, p, \dots, p} \cdot (\max_i \mathcal{C}(f_i^{(1)} - f_i^{(2)})) \cdot (\max_i \{\mathcal{C}(f_i^{(1)}), \mathcal{C}(f_i^{(2)})\})^{q-1} \\
&\le q^2 \cdot 2^{q-1} \cdot \binom{qp}{p, p, \dots, p} \cdot (\max_i \mathcal{C}(f_i^{(1)} - f_i^{(2)})) \cdot (\max_i \mathcal{C}(f_i^{(1)}) + \max_i \mathcal{C}(f_i^{(1)} - f_i^{(2)}))^{q-1}
\end{aligned}$$

Summing up over all monomials finishes the proof. $\square$

I.5 Properties on Matrix Factorization

Claim I.8. *Suppose we have matrices $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{k \times m_1}$ and $\mathbf{B}, \mathbf{D} \in \mathbb{R}^{k \times m_2}$ for some $m_1, m_2 \ge k$ and $\|\mathbf{A}^\top \mathbf{B} - \mathbf{C}^\top \mathbf{D}\|_F \le \varepsilon$. Then, there exists some matrix $\mathbf{P} \in \mathbb{R}^{k \times k}$ so that:*

- $\|\mathbf{A}^\top - \mathbf{C}^\top \mathbf{P}\|_F \le \frac{\varepsilon}{\sigma_{\min}(\mathbf{B})}$,
- $\|\mathbf{B} - \mathbf{P}^{-1} \mathbf{C}\|_F \le \frac{2\varepsilon \cdot (\sigma_{\max}(\mathbf{B}))^2}{\sigma_{\min}(\mathbf{B}) \sigma_{\min}(\mathbf{C}) \sigma_{\min}(\mathbf{D})}$, *and*
- *the singular values of $\mathbf{P}$ are within $\big[\frac{\sigma_{\min}(\mathbf{D})}{\sigma_{\max}(\mathbf{B})}, \frac{\sigma_{\max}(\mathbf{D})}{\sigma_{\min}(\mathbf{B})}\big]$.*

Proof of Claim I.8. We also refer to [1] for the proof.

Suppose $\mathbf{A} = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1$, $\mathbf{B} = \mathbf{U}_2 \mathbf{\Sigma}_2 \mathbf{V}_2$, $\mathbf{C} = \mathbf{U}_3 \mathbf{\Sigma}_3 \mathbf{V}_3$, $\mathbf{D} = \mathbf{U}_4 \mathbf{\Sigma}_4 \mathbf{V}_4$ are the SVD decompositions. We can write

$$\begin{aligned}
&\|\mathbf{V}_1^\top \mathbf{\Sigma}_1^\top \mathbf{U}_1^\top \mathbf{U}_2 \mathbf{\Sigma}_2 \mathbf{V}_2 - \mathbf{V}_3^\top \mathbf{\Sigma}_3^\top \mathbf{U}_3^\top \mathbf{U}_4 \mathbf{\Sigma}_4 \mathbf{V}_4\|_F \le \varepsilon \\
\implies &\|\mathbf{V}_3 \mathbf{V}_1^\top \mathbf{\Sigma}_1^\top \mathbf{U}_1^\top \mathbf{U}_2 \mathbf{\Sigma}_2 \mathbf{V}_2 \mathbf{V}_4^\top - \mathbf{\Sigma}_3^\top \mathbf{U}_3^\top \mathbf{U}_4 \mathbf{\Sigma}_4\|_F \le \varepsilon
\end{aligned}$$

Now note that $\mathbf{\Sigma}_3^\top \mathbf{U}_3^\top \mathbf{U}_4 \mathbf{\Sigma}_4$ is of dimension $m_1 \times m_2$ and only its top left $k \times k$ block is non-zero. Let us write $\mathbf{\Sigma}_4 = (\overline{\mathbf{\Sigma}}_4, \mathbf{0})$ for $\overline{\mathbf{\Sigma}}_4 \in \mathbb{R}^{k\times k}$. Let us write $\mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2\mathbf{V}_4^\top = (\mathbf{E}, \mathbf{F})$ for $\mathbf{E} \in \mathbb{R}^{k\times k}$. Then, the above Frobenius bound also implies (by ignoring the last $m_2 - k$ columns)

$$\|\mathbf{V}_3\mathbf{V}_1^\top \mathbf{\Sigma}_1^\top \mathbf{U}_1^\top \mathbf{E} - \mathbf{\Sigma}_3^\top \mathbf{U}_3^\top \mathbf{U}_4 \overline{\mathbf{\Sigma}}_4\|_F \le \varepsilon$$

