Learning Over-Parametrized Two-Layer ReLU Neural Networks beyond NTK

Yuanzhi Li   YUANZHIL@ANDREW.CMU.EDU
Carnegie Mellon University

Tengyu Ma   TENGYUMA@STANFORD.EDU
Stanford University

Hongyang R. Zhang   HONGYANG90@GMAIL.COM
University of Pennsylvania

Abstract

We consider the dynamic of gradient descent for learning a two-layer neural network. We assume the input $x \in \mathbb{R}^d$ is drawn from a Gaussian distribution and the label of x satisfies $f^\star(x) = a^\top |W^\star x|$, where $a \in \mathbb{R}^d$ is a nonnegative vector and $W^\star \in \mathbb{R}^{d \times d}$ is an orthonormal matrix. We show that an *over-parametrized* two-layer neural network with ReLU activation, trained by gradient descent from *random initialization*, can provably learn the ground truth network with population loss at most $o(1/d)$ in polynomial time with polynomial samples. On the other hand, we prove that any kernel method, including Neural Tangent Kernel, with a polynomial number of samples in d, has population loss at least $\Omega(1/d)$.

1. Introduction

Gradient-based optimization methods are the method of choice for learning neural networks. However, it has been challenging to understand their working on non-convex functions. Prior works prove that stochastic gradient descent provably convergences to an approximate local optimum (Ge et al. (2015); Sun et al. (2015); Lee et al. (2017); Kleinberg et al. (2018)). Remarkably, for many highly complex neural net models, gradient-based methods can also find high-quality solutions (Sun (2019)) and interpretable features (Zeiler and Fergus (2014)).

Recent studies made the connection between training wide neural networks and Neural Tangent Kernels (NTK) (Jacot et al. (2018); Arora et al. (2019b); Cao and Gu (2019); Du et al. (2018c)). The main idea is that training neural networks with gradient descent with a particular initialization is equivalent to using kernel methods. However, the NTK approach has not yet provided a fully satisfactory theory for explaining the success of neural networks. Empirically, there seems to be a non-negligible gap between the test performance of neural nets trained by SGD and that of the NTK (Arora et al. (2019a); Li et al. (2019b)). Recent works have suspected that the gap stems from that the NTK approach has difficulty dealing with non-trivial explicit regularizers or does not sufficiently leverage the implicit regularization of the algorithm (Wei et al. (2019); Chizat and Bach (2018b); Li et al. (2019a); HaoChen et al. (2020)).

In this work, we provide a new convergence analysis of the gradient descent dynamic on an over-parametrized two-layer ReLU neural network. We prove that for learning a certain two-layer

target network with orthogonal ground truth weights, gradient descent is provably more accurate than any kernel method that uses polynomially large feature maps.

1.1. Setup and Main Result

We assume that the input $x \in \mathbb{R}^d$ is drawn from the Gaussian distribution $\mathcal{N}(0, \text{Id}_{d \times d})$. We focus on the realizable setting, i.e. the label of x is generated according to a target network $f^\star$ with d neurons. We study a two-layer target neural network with absolute value activation:

$$f^\star(x) = \sum_{i=1}^{d} a_i \left| w_i^{\star \top} x \right|, \tag{1}$$

where a_i is in $[\frac{1}{\kappa d}, \frac{\kappa}{d}]$ for an absolute constant $\kappa \geq 1$ and satisfies $\sum_{i \in [d]} a_i = 1$, and $\{w_i^\star\}_{i=1}^d$ forms an orthonormal basis. Note that equation (1) can be also written as the sum of $2d$ neurons with of relu activation:

$$f^\star(x) = \sum_{i=1}^{d} a_i \left(\text{ReLU}(w_i^{\star \top} x) + \text{ReLU}(-w_i^{\star \top} x) \right).$$

We consider a training dataset of N i.i.d. samples $\mathcal{Z} = \{(x_j, y_j)\}_{j=1}^N$, where every x_j is drawn from the Gaussian distribution with identity covariance and $y_i = f^\star(x_j)$.

We learn the target network using an over-parametrized two-layer ReLU network with $m \geq 2d$ neurons $W = \{w_i\}_{i=1}^m$, given by:

$$f_W(x) = \frac{1}{m} \sum_{i=1}^{m} \|w_i\| \cdot \text{ReLU}(w_i^\top x). \tag{2}$$

Note that we have re-parametrized the output layer with the norm of the corresponding neuron, so that we only have one set of parameters W. This is without loss of generality because when $a_i \geq 0$, $a_i \cdot \text{ReLU}(w_i^\top x)$ is equal to $\|w_i'\| \cdot \text{ReLU}(w_i'^\top x)$ where $w_i' = \sqrt{a_i/\|w_i\|} \cdot w_i$. Given a training dataset $\mathcal{Z} = \{(x_i, y_i)\}_{i=1}^N$, we learn the target network by minimizing the following empirical loss:

$$\hat{L}(W) = \frac{1}{N} \sum_{i=1}^{N} \left(f_W(x_i) - y_i \right)^2.$$

Let $L(W)$ denote the population loss given by the expectation of the above equation over $\mathcal{Z}$.

Algorithm. We focus on truncated gradient descent with random initialization. Algorithm 1 describes the procedure.An interesting feature is that when a neuron becomes larger than a certain threshold, we no longer update the neuron. This is a variant of gradient clipping often used in training recurrent neural networks (e.g. Merity et al. (2017); Gehring et al. (2017); Peters et al. (2018)) — here we drop the gradients of the large weights instead of re-scaling them. The truncation allows us to upper bound the norm of every neuron. Our main result shows that Algorithm 1 learns the target network accurately in polynomially many iterations.

Theorem 1 (Main result) *Let $\mathcal{Z}$ be a training dataset with $N = \text{poly}_\kappa(d)$ samples generated by the model described above.[1] Let $C(\kappa)$ be a sufficiently large constant that only depends on κ. Let*

---

1. Let $\text{poly}(d)$ denote a polynomial of d and $\text{poly}_\kappa(d)$ denote a polynomial whose degree may depend on κ.

Algorithm 1 Truncated gradient descent for two-layer neural nets

---

Input: A training dataset $\mathcal{Z}$.
Require: Network width m, learning rate η, truncation parameters λ_0, λ_1.
Output: The final learned network $\hat{W} = \left\{ w_i \in \mathbb{R}^d \right\}_{i=1}^m$.

Initialization. Initialize $w_i^{(0)} \sim \mathcal{N}\left(0, \frac{1}{d} \cdot \mathrm{Id}_{d \times d}\right)$, for $1 \leq i \leq m$.

Stage 1. Let $\lambda = \lambda_0 = \Theta(\frac{1}{\mathrm{poly}(d)})$. For $t \leq \Theta\left(\frac{d^2}{\eta C(\kappa) \log d}\right)$, update every neuron as follows:

$$w_i^{(t+1)} = w_i^{(t)} - \eta \cdot \mathbb{1}_{\|w_i^{(t)}\|_2^2 \leq \frac{1}{\lambda_0}} \cdot \nabla_{w_i} \hat{L}(W), \text{ for every } 1 \leq i \leq m.$$

Stage 2. Let $\lambda = \lambda_1 = \Theta(\frac{1}{\mathrm{poly}_\kappa(d)})$. For $t \leq \Theta\left(\frac{d^{1+10Q}}{\eta}\right)$, update every neuron as follows:

$$w_i^{(t+1)} = w_i^{(t)} - \eta \cdot \mathbb{1}_{\|w_i^{(t)}\|_2^2 \leq \frac{1}{\lambda_1}} \cdot \nabla_{w_i} \hat{L}(W), \text{ for every } 1 \leq i \leq m.$$

$0 < Q < 1/100$ be a sufficiently small absolute constant that does not depend on κ. Let λ_0 be a sufficiently small value on the order of $1/\mathrm{poly}(d)$ and $\lambda_1 \leq \lambda_0/O(\mathrm{poly}_\kappa(d))$ be a sufficiently small value on the order of $1/\mathrm{poly}_\kappa(d)$. For a learning rate $\eta < \min\left(\lambda_0^2, O(1/\mathrm{poly}_\kappa(d))\right)$, a network width $m \geq \Omega(\mathrm{poly}(d)/\mathrm{poly}(\lambda_1))$, and truncation parameters λ_0, λ_1, let $\hat{W}$ be the final network learnt by Algorithm 1. With probability $1 - 1/\mathrm{poly}(d)$ over the choice of the random initialization, we have that the population loss of $\hat{W}$ satisfies

$$L(\hat{W}) \leq O(1/d^{1+Q}).$$

Our result builds on a connection between the popluation $L(W)$ and tensor decomposition for Gaussian inputs (Ge et al. (2017, 2018)). By expanding the population loss in the Hermite polynomial basis, the optimization problem becomes an infinite sum of tensor decompositions problems (cf. equation (4) in Section 2.) To analyze the gradient descent on the infinite sum tensor decomposition objective, we first analyze the case when m goes to infinity. We establish a conditional-symmetry condition on the population of neurons, which greatly simplifies the analysis. Informally, we leverage the fact that our input distribution is symmetric and our labeling function uses the absolute value activation, which is symmetric. Our analysis uncovers a stage-wise convergence of the gradient descent dynamic as follows, which matches our observations in simulations.

- First, Algorithm 1 minimizes the 0th and 2nd order tensor decompositions. Informally, the distribution of neurons is fitting to the 0th moment and the 2nd moment of $\{w_i^\star\}_{i=1}^d$.

- Second, Algorithm 1 minimizes the 4th and higher order tensor decompositions. Initially, there is a long plateau where the evolution is slow, but after a certain point gets faster. As a remark, this behavior has been observed for randomly initialized tensor power method (Anandkumar et al. (2017)). Because the solution to the 4th and higher order orthogonal tensor decomposition problems is unique, we can learn the ground truth weights $\{w_i^\star\}_{i=1}^d$.

Then we show that the approximation error between the population and infinite-width case and the finite-sample and finite-width case is small. The finite-width case can be viewed of as a sample of the infinite-width case. As the number of neurons increases, the approximation error reduces. In Section 3 and 4, we present a proof overview. The full proof is given in Section A and B.

As a complement, we show that the generalization error bound of Theorem 1 cannot be achieved by kernel functions with polynomially large feature map. Hence, by minimizing the higher order tensor decomposition terms, the learned neural network is provably more accurate than kernel functions which simply fit the lower order terms. Our result is stated as follows.

Theorem 2 (Lower bound) *Under either of the following two situations,*

1. *We use a feature map $\phi(x) : \mathbb{R}^d \to \mathbb{R}^N$ with $N = \mathrm{poly}(d)$*

2. *We use kernel method with any kernel $K : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ with $N = \mathrm{poly}(d)$ samples.*

There exists a set of orthonormal weights $\{w_i^*\}_{i \in [d]}$ and $\{a_i\}_{i \in [d]}$ where $a_i \in [\frac{1}{2d}, \frac{2}{d}]$ for all $1 \le i \le d$ satisfying $\sum_{i \in [d]} a_i = 1$, such that the following holds: With probability at least 0.999 over the training set $\mathcal{Z}$, for any $w_R, w_K \in \mathbb{R}^N$ with $\mathcal{R}(x) := w_R^\top \phi(x)$ and $\mathcal{K}(x) := w_K^\top [K(x, x_i)]_{i=1}^N$, the population loss of the feature map $\phi(x)$ and kernel $\mathcal{K}(x)$, denoted by $L(\mathcal{R}) = \mathbb{E}_{x \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})} (f^\star(x) - \mathcal{R}(x))^2$ and $L(\mathcal{K}) = \mathbb{E}_{x \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})} (f^\star(x) - \mathcal{R}(x))^2$, satisfies

$$L(\mathcal{R}) = \Omega\left(\frac{1}{d}\right) \quad and \quad L(\mathcal{K}) = \Omega\left(\frac{1}{d}\right). \tag{3}$$

Comparing the above result with Theorem 1, we conclude that provided with polynomially many samples, Algorithm 1 can recover the target two-layer neural network more accurately than the feature map and kernel method described above. Section C shows how to prove Theorem 2.

1.2. Related Work

Neural tangent kernel (NTK). A sequence of work shows that the learning process of gradient descent on over-parametrized neural networks, under certain initializations, reduces to the learning process of the associated neural tangent kernel. See Jacot et al. (2018); Arora et al. (2019b); Cao and Gu (2019); Du et al. (2018c); Arora et al. (2019a); Allen-Zhu and Li (2019b); Allen-Zhu et al. (2019c,b); Li and Liang (2018); Zou et al. (2018); Du et al. (2018a); Daniely et al. (2016); Ghorbani et al. (2019); Li et al. (2019a); Hanin and Nica (2019); Yang (2019) and the references therein. For NTK based results, the learning process of gradient descent can be viewed as solving (convex) kernel regression. Our work differs by analyzing a non-convex objective that involves an infinite sum of tensor decomposition problems. By analyzing the higher order tensor decompositions in our setting, we can achieve a smaller generalization error than kernel methods.

Allen-Zhu and Li (2019a, 2020a) show that over-parametrized neural networks can learn certain concept class more efficient than any kernel method. Their work assumes the target network satisfies a certain "information gap" assumption between the first and second layer, while our target network does not require such gaps. Allen-Zhu et al. (2019a); Bai and Lee (2019) go beyond NTK by studying quadratic approximations of neural networks. By contrast, our work analyzes higher-order tensor decompositions that cannot be achieved by quadratic approximations.

Two-layer neural networks given Gaussian inputs. There are a large body of works focusing on how to learn two-layer neural networks, such as Kawaguchi (2016); Soudry and Carmon (2016); Xie et al. (2016); Ge et al. (2017); Soltanolkotabi et al. (2017); Tian (2017); Brutzkus and Globerson (2017); Zhong et al. (2017); Boob and Lan (2017); Li et al. (2018); Vempala and Wilmes (2018); Ge et al. (2018); Bakshi et al. (2018); Oymak and Soltanolkotabi (2019); Yehudai and Shamir (2019);

Li and Yuan (2017); Zhang et al. (2018); Li and Liang (2017); Li et al. (2016); Li and Dou (2020); Allen-Zhu and Li (2020b) . Li and Yuan (2017); Zhong et al. (2017) consider learning two-layer networks with ReLU activations with a warm start (tensor initialization), as opposed to from a random initialization. Du et al. (2017) consider learning a target function consisting of a single ReLU activation. Brutzkus and Globerson (2017); Tian (2017) study the case where the weight vector for each neuron has disjoint support. Apart from the gradient descent algorithm, the method of moments has also been shown to be an effective strategy with provable guarantees (e.g. Bakshi et al. (2018); Ge et al. (2018)).

The closest work to ours is Ge et al. (2017) that considers a similar concept class. However, their work requires designing a complicated loss function, which is different from the mean squared loss. The learner network also uses a low-degree activation function as opposed to the ReLU activation. Ge et al. (2017) have stated the question of analyzing the gradient descent dynamic for minimizing the sum of second and fourth order tensor decompositions as a challenging open question. Our analysis not only applies to this setting, but also allows for more even order tensor decompositions. Apart from ReLU activations, quadratic activations have been studied in Li et al. (2018); Oymak and Soltanolkotabi (2019); Soltanolkotabi et al. (2017).

Infinite-width neural networks. Previous work such as Mei et al. (2018); Chizat and Bach (2018a) show that as the hidden layer size goes to infinity, gradient descent approaches the Wasserstein gradient flow. Mei et al. (2018) use tools from partial differential equations to prove the global convergence of the gradient descent. Both of these results do not provide explicit convergence rates. Wei et al. (2018) show that under a certain regularity assumption on the activation function, the Wasserstein gradient flow converges in polynomial iterations for infinite-width neural networks..

Organizations. The rest of the paper is organized as follows. In Section 2, we reduce our setting to learning a sum of tensor decomposition problems. In Section 3, we describe an overview of the analysis for the infinite-width case. In Section 4, we show how to connect the above case to the gradient descent dynamic on the empirical loss for polynomially-wide networks. Finally we validate our theoretical insight on simulations in Section 5. In Section A, we provide the proof of the infinite-width case. In Section B, we describe the error analysis of the above case and complete the proof of Theorem 1. In Section C, we present the proof of Theorem 2.

2. Preliminaries

Recall that the ground-truth weights $\{w_i^*\}_{i=1}^d$ forms an orthonormal basis. Since the input distribution $x \sim \mathcal{N}(0, I_{d \times d})$ and the initialization $\{w_i\}_{i=1}^m \sim \mathcal{N}\left(0, \frac{1}{d} \cdot \mathrm{Id}_{d \times d}\right)$ are all rotation invariant, without loss of generality we can assume that $w_i^* = e_i$, for all $1 \leq i \leq d$.