Finally, using $\|\mathbf{M}\mathbf{N}\|_F \le \|\mathbf{M}\|_F \cdot \sigma_{\max}(\mathbf{N})$, we have

$$\|\mathbf{V}_1^\top \mathbf{\Sigma}_1^\top \mathbf{U}_1^\top - \mathbf{V}_3^\top \mathbf{\Sigma}_3^\top \mathbf{U}_3^\top \mathbf{U}_4 \overline{\mathbf{\Sigma}}_4 \mathbf{E}^{-1}\|_F \le \frac{\varepsilon}{\sigma_{\min}(\mathbf{E})} = \frac{\varepsilon}{\sigma_{\min}(\mathbf{B})}$$

Let us define $\mathbf{P} = \mathbf{U}_4\overline{\mathbf{\Sigma}}_4\mathbf{E}^{-1}$, so we have $\sigma_{\max}(\mathbf{P}) \le \frac{\sigma_{\max}(\mathbf{D})}{\sigma_{\min}(\mathbf{B})}$ and $\sigma_{\min}(\mathbf{P}) \ge \frac{\sigma_{\min}(\mathbf{D})}{\sigma_{\max}(\mathbf{B})}$.

From the above derivation we have

$$\|\mathbf{A}^\top \mathbf{B} - \mathbf{C}^\top \mathbf{P}\mathbf{B}\|_F \le \|\mathbf{A}^\top - \mathbf{C}^\top \mathbf{P}\|_F \cdot \sigma_{\max}(\mathbf{B}) \le \frac{\varepsilon \sigma_{\max}(\mathbf{B})}{\sigma_{\min}(\mathbf{B})}$$

By triangle inequality, this further implies

$$\|\mathbf{C}^\top \mathbf{P}\mathbf{B} - \mathbf{C}^\top \mathbf{P}\mathbf{P}^{-1}\mathbf{D}\|_F \le \varepsilon + \frac{\varepsilon\sigma_{\max}(\mathbf{B})}{\sigma_{\min}(\mathbf{B})} \implies \|\mathbf{B} - \mathbf{P}^{-1}\mathbf{D}\|_F \le \left(\varepsilon + \frac{\varepsilon\sigma_{\max}(\mathbf{B})}{\sigma_{\min}(\mathbf{B})}\right) \cdot \frac{1}{\sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{P})}$$

□

Claim I.9. *Suppose we have matrices $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{k\times m_1}$ and $\mathbf{B}, \mathbf{D} \in \mathbb{R}^{k\times m_2}$ for some $m_1, m_2 \ge k$ and $\|\mathbf{A}^\top \mathbf{B} - \mathbf{C}^\top \mathbf{D}\|_F \le \varepsilon < \sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{D})$. Then, there exists some matrix $\mathbf{P} \in \mathbb{R}^{k\times k}$ so that:*

- $\|\mathbf{A}^\top - \mathbf{C}^\top \mathbf{P}\|_F \le \frac{\varepsilon\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{D})-\varepsilon}$,
- $\|\mathbf{B} - \mathbf{P}^{-1}\mathbf{C}\|_F \le \frac{2\varepsilon\cdot(\sigma_{\max}(\mathbf{B}))^2\sigma_{\max}(\mathbf{A})}{(\sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{D})-\varepsilon)^2}$, *and*
- *the singular values of $\mathbf{P}$ are within* $\left[\frac{\sigma_{\min}(\mathbf{D})}{\sigma_{\max}(\mathbf{B})}, \frac{\sigma_{\max}(\mathbf{D})\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{D})-\varepsilon}\right]$.

Proof of Claim I.9. Without loss of generality (by left/right multiplying a unitary matrix), let us assume that $\mathbf{C} = (\overline{\mathbf{C}}, \mathbf{0})$ and $\mathbf{D} = (\overline{\mathbf{D}}, \mathbf{0})$ for $\overline{\mathbf{C}}, \overline{\mathbf{D}} \in \mathbb{R}^{k\times k}$. Let us write $\mathbf{A} = (\overline{\mathbf{A}}, *)$ and $\mathbf{B} = (\overline{\mathbf{B}}, *)$ for $\overline{\mathbf{A}}, \overline{\mathbf{B}} \in \mathbb{R}^{k\times k}$. We have the following relationships

$$\sigma_{\min}(\overline{\mathbf{C}}) = \sigma_{\min}(\mathbf{C}) \ , \quad \sigma_{\min}(\overline{\mathbf{D}}) = \sigma_{\min}(\mathbf{D}) \ , \quad \sigma_{\max}(\overline{\mathbf{A}}) \le \sigma_{\max}(\mathbf{A}) \ , \quad \sigma_{\min}(\overline{\mathbf{B}}) \le \sigma_{\min}(\mathbf{B}) \ .$$

Now, the bound $\|\mathbf{A}^\top \mathbf{B} - \mathbf{C}^\top \mathbf{D}\|_F \le \varepsilon$ translates to (by only looking at its top-left $k \times k$ block) $\|\overline{\mathbf{A}}^\top \overline{\mathbf{B}} - \overline{\mathbf{C}}^\top \overline{\mathbf{D}}\|_F \le \varepsilon$. Since these four matrices are square matrices, we immediately have $\sigma_{\min}(\overline{\mathbf{B}}) \ge \frac{\sigma_{\min}(\overline{\mathbf{C}})\sigma_{\min}(\overline{\mathbf{D}})-\varepsilon}{\sigma_{\max}(\overline{\mathbf{A}})}$. Plugging in the above relationships, the similar bound holds without the hat notion:

$$\sigma_{\min}(\mathbf{B}) \ge \frac{\sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{D}) - \varepsilon}{\sigma_{\max}(\mathbf{A})} \ .$$

Plugging this into the bounds of Claim I.8, we finish the proof. □

Claim I.10. *Suppose we have matrices $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{k\times m}$ for some $m \ge k$ and $\|\mathbf{A}^\top \mathbf{A} - \mathbf{C}^\top \mathbf{C}\|_F \le \varepsilon \le \frac{1}{2}(\sigma_{\min}(\mathbf{C}))^2$, then there exists some unitary matrix $\mathbf{U} \in \mathbb{R}^{k\times k}$ so that*

$$\|\mathbf{A}^\top - \mathbf{C}^\top \mathbf{U}\|_F \le \frac{7\varepsilon(\sigma_{\max}(\mathbf{A}) + \sigma_{\max}(\mathbf{C}))^2(\sigma_{\max}(\mathbf{C}))^3}{(\sigma_{\min}(\mathbf{C}))^6} \ .$$

Proof of Claim I.10. Applying Claim I.9, we know there exists matrix $\mathbf{P} \in \mathbb{R}^{k\times k}$ so that:

- $\|\mathbf{A}^\top - \mathbf{C}^\top \mathbf{P}\|_F \le \frac{2\varepsilon\sigma_{\max}(\mathbf{A})}{(\sigma_{\min}(\mathbf{C}))^2}$,

- the singular values of $\mathbf{P}$ are within $\left[\frac{\sigma_{\min}(\mathbf{C})}{\sigma_{\max}(\mathbf{A})}, \frac{2\sigma_{\max}(\mathbf{C})\sigma_{\max}(\mathbf{A})}{(\sigma_{\min}(\mathbf{C}))^2}\right]$.

They together imply

$$\begin{aligned}\|\mathbf{A}^\top\mathbf{A} - \mathbf{C}^\top\mathbf{P}\mathbf{P}^\top\mathbf{C}\|_F &\le \frac{2\varepsilon\sigma_{\max}(\mathbf{A})}{(\sigma_{\min}(\mathbf{C}))^2}\cdot(\sigma_{\max}(\mathbf{A}) + \sigma_{\max}(\mathbf{C})\sigma_{\max}(\mathbf{P}))\\ &\le \frac{2\varepsilon\sigma_{\max}(\mathbf{A})}{(\sigma_{\min}(\mathbf{C}))^2}\cdot\frac{3(\sigma_{\max}(\mathbf{C}))^2\sigma_{\max}(\mathbf{A})}{(\sigma_{\min}(\mathbf{C}))^2} \le \frac{6\varepsilon(\sigma_{\max}(\mathbf{A}))^2(\sigma_{\max}(\mathbf{C}))^2}{(\sigma_{\min}(\mathbf{C}))^4}\end{aligned}$$

By triangle inequality we have

$$\|\mathbf{C}^\top\mathbf{C} - \mathbf{C}^\top\mathbf{P}\mathbf{P}^\top\mathbf{C}\|_F \le \frac{7\varepsilon(\sigma_{\max}(\mathbf{A}) + \sigma_{\max}(\mathbf{C}))^2(\sigma_{\max}(\mathbf{C}))^2}{(\sigma_{\min}(\mathbf{C}))^4}$$