We can average out the randomness in x by applying Theorem 2.1 of Ge et al. (2017) on the loss function $L(W)$ (which expands the activations function in the Hermite basis).

$$
\begin{aligned}
L(W) = & c_0 \left\| \frac{1}{m} \sum_{i=1}^m \|w_i\|^2 - \sum_i a_i \right\|_F^2 + c_1 \left\| \frac{1}{m} \sum_{i=1}^m \|w_i\| w_i \right\|_F^2 + c_2 \left\| \frac{1}{m} \sum_{i=1}^m w_i^{\otimes 2} - \sum_i a_i e_i e_i^\top \right\|_F^2 \\
& + \sum_{j \geq 2} c_{2j} \left\| \frac{1}{m} \sum_{i=1}^m w_i^{\otimes 2} \otimes \bar{w}_i^{\otimes(2j-2)} - \sum_{i=1}^m a_i e_i^{\otimes 2j} \right\|_F^2,
\end{aligned} \tag{4}
$$

where $c_k = \frac{2[(k-3)!!]^2}{\pi \cdot k!} = \Theta(\frac{1}{k^3})$ is the Hermite coefficients of the absolute value function. We observe that the population risk can be written as a infinite sum of tensor decomposition problems. Intuitively, the 0-th order tensor decomposition term adjusts the l_2-norm of the weights. The k-order term concerns about the k-th moment of the weights.

The distribution of neuron weights. We start by considering an infinite-width neural network and then we extend the proof to polynomial-size neural networks. Following Wei et al. (2018), an infinite-width neural network specifies a distribution of its neuron weights. Let $\mathcal{P}$ be a distribution over $\mathbb{R}^d$. We can write equation (2) in terms of $\mathcal{P}$ as follows:

$$f_{\mathcal{P}}(x) = \mathop{\mathbb{E}}_{w \sim \mathcal{P}} \left[\|w\|_2 \cdot \text{ReLU}\left(w^\top x\right) \right]. \tag{5}$$

Correspondingly, the *population loss* of the neural network defined by $\mathcal{P}$ is given as

$$L_\infty(\mathcal{P}) = c_0 \left\| \mathop{\mathbb{E}}_{w \sim \mathcal{P}} \|w\|^2 - \sum_i^d a_i \right\|_F^2 + c_1 \left\| \mathop{\mathbb{E}}_{w \sim \mathcal{P}} w\|w\|_2 \right\|_F^2 + c_2 \left\| \mathop{\mathbb{E}}_{w \sim \mathcal{P}} w^{\otimes 2} - \sum_{i=1}^d a_i e_i e_i^\top \right\|_F^2$$

$$+ \sum_{j \geq 2} c_{2j} \left\| \mathop{\mathbb{E}}_{w \sim \mathcal{P}} w^{\otimes 2} \otimes \bar{w}^{\otimes(2j-2)} - \sum_{i=1}^d a_i e_i^{\otimes 2j} \right\|_F^2. \tag{6}$$

As a remark, for the case of finite neurons with weights W, the distribution $\mathcal{P}$ corresponds to a uniform distribution over $\{w_i\}_{i \in [m]}$.

Gradient descent update. It has been shown in prior works that gradient descent in the (natural) parameter space corresponds to Wasserstein gradient descent in the distributional space. However, we found that the Wasserstein gradient perspective is not particularly helpful for us to analyze our algorithms and therefore we work with the update in the parameter space W. The distribution $\mathcal{P}$ can be view as a collection of infinitesimal particles and the update of $\mathcal{P}$ can be viewed as the effect of updating each particle in the system. The update for each particle v, denoted by $\nabla_v L_\infty(\mathcal{P})$, is given by computing the gradient of the objective $L(W)$ w.r.t a particle v assuming the rest of the particles follow the distribution $\mathcal{P}$.

$$\nabla_v L_\infty(\mathcal{P}) := b_0 \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \|w\|_2^2 - 1 \right) v + b_1 \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \|w\|_2 w \|v\|_2 + \|w\|_2 \langle w, v \rangle \bar{v} \right) \tag{7}$$

$$+ b_2 \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \langle w, v \rangle w - \sum_{i=1}^d a_i \langle e_i, v \rangle e_i \right) + \sum_{j \geq 2} b_{2j} \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \langle w, v \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} w - \sum_{i=1}^d a_i \langle e_i, v \rangle \langle e_i, \bar{v} \rangle^{2j-2} e_i \right)$$

$$+ \sum_{j \geq 2} b'_{2j} \, \Pi_{v^\perp} \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \langle w, v \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} w - \sum_{i=1}^d a_i \langle e_i, v \rangle \langle e_i, \bar{v} \rangle^{2j-2} e_i \right), \tag{8}$$

where $b_0 = 4c_0, b_1 = 2c_1, b_{2j} = (4j) \times c_{2j} = \Theta\left(\frac{1}{j^2}\right), b_{2j'} = (4j - 4) \times c_{2j}$. We use $\nabla_v L_\infty$ and ∇_v as a shorthand for $\nabla_v L_\infty(\mathcal{P})$. Based on equation (8), we can further decompose $\nabla_v L_\infty(\mathcal{P})$ into the sum of $\nabla_{2j,v} L_\infty(\mathcal{P})$ for $j \geq 0$, where the $2j$-th term refers to the gradient of the $2j$-th tensor decomposition. As a direct consequence, if we are at a neural network described by $\mathcal{P}^{(t)}$, then after a truncated gradient descent update, the next neural network will be $\mathcal{P}^{(t+1)}$ such that

$$v^{(t+1)} \sim \mathcal{P}^{(t+1)} \Leftrightarrow v^{(t+1)} := v^{(t)} - \eta \mathbb{1}_{\|v^{(t)}\|_2^2 \leq \frac{1}{2\lambda}} \nabla_{v^{(t)}} L_\infty(\mathcal{P}^{(t)}), \text{ for } v^{(t)} \sim \mathcal{P}^{(t)}. \tag{9}$$

Finite-width case. We briefly describe the connection between the above distributional view and the finite-width case. Intuitively, we can think of the finite-width case as sampling m neurons randomly from the neuron population $\mathcal{P}$ in the infinite-width case. There are two sources of sampling error that arise from the above process: (i) the error of the gradients between the finite neuron distribution and the infinite neuron distribution; (ii) the error between the empirical loss and the population loss. Based of gradient truncation, the norm of every neuron is bounded by $1/\lambda$. Therefore, the sampling errors reduces as m and N increases, as shown in the following claim.

Claim 2.1 *For every $\lambda, \delta > 0$, for every distribution $\mathcal{P}$ over $\mathbb{R}^d$ supported on the ball $\{w \in \mathbb{R}^d \mid \|w\|_2^2 \leq \frac{1}{\lambda}\}$, let $W = \{w_i\}_{i=1}^m$ be i.i.d. random samples from $\mathcal{P}$. For every $\delta > 0$, with probability at least $1 - \delta$ over the randomness of W, we have that:*

$$|L(W) - L_\infty(\mathcal{P})| \leq \frac{\text{poly}\left(\frac{1}{\lambda}\right) \log \frac{1}{\delta}}{\sqrt{m}}.$$

With probability at least $1 - \delta$ over the randomness of $\{w_i\}_{i=1}^m$ and the training dataset $\mathcal{Z}$, for every $w \in W$, we have that:

$$\left\|\nabla_w \hat{L}(W) - \nabla_w L_\infty(\mathcal{P})\right\|_2 \leq \text{poly}\left(\frac{1}{\lambda}\right) \log \frac{m}{\delta} \left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{N}}\right).$$

Claim 2.1 can be shown via standard concentration inequalities such as the Chernoff bound.

Notations. Let $a = b \pm c$ denote a number within $[b - |c|, b + |c|]$. Let $[d]$ denote the set including $1, 2, \ldots, d$. Let $\text{Id}_{d \times d} \in \mathbb{R}^{d \times d}$ denote the identity matrix in dimension d. For two matrices A, B with the same dimensions, we use $\langle A, B \rangle = \text{Tr}[A^\top B]$ to denote their inner product. For a vector $w \in \mathbb{R}^d$, let $\|w\|_2$ denote its ℓ_2 norm and $\|w\|_\infty$ denote its ℓ_∞ norm. For $i \in [d]$, let w_i denote the i-th coordinate of w and w_{-i} denote the vector which zeroes out the i-th coordinate of w. We define $\bar{w} = \frac{w}{\|w\|_2}$ to be the normalized vector, and $\Pi_{w^\perp} = (\text{Id} - \bar{w}\bar{w}^\top)$ to be the projection onto the orthogonal complement of w. For a matrix M, let $\|M\|_2$ denote the spectral norm of a matrix M.

3. Overview of the Infinite-Width Case

We begin by studying Algorithm 1 for minimizing the population loss using an infinite-width neural network. The infinite-width case plays a central role in our analysis. First, the infinite-width case allows us to simplify the gradient update rule through a conditional-symmetry condition that we describe below. Second, the finite-width case can be reduced to the infinite-width case by bounding the approximation error of the two cases — we describe the reduction in Section 4.

A natural starting point for the infinite-width case is to simply set the network width m to infinity in Theorem 1. However, this will include negligible outliers such as those with large norms in the Gaussian distribution. Therefore, we focus on a truncated probability measure $\mathcal{P}^{(0)}$ of $\mathcal{N}(0, \text{Id}_{d \times d})$ by enforcing a certain bounded condition. The precise definition of $\mathcal{P}^{(0)}$ is presented in Definition 5. For the purpose of providing an overview of the analysis in this section, it suffices to think of $\mathcal{P}^{(0)}$ as satisfying the following conditional-symmetry condition.

Definition 3 (Conditional-symmetry) *We call a distribution $\mathcal{P}$ over $\mathbb{R}^d$ conditionally-symmetric if for every $i \in [d]$ and every $v \in \mathbb{R}^d$, the following is true.*

$$\Pr_{w \sim \mathcal{P}}[w_i = v_i \mid w_{-j} = v_{-j}] = \Pr_{w \sim \mathcal{P}}[w_i = -v_i \mid w_{-j} = v_{-j}]. \tag{10}$$

We will prove the following result for the infinite-width case.

Theorem 4 *In the setting of Theorem 1, let the number of samples N go to infinity. Starting from the initialization $W^{(0)} = \mathcal{P}^{(0)}$, let $\hat{W}$ be the final output network by Algorithm 1. The population loss of $\hat{W}$ satisfies $L(\hat{W}) \leq O(1/d^{1+Q})$.*

For the rest of this section, we present an overview of how to prove Theorem 4 and the proof detail can be found in Section A. First, we provide a simplifying formula for the gradient of $L_\infty(\mathcal{P})$. Then, recall that Algorithm 1 consists of two stages that differ by the gradient truncation parameter. We describe an overview of each stage in Section 3.1 and 3.2, respectively.

Now we show how to simplify the gradient of $L_\infty(\mathcal{P})$ in equation (8). Building on the setup of Section 2, we can view the weights $W^{(t)} = \mathcal{P}^{(t)}$ in the t-th iteration as a distribution over $\mathbb{R}^d$. Note that $\mathcal{P}^{(0)}$ is conditionally-symmetric by definition. Our main observation is that when $\mathcal{P}$ is conditionally-symmetric, the conditionally-symmetric property is preserved during the update.

Claim 3.1 *Suppose the update rule of $\mathcal{P}^{(t)}$ is given in equation (9). If $\mathcal{P}^{(t)}$ is conditionally-symmetric, then $\mathcal{P}^{(t+1)}$ is also conditionally-symmetric.*

To see that Claim 3.1 is true, we first observe that the 1-st order gradient is always zero when $\mathcal{P}^{(t)}$ is conditionally symmetric. Now, we further observe that for every neuron v in $\mathcal{P}^{(t)}$ and every $1 \leq j \leq d$, subject to v_{-j} being fixed, $\nabla_v L_\infty(\mathcal{P})$ is a polynomial of v_j that only involves odd degree monomials. Therefore, as long as $\mathcal{P}^{(t)}$ is conditionally-symmetric, then $\mathcal{P}^{(t+1)}$ is still conditionally-symmetric. Given the conditionally-symmetric property, we can simplify the gradient formula of equation (8) as folllows.

Claim 3.2 *Suppose that $\mathcal{P} = \mathcal{P}^{(t)}$ is conditionally-symmetric. For any $j \geq 0$, let $\nabla_{2j,v}$ be a shorthand for the gradient of the 2j-th tensor $\nabla_{2j,v} L_\infty(\mathcal{P})$. For any $1 \leq i \leq d$, let $[\nabla_{2j,v}]_i$ be the i-th coordinate of $\nabla_{2j,v}$. Depending on j, we have that $[\nabla_{2j,v}]_i$ is equal to the following*

$$[\nabla_{0,v}]_i = b_0 \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \|w\|_2^2 - 1 \right) \langle e_i, v \rangle, \quad [\nabla_{2,v}]_i = b_2 \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} w_i^2 - a_i \right) \langle e_i, v \rangle, \tag{11}$$

$$[\nabla_{2j,v}]_i = (b_{2j} + b'_{2j}) \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \left[\sum_{i_1, \cdots, i_j} \left(\prod_{r \in [j-1]} (\bar{w}_{i_r} \bar{v}_{i_r})^2 \right) (w_{i_j})^2 \langle e_{i_j}, v \rangle \right] - a_i \langle e_i, \bar{v} \rangle^{2j-2} \langle e_i, v \rangle \right)$$

$$- b'_{2j} \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \left[\|w\|_2^2 \sum_{i_1, \cdots, i_j} \prod_{r \in [j]} (\bar{w}_{i_r} \bar{v}_{i_r})^2 \right] - \sum_{r \in [d]} a_r \langle e_r, \bar{v} \rangle^{2j} \right) v_i, \forall j \geq 2. \tag{12}$$

The proof of Claim 3.2 is by applying Claim 3.1 to equation (8), which zeroes out the coordinates in w that has an odd order before taking the expectation over $w \sim \mathcal{P}$. In particular, the 1st order tensor in equation (6) becomes zero. For the 2nd order gradient $[\nabla_{2,v}]_i$, we have that

$$[\nabla_{2,v}]_i = b_2 \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} \langle w, v \rangle w_i - a_i \langle e_i, v \rangle e_i \right) = b_2 \left(\mathop{\mathbb{E}}_{w \sim \mathcal{P}} w_i^2 - a_i \right) v_i.$$

Similar arguments apply to higher order gradients.

Claim 3.1 and 3.2 together implies that for the infinite-width case, the gradient descent update is given by equation (11) and (12) . Another implication is that $[\nabla_v]_i$ equals $h(|v|) v_i$ for a fixed

function $h(\cdot)$, where $|v| \in \mathbb{R}^d$ takes the entry-wise absolute value of v. In other words, the updates of the coordinates of v are only correlated through the norm of v. This observation enables us to essentially consider each coordinate of v separately.

3.1. Dynamic during Stage 1

Stage 1.1: learning 0th and 2nd order tensors. We show that Algorithm 1 minimizes the 0th and 2nd order tensor decompositions of the objective L_∞ to zero first.

First, we show that the gradients from 4th and higher order tensors are dominated by $\nabla_{0,v}$ and $\nabla_{2,v}$. For the 0th and 2nd order tensors, their gradients for the i-th coordinate of $v \sim \mathcal{P}^{(0)}$ satisfy

$$|[\nabla_{0,v}]_i| + |[\nabla_{2,v}]_i| = \Theta\left(\frac{1}{d^{1.5}}\right). \tag{13}$$

This is because $\mathcal{P}^{(0)}$ is a suitable truncation of $\mathcal{N}(0, \mathrm{Id}_{d \times d}/d)$, hence we have that

$$\|v\|_\infty^2, \|\bar{v}\|_\infty^2 = \tilde{\Theta}\left(\frac{1}{d}\right) \text{ and } \left|\mathbb{E}_{w \sim \mathcal{P}^{(0)}}[w_i^2] - a_i\right| \le O_\kappa\left(\frac{1}{d}\right).$$

Applying the above to equation (11), we obtain equation (13). For higher order tensors, in Proposition 13, we show that for any $j \ge 2$, $|[\nabla_{2j,v}]_i| = \tilde{O}\left(1/d^{2.5}\right)$. Therefore, the 0th and 2nd order gradients indeed dominate the higher order gradients and Algorithm 1 is simply minimizing the 0th and 2nd order tensor terms of L_∞.

Based on the above observation, we show that the 0th and 2nd order tensor terms converges to approximately zero in Lemma 8. The main intuition is as follows. By equation (11), both the 0th and 2nd order gradient only depends on the i-th coordinates of neurons in $\mathcal{P}$. Hence the update can be viewed as d independent updates on the d dimensions. In Proposition 14, we show that throughout the algorithm, $|\mathbb{E}_{\mathcal{P}}\|w\|_2^2 - 1|$ is smaller than $\max_{i \in [d]}|\mathbb{E}_{\mathcal{P}} w_i^2 - a_i|$. Thus, it suffices to show that the 2nd order term $\max_{i \in [d]}\left|\mathbb{E}_{\mathcal{P}} w_i^2 - a_i\right|$ converges to zero. This case reduces to principal component analysis and in Proposition 15, we show that the 2nd order term indeed converges by a rate of $O_\kappa(1/d)$ using standard techniques.

As shown in Lemma 8, Stage 1.1 finishes within $\tilde{O}_\kappa(d/\eta)$ iterations, when eventually $|[\nabla_{0,v}]_i| + |[\nabla_{2,v}]_i|$ becomes $\tilde{O}(1/d^{2.5})$ for all $1 \le i \le d$, which is the same order as $|[\nabla_{2j,v}]_i|$ for $j \ge 2$. Thus, the process enters the next substage where the gradients of the higher order terms become effective.