Putting $\mathbf{C}$ into its SVD decomposition, one can easily verify that this implies

$$\|\mathbf{I} - \mathbf{P}\mathbf{P}^\top\|_F \le \frac{7\varepsilon(\sigma_{\max}(\mathbf{A}) + \sigma_{\max}(\mathbf{C}))^2(\sigma_{\max}(\mathbf{C}))^2}{(\sigma_{\min}(\mathbf{C}))^6}$$

Putting $\mathbf{P}$ into its SVD decomposition, one can easily verify that this implies the existence of some unitary matrix $\mathbf{U}$ so that[37]

$$\|\mathbf{U} - \mathbf{P}\|_F \le \frac{7\varepsilon(\sigma_{\max}(\mathbf{A}) + \sigma_{\max}(\mathbf{C}))^2(\sigma_{\max}(\mathbf{C}))^2}{(\sigma_{\min}(\mathbf{C}))^6} \ .$$

Finally, we replace $\mathbf{P}$ with $\mathbf{U}$ in the bound $\|\mathbf{A}^\top - \mathbf{C}^\top\mathbf{P}\|_F \le \frac{2\varepsilon\sigma_{\max}(\mathbf{A})}{(\sigma_{\min}(\mathbf{C}))^2}$, and finish the proof. □

I.6 Nonconvex Optimization Theory

Fact I.11. *For every B-second-order smooth function $f : \mathbb{R}^d \to \mathbb{R}$, every $\varepsilon, X_2 > 0$, every fixed vectors $x, x_1 \in \mathbb{R}^d$, suppose there is a random vector $x_2 \in \mathbb{R}^d$ with $\mathbb{E}[x_2] = 0$ satisfying for every sufficiently small $\eta > 0$,*

$$\mathbb{E}_{x_2}[f(x + \eta x_1 + \sqrt{\eta}x_2)] \le f(x) - \eta\varepsilon \ .$$

Then, either $\|\nabla f(x)\| \ge \frac{\varepsilon}{2\|x_1\|}$ or $\lambda_{\min}(\nabla^2 f(x)) \le -\frac{\varepsilon}{\mathbb{E}[\|x_2\|^2]}$, where $\lambda_{\min}$ is the minimal eigenvalue.

Proof of Fact I.11. We know that

$$\begin{aligned}&f(x + \eta x_1 + \sqrt{\eta}x_2)\\ &= f(x) + \langle\nabla f(x), \eta x_1 + \sqrt{\eta}x_2\rangle + \frac{1}{2}(\eta x_1 + \sqrt{\eta}x_2)^\top\nabla^2 f(x)(\eta x_1 + \sqrt{\eta}x_2) \pm O(B\eta^{1.5}).\end{aligned}$$

Taking expectation, we know that

$$\mathbb{E}[f(x + \sqrt{\eta}x_2)] = f(x) + \eta\langle\nabla f(x), x_1\rangle + \eta\frac{1}{2}\mathbb{E}\left[x_2^\top\nabla^2 f(x)x_2\right] \pm O(B\eta^{1.5})$$

Thus, either $\langle\nabla f(x), x_1\rangle \le -\varepsilon/2$ or $\mathbb{E}\left[x_2^\top\nabla^2 f(x)x_2\right] \le -\varepsilon$, which completes the proof. □

J Details on Empirical Evaluations

We explain how Figure 1 is achieved. Recall AlexNet has 5 convolutional layers with ReLU activation, connected sequentially. The output of AlexNet is a linear function over its 5th convolutional

[37]Indeed, if the singular values of $\mathbf{P}$ are $p_1, \dots, p_k$, then $\|\mathbf{I} - \mathbf{P}\mathbf{P}^\top\|_F \le \delta$ says $\sum_i(1-p_i^2)^2 \le \delta^2$, but this implies $\sum_i(1-p_i)^2 \le \delta^2$.

layer. To make AlexNet more connected to the language of this paper, we redefine its network output as a linear functions over all the five convolutional layers. We only train the weights of the convolutional layers and keep the weights of the linear layer unchanged.

We use fixed learning rate 0.01, momentum 0.9, batch size 128, and weight decay 0.0005. In the first 80 epochs, we freeze the (randomly initialized) weights of the 2nd through 5th convolutional layers, and only train the weights of the first layer). In the next 120 epochs, we unfreeze those weights and train all the 5 convolutional layers together.

As one can see from Figure 1, in the first 80 epochs, we have sufficiently trained the first layer (alone) so that the features do not move significantly anymore; however, as the 2nd through 5th layers become trained together, the features of the first layer gets significantly improved.

We explain how Figure 3 is achieved. We use the vanilla ResNet architecture which requires the number of layers to be $6n + 2$. For a fair comparison, we stick to "basic blocks" without using "bottleneck blocks" for all the depths. Following the tradition in training recipes,[38] we use learning rate 0.1, momentum 0.1, batch size 128, and weight decay 10^{-4}. We train for 150 epochs, and decrease the learning rate by 10 twice, at epochs 100 and 125. For each depth, we run the experiment 21 times, and plot the medium together with an 95% percentile error bar.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. LazySVD: even faster SVD decomposition yet without agonizing pain. In *NeurIPS*, pages 974–982, 2016.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. What Can ResNet Learn Efficiently, Going Beyond Kernels? In *NeurIPS*, 2019. Full version available at `http://arxiv.org/abs/1905.10337`.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Can SGD Learn Recurrent Neural Networks with Provable Generalization? In *NeurIPS*, 2019. Full version available at `http://arxiv.org/abs/1902.01028`.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. In *NeurIPS*, 2019. Full version available at `http://arxiv.org/abs/1811.04918`.

[5] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *NeurIPS*, 2019. Full version available at `http://arxiv.org/abs/1810.12065`.

[6] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via overparameterization. In *ICML*, 2019. Full version available at `http://arxiv.org/abs/1811.03962`.

[7] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

[8] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *CoRR*, abs/1901.08584, 2019. URL `http://arxiv.org/abs/1901.08584`.

[9] Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. *arXiv preprint arXiv:1811.01885*, 2018.

[10] Digvijay Boob and Guanghui Lan. Theoretical properties of the global optimizer of two layer neural network. *arXiv preprint arXiv:1710.11241*, 2017.

[11] Jacob V Bouvrie. *Hierarchical learning: Theory with applications in speech and vision.* PhD thesis, Massachusetts Institute of Technology, 2009.

[12] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

[13] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*, 2019.

[38] Specifically, we used the code base from `https://github.com/bearpaw/pytorch-classification`.

[14] Amit Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

[15] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2253–2261, 2016.

[16] Simon S Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. *arXiv preprint arXiv:1901.08572*, 2019.

[17] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, November 2018.

[18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[19] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

[20] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29 (2-3):131–163, 1997.

[21] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[22] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

[23] Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. *arXiv preprint arXiv:1810.06793*, 2018.

[24] Dileep George. *How the brain might work: A hierarchical and temporal model for learning and recognition*. Stanford University Palo Alto, California, 2008.

[25] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

[26] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[27] Richard LT Hahnloser. On the piecewise analysis of networks of linear threshold neurons. *Neural Networks*, 11(4):691–697, 1998.

[28] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019.

[29] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[32] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[33] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. *arXiv preprint arXiv:1909.08156*, 2019.

[34] Lei Huang, Xianglong Liu, Bo Lang, Adams Wei Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

[36] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[38] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*, pages 3675–3683, 2016.

[39] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM, 2009.

[40] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

[41] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607. http://arxiv.org/abs/1705.09886, 2017.

[42] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *COLT*, 2018.

[43] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *arXiv preprint arXiv:1907.04595*, 2019.

[44] Shachar Lovett. An elementary proof of anti-concentration of polynomials in gaussian variables. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 17, page 182, 2010.

[45] Eran Malach and Shai Shalev-Shwartz. A provably correct algorithm for deep learning that actually works. *arXiv preprint arXiv:1803.09522*, 2018.

[46] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.

[47] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

[48] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 437–446. Society for Industrial and Applied Mathematics, 2012.

[49] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[50] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

[51] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[52] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

[53] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560*, 2017.

[54] Santosh Vempala and John Wilmes. Polynomial convergence of gradient descent for training one-hidden-layer neural networks. *arXiv preprint arXiv:1805.02677*, 2018.

[55] Heather Woltman, Andrea Feldstain, J Christine MacKay, and Meredith Rocchi. An introduction to hierarchical linear modeling. *Tutorials in quantitative methods for psychology*, 8(1):52–69, 2012.

[56] Bo Xie, Yingyu Liang, and Le Song. Diversity leads to generalization in neural networks. *arXiv preprint Arxiv:1611.03131*, 2016.

[57] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior,

gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[58] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*, 2019.

[59] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[60] Chiyuan Zhang, Samy Bengio, and Yoram Singer. Are all layers created equal? *arXiv preprint arXiv:1902.01996*, 2019.

[61] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

[62] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.