Stage 1.2: learning higher order tensor decompositions. After the 0th and 2nd order terms converge to a small value, the gradients from higher order terms will start to dominate the update. In Lemma 9, we show that for a small fraction of neurons $v \in \mathbb{R}^d$, their norms become much larger than an average neuron — a phenomenon that we term as "winning the lottery ticket". We provide an informal analysis below.

In Proposition 16, we show that the gradient of most neurons v except a small fraction can be approximated by a signal term from the 4th order gradient plus a $1/d^2$ error term:

$$|[\nabla_v]_i| = \left(b_4 + b_4'\right) a_i \langle e_i, v \rangle \langle e_i, \bar{v} \rangle^2 \pm \frac{C_t(\kappa) \log d}{d^2} |v_i|, \tag{14}$$

where $C_t(\kappa)$ is a function of κ that grows with t. To intuitively see this, first, except for a fraction of $1/d^\alpha$ of the neurons, any neuron w satisfies $\|w\|_\infty^2 \le \alpha \log d/d$. Second, for the fraction of $1/d^\alpha$

of the neurons, because we clip the gradient when a neuron grows larger than $1/\lambda_0$, the norm of any neuron is no more than $1/\lambda_0$. Thus, for a sufficiently large α, the contribution of these neurons to the gradient is negligible. Combined together, we prove equation (14) in Proposition 16.

Based on equation (14), we reduce the dynamic to tensor power method. We observe that the update of v_i is approximately $v_i^{(t+1)} \approx v_i^{(t)} + \eta \cdot a_i \cdot (v_i^{(t)})^3$, which is analogous to power method for fourth order tensor decomposition. Hence for a larger initialization of $v_i^{(0)}$, the growth of $v_i^{(0)}$ is also faster. Based on the above intuition, we introduce the set of "basis-like" neurons $\mathcal{S}_{i,good}$ in the population $\mathcal{P}$, which are defined more precisely in Lemma 9. Intuitively, $\mathcal{S}_{i,good}$ includes neuron v that satisfies $[v_i^{(0)}]^2 \geq C^2 \log d/d$, which has probability measure at least $1/d^{C^2}$ by standard anti-concentration inequalities. Following equation (14), we show that the neurons in $\mathcal{S}_{i,good}$ keeps growing until they become roughly equal to $e_i/(\lambda_0 \operatorname{poly}(d))$.

As shown in Lemma 9, the above process goes through a long plateau of $O_\kappa(d^2/(\eta \operatorname{poly} \log(d))$ iterations, until the neurons of $\mathcal{S}_{i,good}$ are sufficiently large. Intuitively, the scaling of d^2 arises from the $1/d^2$ increment in equation (14). This concludes Stage 1 and the update of these basis-like neurons will be the focus of Stage 2.

3.2. Dynamic during Stage 2

In the second stage, we reduce the truncation parameter in Algorithm 1 from $\lambda_0 = \Theta(1/\operatorname{poly}(d))$ to a smaller value $\lambda_1 = \Theta(1/\operatorname{poly}_\kappa(d))$. This allows the neurons that are close to basis vectors to further fit the target network more accurately.

Stage 2.1: obtaining a warm start initialization. This substage reduces the gradient truncation parameter of the first stage. In Lemma 11, we show that after $\Theta(d \log d/\eta)$ iterations, the population loss reduces to less than $o(1/(d \log^{0.01} d))$. The proof of Lemma 11 involves analyzing the 0th and 2nd order tensor decompositions, similar to Stage 1.1.

At the end of Stage 2.1, the weights of the learner neural network form a "warm start" initialization, meaning that its population loss is less than $o(1/d)$ (Li and Yuan (2017); Zhong et al. (2017)). The final substage will show that the population loss can be further reduced from $o(1/(d \log^{0.01} d))$ to $O(1/d^{1+Q})$, where Q is a fixed constant defined in Theorem 1.

Stage 2.2: the final substage. In Lemma 12, we show that the population loss further reduces to $O(1/d^{1+Q})$ after $\Theta(d^{1+10Q}/\eta)$ iterations. We describe an informal argument by contrasting the gradient update of neurons in $\mathcal{S}_{i,good}$ and the rest of the neurons for a particular basis vector e_i.

For neurons $v \in \mathcal{S}_{i,good}$, in Claim A.10, we show that v follows the following update during the dynamic (cf. equation (60)):

$$[\nabla_v]_i \approx b_0 \left(\mathbb{E}_{\mathcal{P}} \|w\|_2^2 - 1 \right) v_i + b_2 \left(\mathbb{E}_{\mathcal{P}} w_i^2 - a_i \right) v_i - \eta \frac{c_t C(\kappa)}{d^2} v_i, \tag{15}$$

where c_t is a function that grows with t and $C(\kappa)$ is a fixed function of κ. For neurons $v \notin \mathcal{S}_{i,good}$, in Claim A.10, we show that v_i follows a similar update but its corresponding value of c_t is much smaller than the neurons in $\mathcal{S}_{i,good}$. Thus, the basis-like neurons grow faster than the rest of the neurons by an additive factor that scales with c_t/d^2.

Based on the above intuition, we analyze the dynamic following equation (15) using standard techniques for analyzing the convergence of gradient descent. In Lemma 12, we show for after $O(d^{1+Q}/\eta)$ iterations, the population loss of the 0th order tensor $b_0 \left(\mathbb{E}_{\mathcal{P}} \|w\|_2^2 - 1 \right)$ and the 2nd

order tensor $b_2 \left(\mathbb{E}_{\mathcal{P}} \, w_i^2 - a_i \right)$ reduces to below d^{1+Q}, which corresponds to the scaling of c_t/d in the end. This concludes Lemma 12.

Once Lemma 12 is finished, the gradient descent process has learned an accurate approximation of $f^\star(\cdot)$ and we can conclude the proof of Theorem 4. Specifically, we show that the population loss has also reduced to below d^{1+Q} (cf. equation (25)). Thus, we have finished the analysis of Algorithm 1 for $L_\infty(\mathcal{P})$. We provide the proof details of Theorem 4 in Section A.

4. Overview of the Finite-Width Case

Based on the analysis of the infinite-width case, we reduce the finite-width case to the infinite-width case. By applying Claim 2.1 with $\mathcal{P} = \mathcal{P}^{(t)}$, when $\{w_i^{(t)}\}_{i=1}^m$ are i.i.d. samples from $\mathcal{P}^{(t)}$, the empirical losses as well as their gradients are tightly concentrated around the population losses and their corresponding gradients. Furthermore, as we increase the number of neurons m and the number of samples N, the sampling errors reduce. Therefore, the goal of our reduction is to show that the sampling errors remain small throughout the iterations of Algorithm 1. We describe our reduction informally and leave the details to Section B.

We describe the connection between the dynamic of the finite-width case and the infinite-width case. For a neuron $w^{(t)}$ sampled from $\mathcal{P}^{(t)}$, our analysis in Section 3 has shown the dynamic of $w^{(t)}$ in the infinite-width case starting from $w^{(0)}$. In the finite-width case, we introduce the notation $\tilde{w}^{(t)}$ as the t-th iterate starting from the same initialization $w^{(0)}$ using Algorithm 1. Our goal is to show that $\xi_w^{(t)} := \tilde{w}^{(t)} - w^{(t)}$ does not blow up exponentially large before Algorithm 1 finishes.

Based on the above connection, in Stage 1, we show that the propagation of the error $\xi_w^{(t)}$ remains polynomially small in Lemma 18. Our analysis involves a bound on the average error of all neurons $\mathbb{E}_{w \sim W}[\|\xi_w^{(t+1)}\|_2^2]$ and a bound on the individual error of every neuron $\max_{w \in W} \|\xi^{(t+1)}\|_2^2$. First, in Proposition 23, we show that it suffices to consider the first order errors in $\xi_w^{(t+1)}$, i.e. those that involve at most one of $\xi_w^{(t)}$. Based on the result, in Proposition 21 and 22, we bound the average error and individual error as

$$
\mathbb{E}_{w \sim W}[\|\xi_w^{(t+1)}\|_2^2] \leq (1 \pm o(1)) \left(1 + \frac{\eta}{\text{poly}(d)} \right) \mathbb{E}[\|\xi_w^{(t)}\|_2^2],
$$
$$
\max_{w \in W} \|\xi_w^{(t+1)}\|_2^2 \leq \text{poly}(d) \, \mathbb{E}_{w \sim W} \|\xi_w^{(t+1)}\|_2^2.
$$

Combined together, we show in Lemma 18 that $\xi_w^{(t)}$ indeed remains polynomially small. For Stage 2, we analyze the propagation of error in Lemma 19 and 20 using similar arguments.

Combining the above three lemmas on error propagation and Theorem 4, we complete the proof of Theorem 1 in Section B.

5. Simulations

We provide simulations to complement our theoretical result. We consider a setting where $w_i^\star = e_i$ and $a_i = 1/d$, for $1 \leq i \leq d$. The input is drawn from the Gaussian distribution. For the ith order tensor, we measure the corresponding tensor term from the loss $L(W)$.

Stage-wise convergence. We validate the insight of our analysis, which shows that the convergence of gradient descent has two stages. We use the labeling function of equation (1) and a learner
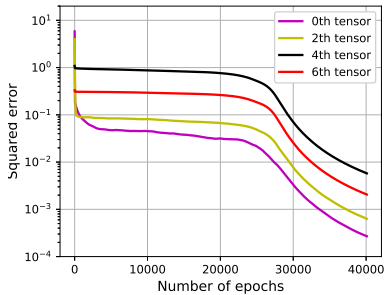
Figure 1: Illustrating the convergence of each tensor during the gradient descent dynamic using absolute value activations.
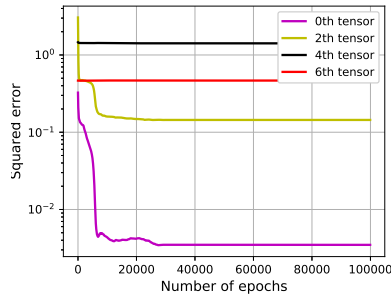

Figure 2: For properly parametrized gradient descent, the 4th and 6th order tensors get stuck using absolute value activations.

network with absolute value activation functions as in Section 3 and A to validate the dynamic. In the first stage, the 0th and 2nd order tensors converge to zero quickly. In the second stage, the 4th and higher order tensors converge to zero followed by a long plateau. Figure 1 shows the result. Here we use $d = 30$ and $m = 100 > 2d$. The number of samples is 10^4.

We can see that initially, 0th and 2nd order tensor has higher loss than the 4th and higher ones. However, at the beginning of the training, both the 0th and the 2nd order tensors decrease significantly from the initial value and converge to below 10^{-1} very quickly. Moreover, after a quick warm up period, the 0th order term always stays smaller than the 2nd term, as our theory predicts. This is followed by a long plateau, which corresponds to Stage 1.2 of our analysis. During this stage, the 4th and higher order tensors dominate the loss function, and the gradient descent is simulating tensor power method to obtain basis-like neurons. Eventually, the neural network accumulates enough basis-like neurons from the 4th and higher tensors in the objective function. The 4th and higher order tensors reduce to the order of 10^{-2}. The 0th and 2nd order tensors further reduce to closer to zero. Our theory provides an in-depth explanation of this phenomenon.

Over-parametrization is necessary. It has been observed that for properly parameterized gradient descent, i.e. for $m = 2d$, gradient descent gets stuck (Ge et al. (2017); Du et al. (2018b)). We consider the same setting as the previous experiment but use $m = 2d$. Figure 2 shows the result. We can see that the 0th order tensor term still converges close to zero. However, the 2nd, 4th and 6th order tensor terms all get stuck even after 10^5 epochs.

6. Conclusions and Discussions

In this work, we have shown that for learning a target network with absolute value activations, truncated gradient descent can provably converge in polynomially many iterations starting from a random initialization. The learned network is more accurate compared to kernel functions with polynomially large feature mappings. It would be interesting to extend our result to settings where the input layer weights $W^\star$ are not orthonormal. A fundamental challenge is to analyze the gradient descent dynamic beyond orthogonal tensors.

References

Zeyuan Allen-Zhu and Yuanzhi Li. What Can ResNet Learn Efficiently, Going Beyond Kernels? In *NeurIPS*, 2019a.

Zeyuan Allen-Zhu and Yuanzhi Li. Can SGD Learn Recurrent Neural Networks with Provable Generalization? In *NeurIPS*, 2019b. Full version available at http://arxiv.org/abs/1902.01028.

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020a.

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190*, 2020b.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. In *NeurIPS*, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *NeurIPS*, 2019b.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019c.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics in overcomplete regime. *The Journal of Machine Learning Research*, 18(1):752–791, 2017.

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*, 2019a.

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, 2019b. URL http://arxiv.org/abs/1901.08584.

Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. *arXiv preprint arXiv:1811.01885*, 2018.

Digvijay Boob and Guanghui Lan. Theoretical properties of the global optimizer of two layer neural network. *arXiv preprint arXiv:1710.11241*, 2017.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10835–10845, 2019.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*. arXiv preprint arXiv:1805.09545, 2018a.

Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 8, 2018b.

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.

Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *arXiv preprint arXiv:1712.00779*, 2017.

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 2018a.

Simon S. Du, Jason D. Lee, Yuandong Tian, Barnabás Póczos, and Aarti Singh. Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima. In *International Conference on Machine Learning (ICML)*. http://arxiv.org/abs/1712.00779, 2018b.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018c.

Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems*, pages 3653–3663, 2017.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points: online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. *arXiv preprint arXiv:1810.06793*, 2018.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org, 2017.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019.

Jeff Z. HaoChen, Colin Wei, Jason D. Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance, 2020.

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.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

Yuanzhi Li and Zehao Dou. When can wasserstein gans minimize wasserstein distance? *arXiv preprint arXiv:2003.04033*, 2020.

Yuanzhi Li and Yingyu Liang. Provable alternating gradient descent for non-negative matrix factorization with strong correlations. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2062–2070. JMLR. org, 2017.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

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.

Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of non-negative matrix factorization via alternating updates. In *Advances in neural information processing systems*, pages 4987–4995, 2016.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *COLT*, 2018.

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*, 2019a.

Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019b.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, 2017.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

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.

Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Ruoyu Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

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.

Santosh Vempala and John Wilmes. Polynomial convergence of gradient descent for training one-hidden-layer neural networks. *arXiv preprint arXiv:1805.02677*, 2018.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369*, 2018.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9709–9721, 2019.

Bo Xie, Yingyu Liang, and Le Song. Diversity leads to generalization in neural networks. *arXiv preprint Arxiv:1611.03131*, 2016.

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.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*, 2019.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. *arXiv preprint arXiv:1806.07808*, 2018.

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.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

Organizations. The appendix provides complete proofs to Theorem 1 and 2.

- In Section A, we describe the proof of Theorem 4 for the infinite-width case. This section comprises the bulk of the appendix.

- In Section B, we describe the proof of Theorem 1, by reducing the finite-width case to the infinite-width case.

- In Section C, we prove Theorem 2 using ideas from the work of Allen-Zhu and Li (2019a).

Appendix A. Proof of the Infinite-Width Case

We provide the proof of Theorem 4, which shows that running truncated gradient descent on an infinite-width network can recovery the target network with population loss at most d^{1+Q}, where Q is a sufficiently small constant defined in Theorem 4. Recall from Section 3 that our analysis begins by setting up the random initialization, and then proceeds in two stages. We fill in the proof details missing from Section 3. The rest of this section is organized as follows.

- **Initialization:** We set up the random initialization of the neuron distribution.

- **Stage 1:** We fill in the proof details of the dynamic during Stage 1, which subsumes Stage 1.1 and Stage 1.2 described in Section 3.1. This stage runs for $\Theta(\frac{d^2}{\eta C(\kappa) \log d})$ iterations.

- **Stage 2:** We fill in the proof details of the dynamic during Stage 2, which subsumes Stage 2.1 and Stage 2.2 described in Section 3.2. This stage runs for $\Theta(\frac{d^{1+10Q}}{\eta})$ iterations.

Initialization. Recall that for the infinite-width case, our initialization of the neuron distribution is a probability measure truncated from a Gaussian distribution with identity covariance. We formally define the truncation and the initialization, denoted by $\mathcal{P}^{(0)}$, as follows.

Definition 5 (Truncated neuron space) *Let $\mathcal{S}_g \subseteq \mathbb{R}^d$ be the set of all $w \in \mathbb{R}^d$ that satisfy the following properties:*

- *The maximum entry of w is bounded: $\|w\|_\infty \leq \frac{\text{poly} \log(d)}{\sqrt{d}}$.*

- *The entrywise square of w satisfies*

$$
\|w\|_2^2 \text{ and } \sum_{i=1}^{d} a_i d \cdot w_i^2 \text{ are both in the range } \left[1 - \frac{\text{poly} \log(d)}{\sqrt{d}}, 1 + \frac{\text{poly} \log(d)}{\sqrt{d}} \right]. \quad (16)
$$

- *There are at most $O(\log^{0.01}(d))$ coordinates $i \in [d]$ of w such that $w_i^2 \geq \frac{\log d}{d}$.*

We define $\mathcal{P}^{(0)}$ as the probability measure of $\mathcal{N}(0, \text{Id}_{d \times d} / d)$ conditional on the support set $\mathcal{S}_g$.

Remark. For our purpose of proving the finite-width case later in Section B, it suffices to consider $\mathcal{P}^{(0)}$ as the initialization as opposed to $\mathcal{N}(0, \text{Id}_{d \times d} / d)$. This is because when Algorithm 1 samples $m = \text{poly}_\kappa(d)$ neurons from $\mathcal{N}(0, \text{Id}_{d \times d} / d)$, with high probability all the m samples are in the set $\mathcal{S}_g$. To see this, by standard concentration inequalities for the Gaussian distribution, we can show

that the set $\mathcal{S}_g$ has probability measure at least $\mu(\mathcal{S}_g) \geq 1 - \frac{1}{d^{\omega(1)}}$. Thus by union bound, with high probability all m samples are in $\mathcal{S}_g$.

As stated in Section 3, we are going to heavily use the conditionally-symmetric property defined in Definition 3 throughout the analysis of the population loss $L_\infty(\mathcal{P})$. We observe that the initialization $\mathcal{P}^{(0)}$ is indeed conditionally-symmetric. This is because $\mathcal{N}(0, \mathrm{Id}_{d \times d}/d)$ satisfies the conditionally-symmetric property and our truncation in Definition 5 only involves restrictions on the squares of the coordinates of w. Hence the truncation of $\mathcal{N}(0, \mathrm{Id}_{d \times d}/d)$ to $\mathcal{S}_g$ preserves the conditionally-symmetric condition.

Notations for gradients. Before describing the analysis, we first introduce several notations for the gradients of neurons. Recall from Claim 3.2 that the gradient of a neuron v in the distribution $\mathcal{P}$ can be simplified under the conditionally-symmetric property. For each coordinate $1 \leq i \leq d$, the gradient of neuron v satisfies that $[\nabla_v]_i = \sum_{j \geq 0} [\nabla_{2j,v}]_i$, where $\nabla_v = \nabla_v L_\infty(\mathcal{P})$, $\nabla_{2j,v} = \nabla_{2j,v} L_\infty(\mathcal{P})$ denotes the gradient of v for the $2j$-th tensor, and $[\nabla_v]_i$ denotes the i-th coordinate of ∇_v. To state the proofs, we require the following notation. Let $B_{1,2j} = b_{2j} + b'_{2j}$ and $B_{2,2j} = b'_{2j}$, where b_{2j} and b'_{2j} are the Hermite coefficients of the $2j$-th tensor given in Section 2. For a vector $w \in \mathcal{S}_g$, let $w^{(0)}$ denote a neuron with initialization w in the initialization $\mathcal{P}^{(0)}$. Let $\mathcal{P}^{(t)}$ denote the t-th iterate of $\mathcal{P}^{(0)}$ following the update rule of equation (9).

Stage 1. Recall from Section 3.1 that the goal of Stage 1 is to show that a small fraction of neurons becomes basis-like, i.e. close to a basis e_i times a scaling factor of $\mathrm{poly}(d)$ at the end of $\Theta_\kappa(d^2/\eta \log d)$ iterations. To facilitate the analysis, we maintain a running inductive hypothesis throughout Stage 1 that provides an upper bound on the norm of a typical neuron during the update. We first introduce the set of neurons that will not become basis-like by the end of Stage 1.

Definition 6 *Let C_0 be a large enough constant. Let $c_0 = C_0 \log d$ and $\mathcal{S}$ be the set of all vectors w in $\mathcal{S}_g$ such that*

$$\|w\|_\infty^2 \leq \frac{c_0}{d} \text{ and } \|\bar{w}\|_\infty^2 \leq \frac{c_0}{d},$$

where $\bar{w} = w/\|w\|$ denotes w after being normalized.

Based on the above definition, we introduce the following inductive hypothesis that shows the neurons in $\mathcal{S}$ remain "small and dense" (i.e. not basis-like) throughout Stage 1. This stage runs for $\Theta(\frac{d^2}{\eta \log d})$ iterations. We use κ_1 to denote a value on the order of $\Theta(\exp(\mathrm{poly}(\kappa)))$.

Proposition 7 (Inductive hypothesis $\mathcal{H}_1$ for Stage 1) *In the setting of Theorem 4, let the number of iterations in the second stage be $T_2 = \Theta(\frac{d^2}{\eta c_0 \exp(\mathrm{poly}(\kappa))})$. There exists an increasing sequence $\{c_t\}_{t=1}^{T_2}$ where $c_t \leq \exp(\mathrm{poly}(\kappa)) \log d$ such that for every neuron $w^{(0)} \in \mathcal{S}$, and every $t \leq T_2$, the t-th iterate of the neuron $w^{(t)}$ satisfies that*

$$\|w^{(t)}\|_\infty^2 \leq \frac{c_t}{d}, \|\bar{w}^{(t)}\|_\infty^2 \leq \frac{c_t}{d}. \tag{17}$$

Furthermore, for every coordinate $i \in [d]$, we have that in expectation,

$$\mathop{\mathbb{E}}_{w^{(t)} \sim \mathcal{P}^{(t)}} [w_i^{(t)^2}] \leq \frac{2\kappa}{d} \text{ and } \mathop{\mathbb{E}}_{w^{(t)} \sim \mathcal{P}^{(t)}} [\bar{w}_i^{(t)^2}] \leq \frac{4\kappa^2}{d}. \tag{18}$$

Equation (17) and (18), which we also refer to as the inductive hypothesis $\mathcal{H}_1$, show that the neurons in $\mathcal{S}$ will not grow beyond $O_\kappa(\log d/d)$. Hence they will not become basis-like at the end of Stage 1.

The set $\mathcal{S}$ consists of most of neurons in $\mathcal{P}^{(0)}$ because by standard anti-concentration inequalities, the measure of the set $\mathcal{S}$ is at least $1 - d^{-O(C_0)}$. Hence, $1 - \mu(\mathcal{S})$ is at most $d^{-O(C_0)}$. Based on this fact, we state a simple fact on the norm of neurons that are not in $\mathcal{S}$ that will be used later in the analysis:

$$\mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}, w^{(0)} \notin \mathcal{S}} \|w^{(t)}\|_2^2 \leq \Lambda := O\left(\frac{1}{\lambda_0}(1 - \mu(\mathcal{S}))\right) \leq \frac{1}{\mathrm{poly}(d)}. \tag{19}$$

To see that equation (19) is true, recall that the truncation of Algorithm 1 ensures that $\|w\|^2 \leq 1/\lambda_0$. Combined with the fact that $1 - \mu(\mathcal{S}) \leq d^{-O(C)}$ and $\lambda_0 = \Theta(1/\mathrm{poly}(d))$, we have that equation (19) holds for a sufficiently large constant C_0. This finishes our introduction of the inductive hypothesis $\mathcal{H}_1$. The proof of Proposition 7 can be found in Section A.2.2.

Given the inductive hypothesis $\mathcal{H}_1$, we can state the formal result that corresponds to Stage 1.1 in Section 3.1. For a neuron distribution $\mathcal{P}$, let us first introduce the following notations, which corresponds to the population loss of the 0th and 2nd order tensors.

$$\Delta_+ := b_0 \sum_{i \in [d]} [\mathbb{E}_{w \sim \mathcal{P}} w_i^2 - a_i]^+, \Delta_- := b_0 \sum_{i \in [d]} [a_i - \mathbb{E}_{w \sim \mathcal{P}} w_i^2]^+,$$

$$\delta_+ := b_2 \max_{i \in [d]} [\mathbb{E}_{w \sim \mathcal{P}} w_i^2 - a_i]^+, \delta_- := b_2 \max_{i \in [d]} [a_i - \mathbb{E}_{w \sim \mathcal{P}} w_i^2]^+.$$

Let $\Delta := \Delta_+ - \Delta_-$ denote an absolute upper bound on the loss of the 0th tensor. Let $\delta_i := b_2(\mathbb{E}_\mathcal{P} w_i^2 - a_i)$ for every $1 \leq i \leq d$. At the t-th iteration, we use $\delta_i^{(t)}$ to denote the value of δ_i given the neuron distribution $\mathcal{P}^{(t)}$, as well as $\Delta^{(t)}$ for Δ, $\delta_+^{(t)}$ for δ_+, and $\delta_-^{(t)}$ for δ_-.

Based on the above notations, we show the following convergence result at the end of Stage 1.1.

Lemma 8 (Stage 1.1: learning 0th and 2nd order tensors) *In the setting of Theorem 4, suppose that Proposition 7 holds. Let $T_1 = \Theta\left(\frac{\mathrm{poly}(\kappa_1) d \log d}{\eta}\right)$. Then, for every $t \geq T_1$, we have that $\Delta^{(t)}, \delta_+^{(t)}, \delta_-^{(t)}$ are all less than $\frac{c_t \mathrm{poly}(\kappa_1)}{d^2}$, where c_t is given in Proposition 7.*

The above result implies that after T_1 iterations, the population loss of equation (6), which can be upper bounded by $\Delta^{(t)} + \delta_+^{(t)} + \delta_-^{(t)}$, remains less than $O(c_t \mathrm{poly}(\kappa_1)/d^2)$. The proof of Lemma 8 can be found in Section A.2.1.

Once Stage 1.1 is finished, recall from Section 3 that the higher order tensors begin to dominate the gradient update. Hence Algorithm 1 enters Stage 1.2. We state the formal result that corresponds to Stage 1.2 in Section 3 below. Let us introduce the following notations in order to state the formal result. Let $T_2' = T_2 - \frac{d^2}{\eta \mathrm{poly} \log(d)}$. For every $1 \leq i \leq d$, let $\Gamma_i = \frac{1}{2B_{1,4}(a_i^2 d)(\eta T_2')}$. Let $\rho = \frac{\mathrm{poly}(\kappa_1) \cdot \log d}{d}$. Here, by our assumption, we know that $a_i^2 = \Theta(1/d^2)$, $T_2' = \Theta(d^2/(\eta \log d))$, so we can see that $\Gamma_i = \Theta(\log d/d)$. Consider a coordinate $i \in [d]$. We define the set of good neurons whose i-th coordinate is larger than $\Gamma_i + \rho$ as

$$\mathcal{S}_{i,good} := \left\{v \in \mathcal{S}_g \mid [v^{(0)}]_i^2 \geq \Gamma_i + \rho \text{ and for all other } j : [v^{(0)}]_j^2 < \Gamma_j - \rho\right\}.$$

19

Then we define the set of bad neurons that have two large coordinates as

$$\mathcal{S}_{i,bad} = \left\{ v \in \mathcal{S}_g \mid [v^{(0)}]_i^2 \geq \Gamma_i - \rho \text{ and there exists } r \neq i : [v^{(0)}]_r^2 \geq \Gamma_r - \rho \right\}.$$

The following lemma shows that, among other statements, the neurons in $\mathcal{S}_{i,good}$ will win the lottery and become basis-like at the end of Stage 1.2 in the sense described below.

Lemma 9 (Stage 1.2: learning higher order tensors) *In the setting of Theorem 4, suppose that Proposition 7 holds. At iteration T_2, the following holds for $\mathcal{S}_{i,good}$ and $\mathcal{S}_{i,bad}$:*

- *For every $i \in [d]$ and $v \in \mathcal{S}_{i,good}$, we have that*

$$|v_i^{(T_2)}|^2 \geq \frac{1}{\lambda_0 \operatorname{poly}(d)} \geq \operatorname{poly}(d), \text{ and for every } j \neq i, |v_j^{(T_2)}| \leq \frac{2(\log d)^2}{\sqrt{d}}.$$

- *For every $1 \leq i \leq d$ and every $v \in \mathcal{S}_g$, if there exists $j \neq i$ such that $|v_i^{(T_2)}|$ and $|v_j^{(T_2)}|$ are both greater than $\frac{2(\log d)^2}{\sqrt{d}}$, then the neuron v is in both $\mathcal{S}_{i,bad}$ and $\mathcal{S}_{j,bad}$.*

- *For every $i \in [d]$, the probability measure of $\mathcal{S}_{i,good}$ and $\mathcal{S}_{i,bad}$ satisfies that*

$$\mu(\mathcal{S}_{i,good}) \geq d^{-\exp(\operatorname{poly}(\kappa_1))} \text{ and } \mu(\mathcal{S}_{i,good}) \geq \mu(\mathcal{S}_{i,bad}) \cdot d^{\exp(\operatorname{poly}(\kappa_1))}.$$

In the above result, the set $\mathcal{S}_{i,good}$ contains neurons that become approximately a large scaling of the basis e_i after T_2 iterations, a phenomenon that we term as winning the lottery ticket. These neurons become much larger than the neurons in $\mathcal{S}$, which are bounded by $O_\kappa(\log d/d)$. The set $\mathcal{S}_{i,bad}$ contains neurons whose coordinate i might be large in the end, but not close to a basis – In other words, these are neurons that "win the lottery ticket" but not close to a scaling of e_i. The final statement in this lemma shows that the measurements of such neurons is small, comparing to those who are basis-like. Lemma 9 is proved in Section A.3. This concludes Stage 1.

Stage 2. The second stage begins by reducing the gradient truncation parameter from $\lambda_0 = \Theta(\frac{1}{\operatorname{poly}(d)})$ to $\lambda_1 = \Theta(\frac{1}{\operatorname{poly}_\kappa(d)})$.[2] Recall from Section 3.2 that the goal of Stage 2 is to allow basis-like neurons to grow until they fit the target network with population loss at most $O(d^{1+Q})$.

- The first substage of the analysis shows that the population loss reduces below $O(\frac{1}{d \log^{0.01} d})$, after $T_3 = \Theta(d \log d/\eta)$ many iterations.

- The second substage of the analysis shows that the population loss further reduces below $O\left(1/d^{1+Q}\right)$, after $T_4 = \Theta(d^{1+10Q}/\eta)$ many iterations.

To facilitate the analysis, we introduce a running inductive hypothesis throughout Stage 2 that describes the behavior of the good and bad neurons. Let us introduce several notations first. Let the union of the bad neurons for all coordinates be given by

$$\mathcal{S}_{bad} := \{v \in \mathcal{S}_g \mid \exists i \neq j \in [d], [v^{(0)}]_i^2 \geq \Gamma_i - \rho \text{ and } [v^{(0)}]_j^2 \geq \Gamma_j - \rho\}.$$

---

2. As a remark, the rational for this technical twist is that the neurons do not grow too large during the process. This is useful for the error analysis later in the finite-width case.

The set of potential neurons is given by

$$\mathcal{S}_{i,pot} = \left\{ v \in \mathcal{S}_g \mid [v^{(0)}]_i^2 \geq \Gamma_i - \rho \right\}.$$

We remark that these are the set of neurons whose coordinate i can have a magnitude larger than $\mathrm{poly}\log(d)/\sqrt{d}$ at the end of Stage 1 (cf. Section A.2.1). As another remark, the set of good neurons, which grows much larger larger than $\mathrm{poly}\log(d)/\sqrt{d}$ for a particular coordinate, is a subset of potential neurons. Let the union of the potential neurons for all coordinates be given by

$$\mathcal{S}_{pot} := \{v \in \mathcal{S}_g \mid \exists i \in [d], [v^{(0)}]_i^2 \geq \Gamma_i - \rho\}.$$

We maintain the following running hypothesis that, among other things, specifies the behavior of the potential, good, and bad neurons in detail.

Proposition 10 (Inductive hypothesis $\mathcal{H}_2$ for Stage 2) *In the setting of Theorem 4, there exists $c_{T_2} = \mathrm{poly}(\log d)$ and a monotonically increasing sequence $c_{T_2} \leq c_t \leq d^{O(Q)} \leq d^{1/10}$ such that for every $T_2 < t \leq T_4$, the following list of properties hold for the neuron distribution $\mathcal{P}^{(t)}$ at the t-th iteration.*

1. *For every $v \in \mathcal{S}_g$, $\|v^{(t)}\|_2^2 \leq 1/\lambda_1$. As a result, gradient truncation never happens during this stage.*

2. *For every $v \notin \mathcal{S}_{pot}$,*

$$\|\bar{v}^{(t)}\|_\infty^2 \leq \frac{c_t}{d} \text{ and } \|v^{(t)}\|_\infty^2 \leq \frac{c_t}{d}. \tag{20}$$

   For every $i \in [d]$, every $v \in \mathcal{S}_{i,pot} \backslash \mathcal{S}_{bad}$, and $j \neq i$

$$\|v_j^{(t)}\|_2^2 \leq \frac{c_t}{d}. \tag{21}$$

3. *The mass of the set of bad neurons is small:*

$$\mathop{\mathbb{E}}_{v^{(t)} \sim \mathcal{P}^{(t)}, v \in \mathcal{S}_{bad}} \|v^{(t)}\|_2^2 \leq \frac{1}{\mathrm{poly}(d)}. \tag{22}$$

4. *For every $i \in [d]$ and every $v \in \mathcal{S}_{i,good}$, we have that $\|v_i^{(t)}\|_2^2 \geq \frac{1}{\lambda_0 \mathrm{poly}(d)}$.*

5. *For every $i \in [d]$, the following masses regarding the potential and bad neurons hold:*

$$\gamma_i^{(t)} := \mathop{\mathbb{E}}_{v^{(t)} \sim \mathcal{P}^{(t)}, v \in \mathcal{S}_{i,pot} \backslash \mathcal{S}_{bad}} v_i^{(t)2} \leq \frac{\mathrm{poly}(\kappa_2)}{d}, \tag{23}$$

$$\beta_i^{(t)} := \mathop{\mathbb{E}}_{v^{(t)} \sim \mathcal{P}^{(t)}, v \notin \mathcal{S}_{pot}} v_i^{(t)2} \leq \frac{\mathrm{poly}(\kappa_2)}{d}. \tag{24}$$

   where κ_2 denotes $\exp(\mathrm{poly}(\kappa_1))$ and κ_1 denotes $\exp(\mathrm{poly}(\kappa))$.

We remark that in the above inductive hypothesis, equation (20) and (21) show similar conditions as equation (17) provided in Proposition 7. For the rest of the section, we refer to the conclusion of Proposition 10 as inductive hypothesis $\mathcal{H}_2$. The proof of Proposition 10 can be found in Section A.4.1.

Given the inductive hypothesis, we can state the formal result that corresponds to Stage 2.1 in Section 3.1. We introduce the notation $\Delta^{(t)} = 2b_0 \left(\sum_{i=1}^{d} (\gamma_i^{(t)} + \beta_i^{(t)}) - \sum_{i=1}^{d} a_i \right)$ that measures the average error of the neurons across all coordinates. We show that by the end of $T_3 = T_2 + \Theta(d \log d / \eta)$ iterations, we have obtained a warm start neuron distribution for Δ, $\left\{ \beta_1^{(t)}, \ldots, \beta_d^{(t)} \right\}$, and $\left\{ \gamma_1^{(t)}, \ldots, \gamma_d^{(t)} \right\}$. We state the result below.

Lemma 11 (Stage 2.1: Obtaining a warm start initialization) *In the setting of Theorem 4, suppose Proposition 10 holds. There exists an iteration $T_3 = T_2 + \Theta(d \log d / \eta)$ such that at iteration $t = T_3$, the following holds:*

$$\text{For any } i \in [d], \ \ \beta_i^{(t)} \leq \frac{1}{d \log^{0.01} d}, \ \ \ |a_i - \gamma_i^{(t)}| \leq \frac{1}{d \log^{0.01} d}; \ \ and \ |\Delta^{(t)}| \leq \frac{1}{d \log^{0.01} d}.$$

The above result implies that the set of potential neurons has fit the i-th coordinate of the target network by an error less than $o(1/d)$. The population loss of the 0th order tensor is also reduced below $o(1/d)$. The proof of Lemma 11 can be found in Appendix A.3.

In the end, we describe the formal result that corresponds to Stage 2.2 in Section 3.2. We construct a potential function to show that $\beta_i^{(t)} + \gamma_i^{(t)}$ converges to a_i beyond iteration $t = T_3$. After running for $T_4 = T_3 + \Theta(\frac{d^{1+10Q}}{\eta})$ many iterations, we show that a certain set of potential neurons has converged to a_i with error at most $O(1/d^{2+Q})$, for every $1 \leq i \leq d$.

The result is shown in Lemma 12 below. We need the following notations for defining the potential function:

$$\delta_-^{(t)} = \max \left\{ \max_{i \in [d]} \left\{ C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + \frac{C_2 \gamma_i^{(t)}}{\beta_i^{(t)} + \gamma_i^{(t)}} \left(a_i - \gamma_i^{(t)} \right) \right\}, 0 \right\},$$

$$\delta_+^{(t)} = \max \left\{ \max_{i \in [d]} \left\{ C_1(\beta_i^{(t)} + \gamma_i^{(t)} - a_i) + \frac{C_2 \gamma_i^{(t)}}{\beta_i^{(t)} + \gamma_i^{(t)}} \left(\gamma_i^{(t)} - a_i \right) \right\}, 0 \right\}.$$

where C_1, C_2 denote two sufficiently large constants. Let $\Delta_+ = \max\{\Delta, 0\}$, and $\Delta_- = \max\{-\Delta, 0\}$. Consider the following functions:

$$\Phi_+^{(t)} = \max \left\{ \delta_+^{(t)}, \left(1 + \frac{1}{\text{poly}(\kappa)} \right) \Delta_-^{(t)} \right\},$$

$$\Phi_-^{(t)} = \max \left\{ \delta_-^{(t)}, \left(1 + \frac{1}{\text{poly}(\kappa)} \right) \Delta_+^{(t)} \right\}.$$

Let $\beta_+^{(t)} = \frac{1}{C} \max_{i \in [d]} \{\beta_i^{(t)}\}$. Let $\Phi^{(t)} = \max\{\Phi_+^{(t)}, \Phi_-^{(t)}, \beta_+^{(t)}\}$ be our potential function. Lemma 11 implies that by the end of $t = T_3$ iterations, we have $\delta_-^{(t)}, \delta_+^{(t)}, \beta_+^{(t)}, \Delta_+^{(t)}, \Delta_-^{(t)}$ are all less than $O\left(1/d \log^{0.01} d\right)$. Hence $\Phi^{(T_3)} \leq O(1/(d \log^{0.01} d))$. The result below shows that after iteration T_3, $\Phi^{(t)}$ further decreases whenever it is at least $O(\frac{\text{poly}(\kappa_2)c_t}{d^2})$.

Lemma 12 (Stage 2.2: the final substage) *In the setting of Theorem 4, suppose that Proposition 10 holds. For any $T_3 < t \le T_4$, as long as $\Phi^{(t)} \ge \frac{\text{poly}(\kappa_2)c_t}{d^2}$ (recalling that c_t is defined in Proposition 10) we have that*

$$\Phi^{(t+1)} \le \Phi^{(t)}\left(1 - \eta\frac{\min\{C_1, 1\}}{8}\Phi^{(t)}\right).$$

By combining the results from Stage 1 and 2, we are ready to prove Theorem 4.

Proof [Proof of Theorem 4] When Proposition 7 and 10 hold, we can obtain the following observation. Using the induction hypothesis in Eq (22), we have that for the infinite-width case, the population loss $L_\infty(\mathcal{P}^{(t)})$ satisfies:

$$L_\infty(\mathcal{P}^{(t)}) = O\left(\left(\sum_{i\in[d]}\left[\gamma_i^{(t)} + \beta_i^{(t)}\right] - \sum_{i\in[d]} a_i\right)^2\right) + O\left(\sum_{i\in[d]}\left[(a_i - \gamma_i^{(t)})^2 + \beta_i^2\right]\right) + \frac{1}{\text{poly}(d)},$$

$$(25)$$

where the first term comes from the zero-order term in the population loss and the second term comes from 2nd and higher order tensors in the population loss. This observation also implies that

$$L_\infty(\mathcal{P}^{(t)}) = O\left(d[\Phi^{(t)}]^2\right) + \frac{1}{\text{poly}(d)}. \tag{26}$$

At the beginning of Stage 2.2, by Lemma 11, we know that $\Phi^{(T_3)} \le 1/(d\log^{0.01} d)$. During Stage 2.2, by Lemma 12, as long as $\Phi^{(t)} \ge O_\kappa(c_t/d^2)$, $\Phi^{t+1} \le \Phi^{(t)} \le \Phi^{(t)}(1 - O(\Phi^{(t)}))$. Hence, after at most $d^{1+O(Q)}/\eta$ iterations (or $T_4 - T_3$ more precisely), $\Phi^{(T_4)}$ reduces to below $O(d^{1+Q})$. Applying this result to equation (26), we conclude that $L_\infty(\mathcal{P}^{(T_4)}) \le O(1/d^{1+Q})$. ∎

A.1. Stage 1.1: Proof of Convergence for 0th and 2nd Order Tensors

This section provides the proof of Lemma 8 is organized as follows.

- In Proposition 13, we first show that the gradients from 4+ order tensor decomposition terms are small compared to that of the 0th and 2nd order tensor terms.

- The above shows that the dynamic is mainly dominated by the 0th and 2nd tensor terms initially. In Proposition 14 and Proposition 15, we show the gradient update of the 0th and 2nd order. Based on these, we show the proof Lemma 8 at the end of this subsection.

Upper bound on the gradient of 4+ order terms. We first show that the 4th and higher order tensor gradients do not have much contribution to the gradient, for all the neurons in $\mathcal{S}$. We introduce the following notations for convenience. For a neuron distribution $\mathcal{P}$, let the following denote the gradient term of v involving only other neurons w.

$$\nabla_{2j,v,n} := (b_{2j} + b'_{2j})\left(\mathbb{E}_{w\sim\mathcal{P}}\langle w, v\rangle\langle\bar{w}, \bar{v}\rangle^{2j-2}w\right) - b'_{2j}\left(\mathbb{E}_{w\sim\mathcal{P}}\langle w, v\rangle\langle\bar{w}, \bar{v}\rangle^{2j-2}\langle w, \bar{v}\rangle\right)\bar{v}. \quad (27)$$

Recall that $\nabla_{2j,v}$ is the gradient of v for the 2j-th tensor (cf. equation (12)). Let

$$\nabla_{\geq 4,v} = \sum_{j \geq 2} \nabla_{2j,v}, \text{ and } \nabla_{\geq 4,v,n} = \sum_{j \geq 2} \nabla_{2j,v,n}.$$

The following result provides an upper bound on the higher order gradients.

Proposition 13 (Upper bound for 4th or higher order gradients) *In the setting of Lemma 8, suppose Proposition 7 holds. Then there exists an absolute constant $C > 0$ such that for every $i \in [d]$ and $v \in \mathcal{S}_g$, at the t-th iteration for $t \leq T_2$, the neuron v from distribution $\mathcal{P}^{(t)}$ satisfies that*

$$\left| [\nabla_{\geq 4,v}]_i \right| \leq \frac{C}{4} \left(\frac{c_t \kappa}{d^2} + \frac{\kappa}{d} \|\bar{v}^{(t)}\|_\infty^2 \right) |v_i^{(t)}|.$$

Moreover, the gradient from the network satisfies

$$\left| [\nabla_{\geq 4,v,n}]_i \right| \leq \frac{C}{4} \cdot \frac{c_t \kappa}{d^2} |v_i^{(t)}|.$$

As a corollary, for every $v \in \mathcal{S} \subseteq \mathcal{S}_g$, we have that

$$\left| [\nabla_{\geq 4,v}]_i \right| \leq \frac{C}{2} \cdot \frac{c_t \kappa}{d^2} |v_i^{(t)}|.$$

Proof Let us focus on $\nabla_{4,v}$ first. We now bound each terms in $[\nabla_{4,v}]_i$ in equation (12) separately.

$$[\nabla_{4,v}]_i = (b_4 + b_4') \left(\mathbb{E}_{w \sim \mathcal{P}} \left[\sum_{i_1,i_2} \left(\prod_{r=1} (\bar{w}_{i_r} \bar{v}_{i_r})^2 \right) (w_{i_j})^2 v_{i_j} \right] - a_i \langle e_i, v \rangle \langle e_i, \bar{v} \rangle^2 \right)$$
$$- b_4' \left(\mathbb{E}_{w \sim \mathcal{P}} \left[\|w\|_2^2 \sum_{i_1,i_2} \prod_{r=1}^2 (\bar{w}_{i_r} \bar{v}_{i_r})^2 \right] - \sum_{r \in [d]} a_r \langle e_r, \bar{v} \rangle^4 \right) v_i.$$

For the second two line of the above,

$$\left| \mathbb{E}_{\mathcal{P}} \left[w_i^2 \sum_{j=1}^d (\bar{w}_j \bar{v}_j)^2 \right] v_i \right| \leq \left| \mathbb{E}_{w \sim \mathcal{P}, w^{(0)} \in \mathcal{S}} \left[w_i^2 \sum_{j=1}^d (\bar{w}_j \bar{v}_j)^2 \right] v_i \right| + \left| \mathbb{E}_{w \sim \mathcal{P}, w^{(0)} \notin \mathcal{S}} \left[w_i^2 \sum_{j=1}^d (\bar{w}_j \bar{v}_j)^2 \right] v_i \right|$$

$$\leq |v_i| \left| \mathbb{E}_{w \sim \mathcal{P}, w^{(0)} \in \mathcal{S}} \left[w_i^2 \sum_{j=1}^d (\bar{w}_j \bar{v}_j)^2 \right] \right| + \Lambda |v_i|$$

$$\leq |v_i| \frac{c_t}{d} \left| \mathbb{E}_{w \sim \mathcal{P}, w^{(0)} \in \mathcal{S}} \left[w_i^2 \sum_{j=1}^d (\bar{v}_j)^2 \right] \right| + \Lambda |v_i|$$

$$\leq |v_i| \left(\frac{2 c_t \kappa}{d^2} + \frac{1}{\text{poly}(d)} \right) \tag{28}$$

where the second inequality uses inequality (19) so $\mathbb{E}_{w \sim \mathcal{P}^{(t)}, w^{(0)} \notin \mathcal{S}} \|w\|_2^2 \leq \Lambda$, and the second last inequality uses $\mathbb{E}_{\mathcal{P}}[w_i^2] \leq \frac{2\kappa}{d}$ as in Eq (17).

For the signal term in the gradient, $a_i\langle e_i, v\rangle\langle e_i, \bar{v}\rangle^2$, because $a_i \leq \kappa/d$, we have

$$\left|a_i\langle e_i, v\rangle\langle e_i, \bar{v}\rangle^2\right| \leq \frac{\kappa}{d}|v_i||\bar{v}_i|^2 \tag{29}$$

Another term in the gradient is (again, using the fact that for w with $w^{(0)} \in \mathcal{S}$, $\|w\|_\infty^2 \leq \frac{c_t}{d}$):

$$\left|\sum_{r,r'} \mathop{\mathbb{E}}_{\mathcal{P}} \frac{w_r^2 v_r^2 w_{r'}^2 v_{r'}^2}{\|w\|_2^2 \|v\|_2^4}\right| |v_i| \leq \left|\sum_{r,r'} \mathop{\mathbb{E}}_{w\sim\mathcal{P}, w^{(0)}\in\mathcal{S}} \frac{w_r^2 v_r^2 w_{r'}^2 v_{r'}^2}{\|w\|_2^2 \|v\|_2^4}\right| |v_i| + \left|\sum_{r,r'} \mathop{\mathbb{E}}_{w\sim\mathcal{P}, w^{(0)}\not\in\mathcal{S}} \frac{w_r^2 v_r^2 w_{r'}^2 v_{r'}^2}{\|w\|_2^2 \|v\|_2^4}\right| |v_i|$$

$$\leq \frac{c_t}{d}|v_i| \sum_{r,r'} \mathop{\mathbb{E}}_{\mathcal{P}} \frac{w_r^2}{\|w\|_2^2} \frac{v_r^2 v_{r'}^2}{\|v\|_2^4} + \Lambda \operatorname{poly}(d)|v_i|$$

$$\leq |v_i| \left(\frac{2\kappa c_t}{d^2} + \frac{1}{\operatorname{poly}(d)}\right) \tag{30}$$

The last term in the gradient is given by:

$$\left|\left(\sum_r a_r\langle e_r, v\rangle\langle e_r, \bar{v}\rangle^3\right)\bar{v}_i\right| \leq \frac{\kappa}{d}\left(\sum_r v_r \bar{v}_r^3\right)|\bar{v}_i|$$

$$\leq \frac{\kappa}{d}\frac{\|v\|_4^4}{\|v\|_2^4}|v_i| \leq \frac{\kappa}{d}\|\bar{v}\|_\infty^2 |v_i| \tag{31}$$

Combining Eq (28), Eq (29), Eq (30) and Eq (31), we obtain that

$$\left|[\nabla_{\geq 4, v}]_i\right| \leq O(1) \times \left(\frac{c_t \kappa}{d^2} + \frac{\kappa}{d}\|\bar{v}\|_\infty^2\right)|v_i|.$$

For $\Delta_{2j, v}$, with $j \geq 3$, we can apply the same calculation as above, and show that

$$\left|[\nabla_{2j, v}]_i\right| \leq O(b_{2j}) \times \left(\frac{c_t \kappa}{d^2} + \frac{\kappa}{d}\|\bar{v}\|_\infty^2\right)|v_i|.$$

Since $\nabla_{\geq 4, v} = \sum_{j\geq 2} \nabla_{2j, v}$ and $\sum_j b_{2j} = O(1)$ we complete the proof. $\blacksquare$

Based on the above result, we describe the dynamic of the 0th order tensor in the following proposition.

Proposition 14 (Learning 0th order tensor) *In the setting of Lemma 8, suppose Proposition 7 holds. Assume that for every $1 \leq i \leq d$, $\mathbb{E}_{w^{(t)}\sim\mathcal{P}^{(t)}}[w^{(t)}{}_i^2] \geq \frac{1}{\kappa_1 d}$. Then for every $t \leq T_2$, at least one of the following holds:*

1. $|\Delta^{(t)}| \leq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$.

2. If $\Delta^{(t)} > 0$, then $\Delta^{(t)} \leq \delta_-^{(t)}\left(1 - \frac{1}{16\kappa^2 d}\right)$. If $\Delta^{(t)} < 0$, then $|\Delta^{(t)}| \leq \left(1 - \frac{1}{5\kappa_1^5 d}\right)\delta_+^{(t)}$.

Moreover, when $\Delta^{(t)} \geq \max\left\{8\kappa^2\delta_+^{(t)}, \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}\right\}$, it holds that

$$\Delta^{(t+1)} \leq \Delta^{(t)} - \eta\frac{1}{4\kappa d}\Delta^{(t)}|\mathcal{S}_+^{(t)}|, \tag{32}$$

where $\mathcal{S}_+^{(t)}$ is the set of all $i \in [d]$ with $\delta_i^{(t)} \geq 0$ and $\left|\mathcal{S}_+^{(t)}\right|$ denotes its cardinality.

Proof Consider the iteration t, we have that for every i and every $v \in \mathcal{S}_g$, the update of $v_i^{(t)}$ is given as:

$$v_i^{(t+1)} = v_i^{(t)} - \eta(\Delta^{(t)} + \delta_i^{(t)})v_i^{(t)} - \eta[\nabla_{\geq 4, v}^{(t)}]_i.$$

Hence, using Proposition 13 and inequality (19), it holds that

$$\mathbb{E}_{\mathcal{P}^{(t+1)}}[w_i^2] = (1 - 2\eta(\Delta^{(t)} + \delta_i^{(t)})) \mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \pm \left(2\eta C \frac{c_t \kappa^2}{d^3} + \eta^2 \frac{1}{\lambda_0} + \eta \Lambda \frac{1}{\lambda_0}\right). \tag{33}$$

This implies that for every sufficiently small $\eta \leq \lambda_0^2$ and $\Lambda \leq \lambda_0^2$, it holds:

$$\delta_i^{(t+1)} = \delta_i^{(t)} - 2\eta \left(\Delta^{(t)} + \delta_i^{(t)}\right) \mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \pm 3\eta C \frac{c_t \kappa^2}{d^3} \tag{34}$$

Let us consider two cases when $\mathrm{Abs}\, \Delta^{(t)} = \Omega(\kappa_1^8 c_t / d^2)$.

Case 1. $\Delta^{(t)} = \Omega\left(\frac{\kappa_1^8 c_t}{d^2}\right)$, we have that for every i with $\delta_i^{(t)} \geq 0$, it holds that $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \geq a_i \geq \frac{1}{\kappa d}$. Hence,

$$\delta_i^{(t+1)} \leq \delta_i^{(t)} \left(1 - \eta \frac{1}{\kappa d}\right) + 3\eta C \frac{c_t \kappa^2}{d^3} - \eta \Delta^{(t)} \frac{1}{\kappa d}$$

$$\leq \delta_i^{(t)} \left(1 - \eta \frac{1}{\kappa d}\right) - \eta \frac{1}{2\kappa d} \Delta^{(t)}.$$

Summing up over all those i gives us

$$\Delta_+^{(t+1)} \leq \Delta_+^{(t)} \left(1 - \eta \frac{1}{\kappa d}\right) - \eta \frac{1}{2\kappa d} \Delta^{(t)} |\mathcal{S}_+^{(t)}|. \tag{35}$$

On the other hand, for every i with $\delta_i^{(t)} \leq 0$, it holds that $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \leq a_i \leq \frac{\kappa}{d}$. Hence,

$$|\delta_i^{(t+1)}| \geq |\delta_i^{(t)}| - \eta \left(|\delta_i^{(t)}| - \Delta^{(t)}\right) \mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] - 3\eta C \frac{c_t \kappa^2}{d^3}. \tag{36}$$

Now, consider a value $\rho = \frac{1}{4\kappa^2 d}$, when $\Delta^{(t)} \geq \delta_-^{(t)}(1 - \frac{\rho}{2})$, it holds that $(1 - \frac{\rho}{2})\Delta^{(t)} \geq \delta_-^{(t)}(1 - \rho)$. Therefore, when $\Delta^{(t)} \geq \delta_-^{(t)}(1 - \frac{\rho}{2})$, Eq (36) implies that

$$|\delta_i^{(t+1)}| \geq |\delta_i^{(t)}| - \eta \rho \left(|\delta_i^{(t)}|\right) \frac{\kappa}{d} - 3\eta C \frac{c_t \kappa^2}{d^3} + \eta \frac{\rho}{2} \Delta^{(t)} \mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2]. \tag{37}$$

Hence, using the assumption that $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \geq \frac{1}{\kappa_1 d}$, it holds that

$$|\delta_i^{(t+1)}| \geq |\delta_i^{(t)}| - \eta \rho \left(|\delta_i^{(t)}|\right) \frac{\kappa}{d} = |\delta_i^{(t)}| \left(1 - \eta \rho \frac{\kappa}{d}\right). \tag{38}$$

Summing up all $\delta_i^{(t)}$ with $\delta_i^{(t)} \leq 0$, this implies that

$$\Delta_-^{(t+1)} \geq \Delta_-^{(t)} \left(1 - \eta \frac{\rho \kappa}{d}\right).$$

Combine the above inequality with inequality (35), we have that (using $\Delta^{(t)} \geq 0$ so that $\Delta_+^{(t)} \geq \Delta_-^{(t)}$):

$$\Delta^{(t+1)} = \Delta_+^{(t+1)} - \Delta_-^{(t+1)} \leq \Delta_+^{(t)} - \Delta_-^{(t)} - \eta \frac{1}{\kappa d}\Delta_+^{(t)} + \eta \frac{\rho\kappa}{d}\Delta_-^{(t)} - \eta \frac{1}{2\kappa d}\Delta^{(t)}|\mathcal{S}_+^{(t)}|$$

$$\leq \Delta^{(t)} - \eta \frac{1}{2\kappa d}\Delta_+^{(t)} - \eta \frac{1}{2\kappa d}\Delta^{(t)}|\mathcal{S}_+^{(t)}|.$$

Therefore we conclude that when $\Delta^{(t)} \geq \Omega\left(\frac{c_t\kappa^8}{d^2}\right)$ and $\Delta^{(t)} \geq \delta_-^{(t)}(1 - \frac{\rho}{2})$, it must holds that

$$\Delta^{(t+1)} \leq \Delta^{(t)}\left(1 - \eta\frac{1}{2\kappa d}\right)$$

$$\delta_-^{(t+1)} \geq \delta_-^{(t+1)}\left(1 - \eta\frac{1}{4\kappa d}\right).$$

Here the second inequality comes from Eq (38). This implies that $\Delta^{(t+1)}$ will decrease faster than $\delta_-^{(t+1)}$ at the next iteration. Hence, when $\Delta^{(t)} \geq \Omega\left(\frac{c_t\kappa^8}{d^2}\right)$, then $\Delta^{(t)} \geq \delta_-^{(t)}(1 - \frac{\rho}{2})$ can never happen. Hence, by our choice of ρ, we conclude that as long as $\Delta^{(t)} \geq \Omega\left(\frac{c_t\kappa^8}{d^2}\right)$, then

$$\Delta^{(t)} \leq \delta_-^{(t)}\left(1 - \eta\frac{1}{16\kappa^2 d}\right).$$

On the other hand, even when $\Delta^{(t)} \leq \delta_-^{(t)}(1 - \frac{\rho}{2})$ but $\Delta^{(t)} = \Omega\left(\frac{\kappa_1^8 c_t}{d^2}\right)$, we still have that for every i with $\delta_i^{(t)} \leq 0$, by Eq (37):

$$|\delta_i^{(t+1)}| \geq |\delta_i^{(t)}| - \eta\left(|\delta_i^{(t)}| - \Delta^{(t)}\right)\underset{\mathcal{P}^{(t)}}{\mathbb{E}}[w_i^2] - 3\eta C\frac{c_t\kappa^2}{d^3}$$

$$\geq |\delta_i^{(t)}|\left(1 - \eta\frac{2\kappa}{d}\right).$$

Hence, as long as $\Delta^{(t)} = \Omega\left(\frac{\kappa_1^8 c_t}{d^2}\right)$, we will always have

$$\Delta_-^{(t+1)} \geq \Delta_-^{(t)}\left(1 - \eta\frac{2\kappa}{d}\right).$$

Combining the above with equation (35), we have that

$$\Delta^{(t+1)} = \Delta_+^{(t+1)} - \Delta_-^{(t+1)} \leq \Delta_+^{(t)} - \Delta_-^{(t)} - \eta\frac{1}{\kappa d}\Delta_+^{(t)} + \eta\frac{2\kappa}{d}\Delta_-^{(t)} - \eta\frac{1}{2\kappa d}\Delta^{(t)}|\mathcal{S}_+^{(t)}|$$

$$\leq \Delta^{(t)} + \eta\frac{2\kappa}{d}\delta_+^{(t)}|\mathcal{S}_+^{(t)}| - \eta\frac{1}{2\kappa d}\Delta^{(t)}|\mathcal{S}_+^{(t)}|.$$

Here, we are using the fact that $\Delta_-^{(t)} \leq \Delta_+^{(t)} \leq \delta_+^{(t)}|\mathcal{S}_+^{(t)}|$. Now, this implies that when $\Delta^{(t)} \geq \max\left\{8\kappa^2\delta_+^{(t)}, \Omega\left(\frac{\kappa_1^8 c_t}{d^2}\right)\right\}$, it also holds that

$$\Delta^{(t+1)} \leq \Delta^{(t)} - \eta\frac{1}{4\kappa d}\Delta^{(t)}|\mathcal{S}_+^{(t)}|.$$

Case 2. In the second case $\Delta^{(t)} \leq -\Omega\left(\frac{c_t \kappa_1^8}{d^2}\right)$, we shall use the similar proof, but with the assumption that $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \in \left[\frac{1}{\kappa_1 d}, \frac{2\kappa}{d}\right]$, it holds that when $|\Delta^{(t)}| \geq \delta_+^{(t)}(1 - \rho)$, for $\rho = \frac{1}{10\kappa_1^5 d}$. Therefore, we can also conclude:

$$-\Delta^{(t+1)} \leq -\Delta^{(t)} - \eta \frac{1}{5\kappa_1^5 d} \Delta_-^{(t)}.$$

The proof follows by a similar argument to Case 1. ∎

Based on the above result, next we describe the dynamic of the 2nd order tensor.

Proposition 15 (Learning 2nd order tensor) *In the setting of Lemma 8, suppose Proposition 7 holds. Assume that for every $1 \leq i \leq d$, $\mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}}[{w^{(t)}}_i^2] \geq \frac{1}{\kappa_1 d}$. For every $t \leq T_2$, we have that*

$$\delta_+^{(t+1)} \leq \delta_+^{(t)}\left(1 - \eta \frac{1}{\operatorname{poly}(\kappa_1)d}\right) + \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^3},$$

$$\delta_-^{(t+1)} \leq \delta_-^{(t)}\left(1 - \eta \frac{1}{\operatorname{poly}(\kappa_1)d}\right) + \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^3}.$$

Moreover, when $\Delta^{(t)} > 0$, we have the following improved bound for $\delta_+^{(t+1)}$:

$$\delta_+^{(t+1)} \leq \delta_+^{(t)}\left(1 - \eta \frac{1}{4\kappa d}\right) + \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^3}.$$

Proof By the update rule, we can obtain (in Eq (34)) that

$$\delta_i^{(t+1)} = \delta_i^{(t)} - 2\eta\left(\Delta^{(t)} + \delta_i^{(t)}\right)\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \pm 3\eta C \frac{c_t \kappa^2}{d^3}.$$

Let us first consider the case when $\Delta^{(t)} > 0$, then, for $\delta_i^{(t)} \geq 0$, it holds that $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \geq a_i \geq \frac{1}{\kappa}$, thus,

$$\delta_+^{(t+1)} \leq \delta_+^{(t)}\left(1 - \eta \frac{1}{\kappa d}\right) + 3\eta C \frac{c_t \kappa^2}{d^3}.$$

Now, for $\delta_-^{(t+1)}$, consider two cases: $\delta_-^{(t)} \geq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$ or $\delta_-^{(t)} \leq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$. In the first case, we have that when $|\Delta^{(t)}| > \left(1 - \frac{1}{5\kappa_1^5 d}\right)\delta_-^{(t)}$, it must be that $|\Delta^{(t)}| \geq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$. By Proposition 14, this can not happen. Therefore, we must have $|\Delta^{(t)}| \leq \left(1 - \frac{1}{5\kappa_1^5 d}\right)\delta_-^{(t)}$. Hence, using $\mathbb{E}_{\mathcal{P}}[w_i^2] \geq \frac{1}{\kappa_1 d}$, we have that

$$\delta_-^{(t+1)} \leq \delta_-^{(t)} - \eta\left(\delta_-^{(t)} - |\Delta^{(t)}|\right)\frac{1}{\kappa_1 d} + 3\eta C \frac{c_t \kappa^2}{d^3}$$

$$\leq \delta_-^{(t)} - \eta \delta_-^{(t)} \frac{1}{5\kappa_1^9 d} + 3\eta C \frac{c_t \kappa^2}{d^3},$$

which proves the condition. On the other hand when $\delta_-^{(t)} \leq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$, we directly completes the proof by choosing a larger poly in $\frac{c_t \operatorname{poly}(\kappa_1)}{d^3}$. We can apply the same argument for δ_+, and the improved bound for the case when $\Delta^{(t)} > 0$. ∎

A.1.1. PROOF OF THE MAIN LEMMA

Now we are ready to show the final convergence lemma. We first provide the following claim that shows on average, each coordinate of the neuron distribution lies in a bounded range. This also proves the first equation of (18) in the inductive hypothesis $\mathcal{H}_1$.

Claim A.1 *In the setting of Lemma 8, for every $t \leq T_2' = \Theta(\frac{d^2}{\eta c_{T_2} \exp(\text{poly}(\kappa))})$ and every $i \in [d]$, we have that $\mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}}[w^{(t)2}_i] \in \left[\frac{1}{\kappa_1 d}, \frac{2\kappa}{d}\right]$.*

Proof The upper bound follows from $\delta_+^{(t)} \leq \delta_+^{(0)} \leq \frac{\kappa}{d}$, so $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \leq a_i + \delta_+^{(t)} \leq \frac{2\kappa}{d}$. For the lower bound, we will prove it by induction. Let us assume that the lower bounds hold for $t \leq T_0$ for some $T_0 < T_2$. For $t = T_0 + 1$, denote $\mathcal{T}$ the iterations $t \leq T_0 + 1$ where $\Delta^{(t)} > 0$, and $\mathcal{T}^c$ be the other iterations. To show the lower bound, for $i \in [d]$, we know that when $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \leq \frac{1}{2\kappa d}$, by the update rule in Eq (33), we can conclude that:

$$t \in \mathcal{T}^c \implies \mathbb{E}_{\mathcal{P}^{(t+1)}}[w_i^2] \geq \mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2], \tag{39}$$

$$t \in \mathcal{T} \implies \mathbb{E}_{\mathcal{P}^{(t+1)}}[w_i^2] \geq \mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2]\left(1 - \eta\Delta^{(t)}\right). \tag{40}$$

Hence, we only need to consider $t \in \mathcal{T}$, for these iterations, by Proposition 15 we know that

$$\delta_+^{(t+1)} \leq \delta_+^{(t)}\left(1 - \eta\frac{1}{4\kappa d}\right) + \eta\frac{c_t \text{poly}(\kappa_1)}{d^3}.$$

On the other hand by Proposition 14, we have that when $\Delta^{(t)} \geq \max\left\{8\kappa^2\delta_+^{(t)}, \frac{c_t \text{poly}(\kappa_1)}{d^2}\right\}$, it holds that:

$$\Delta^{(t+1)} \leq \Delta^{(t)} - \eta\frac{1}{4\kappa d}\Delta^{(t)}|\mathcal{S}_+^{(t)}| \leq \Delta^{(t)} - \eta\frac{1}{4\kappa d}\Delta^{(t)}.$$

Now, let us define $\gamma_0 = \delta_+^0 \leq \frac{2\kappa}{d}$, with $\gamma_{t+1} = \gamma_t\left(1 - \eta\frac{1}{4\kappa d}\right) + \eta\frac{c_t \text{poly}(\kappa_1)}{d^3}$ for every $t \in \mathcal{T}$ and $\gamma_{t+1} = \gamma_t + \eta\frac{c_t \text{poly}(\kappa_1)}{d^3}$ otherwise. We know that as long as $\Delta^{(t)} \geq \frac{c_t \text{poly}(\kappa_1)}{d^2}$, we have:

$$\Delta^{(t)} \leq 8\kappa^2\gamma_t.$$

This implies that

$$\eta\sum_{t\in\mathcal{T}}\Delta^{(t)} \leq \eta\sum_{t\in\mathcal{T}}\left(8\kappa^2\gamma_t + \frac{c_t \text{poly}(\kappa_1)}{d^2}\right) \leq \eta\frac{c_{T_2}\text{poly}(\kappa_1)}{d^2} \times T_2 + \eta 8\kappa^2\sum_{t\in\mathcal{T}}\gamma_t$$

$$\leq 1 + 8\kappa^2(4\kappa d)\left(\gamma_0 + \eta\frac{c_{T_2}\text{poly}(\kappa_1)}{d^3} \times T_2\right) \leq 2 + 64\kappa^4 \tag{41}$$

Using equation (40), we have that

$$\mathbb{E}_{\mathcal{P}^{(T_0+1)}}[w_i^2] \geq \mathbb{E}_{\mathcal{P}^{(0)}}[w_i^2]\exp\left\{-\eta\sum_{s\in\mathcal{T}}\Delta^{(s)} - \eta^2\text{poly}(1/\lambda_0)T_2\right\}.$$

29

Together with Eq (41) we conclude that $\mathbb{E}_{\mathcal{P}^{(T_0+1)}}[w_i^2] \geq \frac{1}{\kappa_1 d}$. By induction we complete the proof. ∎

Based on Claim A.1, we prove Lemma 8.

Proof [Proof of Lemma 8] Clearly, by Proposition 15 once $\delta_+^{(t)}$ or $\delta_-^{(t)} \leq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$, they will stay within the interval for the next iterations. By Proposition 14, after both $\delta_+^{(t)}$ and $\delta_-^{(t)} \leq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$, we know $\Delta^{(t)}$ will be within the interval as well.

Hence, we just need to consider the first time that $\delta_+^{(t)}$ and $\delta_-^{(t)}$ goes outside the interval. Following Proposition 15, we know that when $\delta_+^{(t)} \geq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$, it holds:

$$\delta_+^{(t+1)} \leq \delta_+^{(t)} \left(1 - \eta \frac{1}{\operatorname{poly}(\kappa_1)d}\right),$$

which gives the convergence error rate of δ_+ after T_1 iterations. The same holds for $\delta_-^{(t)}$. ∎

Finally, we have an estimate of how big each coordinate is for the neurons at the end of Stage 1.1, which can be given by the output layer weights $\{a_i\}_{i=1}^d$. We show the following claim, which will be used in the proof of Stage 2.

Claim A.2 (The end of Stage 1.1) *In the setting of Lemma 8, at iteration T_1 (recalling that $T_1 = \Theta(\frac{\operatorname{poly}(\kappa_1)d\log d}{\eta})$), for every $v \in \mathcal{S}_g$ and every $i \in [d]$, we have that*

$$v_i^{(T_1)} = (a_i d)v_i^{(0)} \pm \frac{\operatorname{poly}(\log d)}{d^{3/2}}.$$

Proof Let us first show the upper bound. For every $v^{(0)} \in \mathcal{S}_g$. By the update rule, we have that

$$v_i^{(t+1)} = v_i^{(t)} - \eta(\Delta^{(t)} + \delta_i^{(t)})v_i^{(t)} - \eta[\nabla_{\geq 4,v}^{(t)}]_i.$$

Using Proposition 13, we have that:

$$\left(v_i^{(t+1)}\right)^2 \leq \left(v_i^{(t)}\right)^2 - 2\eta(\Delta^{(t)} + \delta_i^{(t)})\left(v_i^{(t)}\right)^2 + \eta[\nabla_{\geq 4,v}^{(t)}]_i v_i^{(t)} + \eta^2 O\left(\frac{1}{\lambda_0^2}\right)$$

$$\leq \left(v_i^{(t)}\right)^2 \left(1 - 2\eta(\Delta^{(t)} + \delta_i^{(t)})\right) + \eta\frac{C}{4}\left(\frac{c_t\kappa}{d^2} + \frac{\kappa}{d}\|v^{(t)}\|_\infty\|\bar{v}^{(t)}\|_\infty\right)|v_i^{(t)}|^2 + \eta^2 O\left(\frac{1}{\lambda_0^2}\right).$$

On the other hand, we have that by Eq (33), it holds:

$$\mathbb{E}_{\mathcal{P}^{(t+1)}}[w_i^2] \geq (1 - 2\eta(\Delta^{(t)} + \delta_i^{(t)}))\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] - \left(3\eta C\frac{c_t\kappa^2}{d^3}\right)$$

Using Claim A.1, we have that $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \geq \frac{1}{\kappa_1 d}$, which implies

$$\frac{(v_i^{(t+1)})^2}{\mathbb{E}_{\mathcal{P}^{(t+1)}}[w_i^2]} \leq \frac{(v_i^{(t)})^2}{\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2]} + \eta\left(10C\frac{c_t\kappa^2\kappa_1}{d^2} + C\left(\frac{c_t\kappa\kappa_1}{d} + \kappa\kappa_1\|v^{(t)}\|_\infty\|\bar{v}^{(t)}\|_\infty\right)|v_i^{(t)}|^2\right).$$

Similarly, we also have that

$$\frac{(v_i^{(t+1)})^2}{\mathbb{E}_{\mathcal{P}^{(t+1)}}[w_i^2]} \geq \frac{(v_i^{(t)})^2}{\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2]} - \eta \left(10C\frac{c_t\kappa^2\kappa_1}{d^2} + C\left(\frac{c_t\kappa\kappa_1}{d} + \kappa\kappa_1\|v^{(t)}\|_\infty\|\bar{v}^{(t)}\|_\infty\right)|v_i^{(t)}|^2 \right).$$

Using these two inequalities, with the conclude $\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2] \in \left[\frac{1}{\kappa_1 d}, \frac{2\kappa}{d}\right]$ in Claim A.1 and the assumption about initialization of $\mathcal{S}_g$, which says that for every $v \in \mathcal{S}_g$, $\|\bar{v}^{(0)}\|_\infty, \|v^{(0)}\|_\infty \leq \frac{\text{poly}(\log d)}{d}$, we can conclude that for every $v^{(0)} \in \mathcal{S}_g$, we have for every $t \in [T_1]$,

$$\frac{(v_i^{(t)})^2}{\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2]} = \frac{(v_i^{(0)})^2}{\mathbb{E}_{\mathcal{P}_0}[w_i^2]} \pm \frac{\text{poly}(\log d)}{d}. \tag{42}$$

Note that $v_i^{(t)}$ do not change sign during the gradient process (otherwise $v_i^{(t)}$ will be close to zero, violating the above inequality). Since $\mathbb{E}_{\mathcal{P}_0}[w_i^2] = \frac{1}{d}$ and by Lemma 8, $\left|\mathbb{E}_{\mathcal{P}^{(T_1)}}[w_i^2] - a_i\right| \leq \frac{c_t\,\text{poly}(\kappa_1)}{d^2}$. This implies that

$$v_i^{(T_1)} = (a_i d)v_i^{(0)} \pm \frac{\text{poly}(\log d)}{d^{3/2}},$$

which completes the proof. ∎

A.2. Stage 1.2: Proof of Convergence for Higher Order Tensors

In this section, we prove Lemma 9, which shows that by the end of Stage 1, a small fraction of neurons have won the lottery ticket by growing much larger than a typical neuron. This stage runs for approximately $\Theta_\kappa(\frac{d^2}{\eta\log d})$ many iterations (or $T_2 - T_1$ more precisely). The proof of Lemma 9 is organized as follows.

- First, in Proposition 16, we show that the dynamic is mainly determined by the 4th order gradients by bounding the gradients contributed by the 0th and 2nd order terms so that, as described in Section 3.1. Based on this result, we can relate the dynamic of this substage to tensor power method.

- Second, we provide a lower bound on the norm of every neuron in Claim A.3. Based on this result, we prove Claim A.4 that shows the growth of good neurons. This leads to the proof of Lemma 9 in Section A.2.1.

- Finally, we prove the inductive hypothesis $\mathcal{H}_1$ in Section A.2.2.

We describe the following proposition to bound the gradients of 4th or higher tensors.

Proposition 16 (Gradient bound for Stage 1.2) *In the setting of Lemma 9, suppose that Proposition 7 holds. Consider any iteration $t \in [T_1 + 1, T_2]$ and any neuron $v \in \mathcal{S}_g$. Suppose that for every $s \leq t$, $\|v^{(s)}\|_\infty \leq \frac{\text{poly}(\log d)}{\sqrt{d}}$. Then for every $i \in [d]$, the gradient of v at iteration t satisfies*

$$-[\nabla_v]_i = B_{1,4}a_i\langle e_i, v^{(t)}\rangle\langle e_i, \bar{v}^{(t)}\rangle^2 \pm \frac{c_t\,\text{poly}(\kappa_1)}{d^2}|v_i^{(t)}|.$$

Proof The result mainly follows from combining Proposition 13 for the gradient coming from 4+ order terms with Proposition 8 for the gradient of 0, 2 order terms.

The only remaining term is

$$\left| \left(\sum_r a_r \langle e_r, v^{(t)} \rangle \langle e_r, \bar{v}^{(t)} \rangle^3 \right) \bar{v}_i^{(t)} \right| \leq \frac{\kappa}{d} \left(\sum_r v_r^{(t)} (\bar{v}_r^{(t)})^3 \right) |\bar{v}_i^{(t)}| \leq \frac{\kappa}{d} \frac{\|v^{(t)}\|_4^4}{\|v^{(t)}\|_2^4} |v_i^{(t)}|.$$

Hence, we have

$$-[\nabla_{v^{(t)}}]_i = B_{1,4} a_i \langle e_i, v^{(t)} \rangle \langle e_i, \bar{v}^{(t)} \rangle^2 \pm \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} |v_i^{(t)}| \pm \frac{\kappa}{d} \frac{\|v^{(t)}\|_4^4}{\|v^{(t)}\|_2^4} |v_i^{(t)}|. \tag{43}$$

By the definition of $\mathcal{S}_g$, we know that at T_1 every $v \in \mathcal{S}_g$ satisfies

$$\|v^{(T_1)}\|_2^2 \geq \frac{1}{2\kappa}, \text{ and for at most } O(\log d) \text{ many } i \in [d], \ |v_i^{(T_1)}|^2 \geq \frac{\kappa \log d}{d}.$$

We will maintain the following condition by induction.

$$\|v^{(t)}\|_2^2 \geq \frac{1}{4\kappa}, \text{ and for at most } O(\log d) \text{ many } i \in [d], \ |v_i^{(t)}|^2 \geq \frac{c_t \operatorname{poly}(\kappa_1)}{d}. \tag{44}$$

Now suppose the following is true at some iteration $t \geq T_1$, then we have that

$$\frac{\|v^{(t)}\|_4^4}{\|v^{(t)}\|_2^4} \leq \frac{c_t \operatorname{poly}(\kappa_1)}{d}.$$

Thus, for iteration $t+1$, using Eq (43) we know that

$$(v_i^{(t+1)})^2 = (v_i^{(t)})^2 + \eta B_{1,4} a_i |v_i^{(t)}|^2 \langle e_i, \bar{v}^{(t)} \rangle^2 \pm \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} |v_i^{(t)}|^2 \pm \eta^2 O\left(\frac{1}{\lambda_0^2} \right).$$

Hence, we have that for every i with $|v_i^{(t)}|^2 \leq \frac{\operatorname{poly}(\kappa_1) c_t}{d}$, it holds that

$$(v_i^{(t+1)})^2 = (v_i^{(t)})^2 \left(1 \pm \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} \right). \tag{45}$$

Hence for every $t \leq T_2$, as long as $|v_i^{(T_1)}|^2 \leq \frac{\operatorname{poly}(\kappa_1) c_t}{d}$, we have:

$$(v_i^{(t+1)})^2 \in \left[\frac{2}{3} (v_i^{(T_1)})^2, \frac{3}{2} (v_i^{(T_1)})^2 \right].$$

This proves inequality (44) for $t+1$. ∎

Next, we use the following claim to maintain a lower bound on the norm of each neuron.

Claim A.3 (Norm lower bound for Stage 1.2) *In the setting of Lemma 9, suppose Proposition 7 holds. For every $v \in \mathcal{S}_g$, the norm of v at any iteration $t \in [T_1 + 1, T_2]$ satisfies $\|v^{(t)}\|_2^2 \geq \Omega\left(\frac{1}{\kappa} \right)$.*

Proof By the update rule, using Proposition 13 we know that for every $p \in [d]$:

$$
\begin{aligned}
-[\nabla_{v^{(t)}}]_p &= \sum_{j \geq 2} \left(B_{1,2j} a_p \langle e_p, v^{(t)} \rangle \langle e_p, \bar{v}^{(t)} \rangle^{2j-2} - B_{2,2j} \left(\sum_r a_r \langle e_r, v^{(t)} \rangle \langle e_r, \bar{v}^{(t)} \rangle^{2j-1} \right) \bar{v}_p^{(t)} \right) \\
&\quad \pm \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} v_p^{(t)} \\
&= \sum_{j \geq 2} \left(B_{1,2j} a_p \frac{(v_p^{(t)})^{2j-1}}{\|v^{(t)}\|_2^{2j-2}} - B_{2,2j} \frac{\sum_r a_r (v_r^{(t)})^{2j}}{\|v^{(t)}\|_2^{2j}} v_p^{(t)} \right) \pm \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} v_p^{(t)} \\
&= v_p^{(t)} Q_p^{(t)} \pm \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} v_p^{(t)},
\end{aligned}
\tag{46}
$$

where

$$
\begin{aligned}
Q_p^{(t)} &:= \sum_{j \geq 2} \left(B_{1,2j} a_p \frac{(v_p^{(t)})^{2j-2}}{\|v^{(t)}\|_2^{2j-2}} - B_{2,2j} \frac{\sum_r a_r (v_r^{(t)})^{2j}}{\|v^{(t)}\|_2^{2j}} \right) \\
&= \sum_{j \geq 2} \frac{1}{\|v^{(t)}\|_2^{2j-2}} \left(B_{1,2j} a_p (v_p^{(t)})^{2j-2} - B_{2,2j} \frac{\sum_r a_r (v_r^{(t)})^{2j}}{\|v^{(t)}\|_2^2} \right).
\end{aligned}
\tag{47}
$$

We have that

$$
\begin{aligned}
\sum_p \left(v_p^{(t)} \right)^2 Q_p^{(t)} &= \sum_p \sum_{j \geq 2} \frac{1}{\|v^{(t)}\|_2^{2j-2}} \left(B_{1,2j} a_p (v_p^{(t)})^{2j-2} - B_{2,2j} \frac{\sum_r a_r (v_r^{(t)})^{2j}}{\|v^{(t)}\|_2^2} \right) \left(v_p^{(t)} \right)^2 \\
&= \sum_p \sum_{j \geq 2} \frac{1}{\|v^{(t)}\|_2^{2j-2}} \left(B_{1,2j} a_p (v_p^{(t)})^{2j} - B_{2,2j} \left(\sum_r a_r (v_r^{(t)})^{2j} \right) \right) \\
&= \sum_{j \geq 2} \frac{1}{\|v^{(t)}\|_2^{2j-2}} (B_{1,2j} - B_{2,2j}) \left(\sum_r a_r (v_r^{(t)})^{2j} \right) \geq 0
\end{aligned}
\tag{48}
$$

This implies the following

$$
\|v^{(t+1)}\|_2^2 \geq \|v^{(t)}\|_2^2 \left(1 - \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} \right).
\tag{49}
$$

Combined with Proposition 16 , we have that for every neuron v, $\|v^{(t)}\|_2^2 = \Omega\left(\frac{1}{\kappa}\right)$ for every $t \in [T_1, T_2]$. ∎

A.2.1. PROOF OF THE MAIN LEMMA

Provided with the gradient bound and norm lower bound, we are now ready to prove the main result of Stage 1.2. Towards showing Lemma 9, we prove the following claim, which shows that if a neuron has grown beyond $\frac{\operatorname{poly}\log(d)}{d}$ at a certain iteration T_2', then this neuron will become basis-like at iteartion T_2.

Claim A.4 (Growth of good neurons) *In the setting of Lemma 9, suppose that Proposition 7 holds. For every $v \in \mathcal{S}_g$, suppose at iteration T_2' (recalling that $T_2' = T_2 - \frac{d^2}{\eta \operatorname{poly} \log(d)}$), only one coordinate $i \in [d]$ satisfies $|v_i^{(T_2')}| \geq \frac{\log^{10} d}{\sqrt{d}}$ and all the other coordinates satisfies $|v_j^{(T_2')}| \leq \frac{(\log d)^2}{\sqrt{d}}$, then at iteration T_2, we have that*

$$|v_i^{(T_2)}|^2 = \Omega\left(\frac{1}{\lambda_0 \operatorname{poly}(d)}\right) \geq \operatorname{poly}(d), \text{ and for any other } j \neq i, |v_j^{(T_2)}| \leq \frac{2(\log d)^2}{\sqrt{d}}.$$

In other words, the claim says that for neuron v, its i-th coordinate at iteration T_2, denoted by $|v_i^{(T_2)}|$, will be as large as $\operatorname{poly}(d)$, which implies that this neuron has won the lottery. We describe the proof of Claim A.4.

Proof We shall prove the claim by doing an induction. Consider the condition $|v_i^{(t)}| \geq \frac{\operatorname{poly}(\log d)}{\sqrt{d}}$ and all the other coordinates satisfies $|v_j^{(t)}| \leq \frac{2(\log d)^2}{\sqrt{d}}$ for $t \in [T_2', T]$. Suppose it is true up to iteration t, consider iteration $t+1$. When $p = i$, we have that $a_i(v_i^{(t)})^{2j-2} \geq a_r(v_r^{(t)})^{2j-2}$ for every $r \neq i$, hence this implies that (using the fact that $B_{1,2j} > B_{2,2j}$ and $B_{1,4}$ is greater than $B_{2,4}$ plus a fixed constant):

$$Q_i^{(t)} = \Omega\left(\frac{a_i(v_i^{(t)})^2}{\|v^{(t)}\|_2^2}\right) = \Omega\left(\frac{\operatorname{poly}(\log d)}{d^2} + \frac{a_i(v_i^{(t)})^2}{\|v^{(t)}\|_2^2}\right),$$

where $Q_i^{(t)}$ is defined in the proof of Claim A.3. With equation (48), this implies that

$$(v_i^{(t+1)})^2 \geq (v_i^{(t)})^2 \left(1 + \eta\Omega\left(\frac{\operatorname{poly}(\log d)}{d^2} + \frac{a_i(v_i^{(t)})^2}{\|v^{(t)}\|_2^2}\right)\right), \tag{50}$$

which provides a direct the lower bound on $(v_i^{(t+1)})^2$. Now, to show the upper bound of the other coordinates, recall that we have shown $\|v^{(t)}\|_2^2 = \Omega\left(\frac{1}{\kappa}\right)$ for every $t \in [T_1, T_2]$,

$$Q_p^{(t)} \leq \sum_{j \geq 2} \frac{1}{\|v^{(t)}\|_2^{2j-2}} \left(B_{1,2j} a_p(v_p^{(t)})^{2j-2}\right) = O\left(\kappa^2 \frac{\log^8 d}{d^2}\right).$$

This implies that

$$(v_p^{(t+1)})^2 \leq (v_p^{(t)})^2 \left(1 + \eta C^2 O\left(\frac{\log^8 d}{d^2}\right)\right).$$

Hence we prove all the other $p \neq i$ satisfies $|v_p^{(t)}| \leq \frac{2(\log d)^2}{\sqrt{d}}$ as long as $T_2 - T_2' \leq \frac{d^2}{\eta \log^9(d)}$, which complete the induction. In the end, since $|v_i^{(t)}| \geq \frac{\operatorname{poly}(\log d)}{\sqrt{d}}$ and all the other coordinates satisfies $|v_j^{(t)}| \leq \frac{2(\log d)^2}{\sqrt{d}}$ for every $t \in [T_2', T]$, we can further simplify Eq (51) as:

$$(v_i^{(t+1)})^2 \geq (v_i^{(t)})^2 \left(1 + \frac{\kappa(v_i^{(t)})^2}{d\log^5 d}\right), \tag{51}$$

34

which directly gives us the bound $|v_i^{(T_2)}|^2 = \Omega\left(\frac{1}{\lambda_0 \, \text{poly}(d)}\right)$ at iteration T_2. ∎

Now we are ready to prove Lemma 9. We define the union of good neurons as

$$\mathcal{S}_{good} = \left\{ v \in \mathcal{S}_g \mid \exists i \in [d], [v^{(0)}]_i^2 \geq \Gamma_i + \rho \text{ and all other } j : [v^{(0)}]_j^2 < \Gamma_j - \rho \right\},$$

where we recall that Γ_i and ρ have been defined before the statement of Lemma 9. In the proof, we focus on the dynamic of a neuron v until the point that $\|v\|_\infty \geq \frac{\text{poly}(\log d)}{\sqrt{d}}$. The key step is to track the dynamic via a tensor gradient update.

Proof [Proof of Lemma 9] We focus on proving the following three statements.

1. For every $v \notin \mathcal{S}_{pot}$, $\|v^{(t)}\|_\infty \geq \frac{\text{poly}(\log d)}{\sqrt{d}}$ never happen for any $t \leq T_2$.

2. For every $v \in \mathcal{S}_{good}$, $\|v^{(t)}\|_\infty \geq \frac{\text{poly}(\log d)}{\sqrt{d}}$ must happen for some $t \leq T_2'$ and when it happens, the condition in Claim A.4 meets for $i = \arg\max_{j \in [d]}\{v_j^{(0)}\}$.

3. For every $v \in \mathcal{S}_{pot} \backslash \mathcal{S}_{bad}$, $\|v^{(t)}\|_\infty \geq \frac{\text{poly}(\log d)}{\sqrt{d}}$ might happen for some $t \leq T_2'$. If $\|v^{(t)}\|_\infty \geq \frac{\text{poly}(\log d)}{\sqrt{d}}$ happens for some $t \leq T_2$, then the condition in Claim A.4 meets for the coordinate $i = \arg\max_{j \in [d]}\{v_j^{(0)}\}$.

The first and second statement of Lemma 9 follow by combining the above three statements and Claim A.4. The third statement can be proved by standard anti-concentration inequalities for the Gaussian distribution. For the rest of the proof, we focus on proving the above three statements. We know by Proposition 16 that when $\|v^{(t)}\|_\infty \leq \frac{\text{poly}(\log d)}{\sqrt{d}}$ the update of $v^{(t)}$ at every iteration $t \in [T_1, T_2]$ is given by

$$-[\nabla_{v^{(t)}}]_i = B_{1,4} a_i \langle e_i, v^{(t)} \rangle \langle e_i, \bar{v}^{(t)} \rangle^2 \pm \frac{c_t \, \text{poly}(\kappa_1)}{d^2} |v_i^{(t)}|$$

$$= B_{1,4} a_i \frac{(v_i^{(t)})^3}{\|v^{(t)}\|_2^2} \pm \frac{c_t \, \text{poly}(\kappa_1)}{d^2} |v_i^{(t)}|.$$

For every i, consider a process where $p^{(T_1)}, q^{(T_1)} = v_i^{(T_1)}$, with

$$p^{(t+1)} = p^{(t)} + \eta p^{(t)} \left(B_{1,4} a_i (p^{(t)})^2 + \frac{c_t \, \text{poly}(\kappa_1)}{d^2} \right),$$

$$q^{(t+1)} = q^{(t)} + \eta q^{(t)} \left(B_{1,4} a_i (q^{(t)})^2 - \frac{c_t \, \text{poly}(\kappa_1)}{d^2} \right).$$

Along with Eq (53), we can see that for every t where $\|v^{(t)}\|_\infty \leq \frac{\text{poly}(\log d)}{\sqrt{d}}$,

$$|p^{(t)}| \geq |v_i^{(t)}| \times \max\left\{ 1, \frac{1}{\|v^{(t)}\|_2} \right\},$$

$$|q^{(t)}| \leq |v_i^{(t)}| \times \min\left\{ 1, \frac{1}{\|v^{(t)}\|_2} \right\}.$$

To analyze this process, we introduce the following differential equation

$$\frac{dx(t)}{dt} = \tau_1 x^3, \quad x(0)^2 = \tau_2.$$

The solution is given as $x^2(t) = \frac{1}{\frac{1}{\tau_2} - 2\tau_1 t}$. Therefore, we can easily obtain that as long as $\rho = \Omega\left(\frac{c_t \operatorname{poly}(\kappa_1)}{d^2}\right)$, when $\tau_1 = B_{1,4}a_i$, $\eta 2\tau_1 T_2' = \frac{1}{\tau_2}$ which implies that $\tau_2 = \frac{1}{\eta 2\tau_1 T_2'} = \frac{1}{\eta 2(b_4+b_4')a_i T_2'}$, we have that

$$|v_i^{(T_1)}|^2 \geq \tau_2 + \rho \implies |q^{(T_2')}| = +\infty.$$

On the other hand,

$$|v_i^{(T_1)}|^2 \leq \tau_2 - \rho \implies |p^{(T_2')}|^2 = O\left(\frac{\tau_2^2}{\rho}\right) = O\left(\frac{\log^3 d}{d}\right).$$

In the end, by Proposition A.1 and the definition of $\mathcal{S}_g$ (Eq (16)), we know that for every $v \in \mathcal{S}_g$ and every $i \in [d]$, we have that

$$v_i^{(T_1)} = (a_i d)v_i^{(0)} \pm \frac{\operatorname{poly}(\log d)}{d^{3/2}}.$$

Putting into the definition of τ_2 we complete the proof. ∎

In addition, we state the following claim that will be used in Appendix B for the error analysis.

Claim A.5 (Upper bound on gradient norm at the end of Stage 1) *In the setting of Theorem 4, at the first iteration t where $\|v^{(t)}\|_2^2 > \frac{1}{\lambda_0}$, i.e. the threshold where gradients are truncated, we have that*

$$\sum_{s=1}^{t-1} \|\bar{v}^{(s)}\|_\infty^2 \leq O\left(\frac{\operatorname{poly}(\kappa_1)d \log \frac{1}{\lambda_0}}{\eta}\right).$$

Proof When $\|\bar{v}^{(t)}\|_\infty^2 \leq \frac{1}{\operatorname{poly}(\kappa_1)}$, we have that for $p = \arg\max_{r \in [d]}\{a_r(v_r^{(t)})^2\}$, the following holds

$$j = 2: \frac{1}{\|v^{(t)}\|_2^{2j-2}}\left(B_{1,2j}a_p(v_p^{(t)})^{2j-2} - B_{2,2j}\frac{\sum_r a_r(v_r^{(t)})^{2j}}{\|v^{(t)}\|_2^2}\right)$$

$$= \Omega\left(\frac{a_p(v_p^{(t)})^2}{\|v^{(t)}\|_2^2}\right) = \Omega\left(\frac{1}{\operatorname{poly}(\kappa_1)d}\|\bar{v}^{(t)}\|_\infty^2\right).$$

$$j \geq 2: \frac{1}{\|v^{(t)}\|_2^{2j-2}}\left(B_{1,2j}a_p(v_p^{(t)})^{2j-2} - B_{2,2j}\frac{\sum_r a_r(v_r^{(t)})^{2j}}{\|v^{(t)}\|_2^2}\right)$$

$$= O\left(\frac{a_p(v_p^{(t)})^2}{\|v^{(t)}\|_2^2}\right) \times \left(\kappa^2\|\bar{v}^{(t)}\|_\infty^{2j-4}\right).$$

The above implies that as long as $\|\bar{v}^{(t)}\|_\infty^2 \leq \frac{1}{\text{poly}(\kappa_1)}$, we have:

$$\max\{a_r(v_r^{(t+1)})^2\} \geq \max\{a_r(v_r^{(t)})^2\} \left(1 + \eta\Omega\left(\frac{1}{\text{poly}(\kappa_1)d}\|\bar{v}^{(t)}\|_\infty^2\right)\right). \qquad (52)$$

After that, when $\|\bar{v}^{(t)}\|_\infty^2 \geq \frac{1}{\text{poly}(\kappa_1)}$, we have that $\|\bar{v}^{(t)}\|_\infty^4 \geq \frac{1}{\text{poly}(\kappa_1)}$ as well, which implies

$$\frac{\sum_r a_r(v_r^{(t)})^4}{\|v^{(t)}\|_2^2} \geq \frac{1}{\kappa d}\|\bar{v}^{(t)}\|_\infty^4 \geq \frac{1}{\text{poly}(\kappa_1)d}.$$

Hence, as long as $\|\bar{v}^{(t)}\|_\infty^2 \geq \frac{1}{\text{poly}(\kappa_1)}$, Eq (48) implies that

$$(*) : \|v^{(t+1)}\|_2^2 \geq \|v^{(t)}\|_2^2 \left(1 + \eta\frac{1}{\text{poly}(\kappa_1)d}\right).$$

On the other hand, we also have for every iteration, by Eq (49):

$$\|v^{(t+1)}\|_2^2 \geq \|v^{(t)}\|_2^2 \left(1 - \eta\frac{c_t\,\text{poly}(\kappa_1)}{d^2}\right). \qquad (53)$$

The above implies that $(*)$ can only happen for $\frac{\text{poly}(\kappa_1)d}{\eta}\log\frac{1}{\lambda_0}$ iterations until the norm of v is too large and gradient clipping happens. For these iterations when $\|\bar{v}^{(t)}\|_\infty^2 \geq \frac{1}{\text{poly}(\kappa_1)}$, we can also easily see that

$$\max\{a_r(v_r^{(t+1)})^2\} \geq \max\{a_r(v_r^{(t)})^2\} \left(1 - \eta O\left(\frac{\kappa}{d}\right)\right).$$

For all the other iterations when $\|\bar{v}^{(t)}\|_\infty^2 \leq \frac{1}{\text{poly}(\kappa_1)}$, we have Eq (52) holds, which implies that as long as $\|v^{(t-1)}\|_2^2 \leq \frac{1}{2\lambda_0}$:

$$\eta\sum_{s=1}^{t-1}\|\bar{v}^{(s)}\|_\infty^2 \leq \text{poly}(\kappa_1)d\log\frac{1}{\lambda_0} + \eta\frac{\kappa}{d} \times \frac{\text{poly}(\kappa_1)d}{\eta}\log\frac{1}{\lambda_0}$$

$$\leq \text{poly}(\kappa_1)d\log\frac{1}{\lambda_0}.$$

∎

A.2.2. PROOF OF THE INDUCTIVE HYPOTHESIS

Verifying the inductive hypothesis $\mathcal{H}_0$ during Stage 1. **Proof** [Proof of Proposition 7] Note that the first part of equation (18) has been shown in Claim A.1 — the second part can be shown via a similar proof of Claim A.1. For the rest of the proof, we focus on proving equation (17). The construction of the sequence $\{c_t\}_{t=1}^{T_2}$ will be shown below.

By inequality (42) in the proof of Claim A.1, we know that for every $v^{(0)} \in \mathcal{S}_g$ and $t \leq T_1$, it holds that

$$\frac{(v_i^{(t)})^2}{\mathbb{E}_{\mathcal{P}^{(t)}}[w_i^2]} = \frac{(v_i^{(0)})^2}{\mathbb{E}_{\mathcal{P}_0}[w_i^2]} \pm \frac{\text{poly}(\log d)}{d}, \qquad (54)$$

which implies that for every $t \in [T_1]$, $c_t \leq 2\kappa_1 \kappa c_0$. Now, we focus on $t \in [T_1, T_2]$. By Lemma 8 ,we know that for every $t \geq T_1$, we have that

$$\Delta^{(t)}, \delta_+^{(t)}, \delta_-^{(t)} \leq \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$$

By Proposition 13, we have that for every $v^{(0)} \in \mathcal{S}$, $\left| \left[\nabla_{\geq 4, v^{(t)}} \right]_i \right| \leq \frac{C}{2} \frac{c_t \kappa}{d^2} |v_i^{(t)}|$. Hence,

$$[v_i^{(t+1)}]^2 = [v_i^{(t)}]^2 \pm \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} [v_i^{(t)}]^2.$$

This also implies that

$$\|v^{(t+1)}\|_2^2 = \|v^{(t)}\|_2^2 \left(1 \pm \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} \right). \tag{55}$$

Hence the above implies that

$$c_{t+1} \leq \left(1 + \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} \right) c_t.$$

Iterating the above equation over t gives us the sequence $\{c_t\}_{t=1}^{T_2}$. By maintaining that for every $v \in \mathcal{S}$, the norm of v at iteration t satisfies $c_t \leq \operatorname{poly}(\kappa_1)c_0$ and the fact that $T_2 \leq \frac{d^2}{\eta c_0 \operatorname{poly}(\kappa_1)}$, we have verified the running hypothesis $\mathcal{H}_1$. ∎

A.3. Stage 2.1: Obtaining a Warm Start Initialization

At the beginning of Stage 2, we reduce the gradient truncation parameter. This allows the basis-like neurons to continue to grow and we can obtain a warm start initialization at the end of Stage 2.1 in the sense described in Lemma 11. The proof of Lemma 11 consists of the following steps. First, we analyze the 0th order term in Claim A.6 and A.8. Second, We analyze the 2nd order term in Proposition 17. Combined together, we prove Lemma 11 in Section A.3.1.

Notations for gradients. To facilitate the analysis, we introduce several notations on the gradients of a neuron v. We separate the gradient of v into several components at the t-th iteration as $\nabla_{v,2j} = \nabla_{v,2j,sig} + \nabla_{v,2j,\neg pot} + \nabla_{v,2j,bad} + \nabla_{v,2j,pot \backslash bad}$, where each term is given by

$$\nabla_{2j,v,sig} = - B_{1,2j} \left(\sum_i a_i \langle e_i, v \rangle \langle e_i, \bar{v} \rangle^{2j-2} e_i \right) + b'_{2j} \left(\sum_i a_i \langle e_i, v \rangle \langle e_i, \bar{v} \rangle^{2j-1} \right) \bar{v},$$

$$\nabla_{v,2j,\neg pot} = B_{1,2j} \left(\mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}, w \notin \mathcal{S}_{pot}} \langle w^{(t)}, v \rangle \langle \bar{w}^{(t)}, \bar{v} \rangle^{2j-2} w^{(t)} \right)$$
$$- b'_{2j} \left(\mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}, w \notin \mathcal{S}_{pot}} \langle w^{(t)}, v \rangle \langle \bar{w}^{(t)}, \bar{v} \rangle^{2j-2} \langle w^{(t)}, \bar{v} \rangle \right) \bar{v},$$

$$\nabla_{v,2j,bad} = B_{1,2j} \left(\mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}, w \in \mathcal{S}_{bad}} \langle w^{(t)}, v \rangle \langle \bar{w}^{(t)}, \bar{v} \rangle^{2j-2} w^{(t)} \right)$$
$$- b'_{2j} \left(\mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}, w \in \mathcal{S}_{bad}} \langle w^{(t)}, v \rangle \langle \bar{w}^{(t)}, \bar{v} \rangle^{2j-2} \langle w^{(t)}, \bar{v} \rangle \right) \bar{v},$$

$$\nabla_{v,2j,pot \backslash bad} = B_{1,2j} \left(\mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}, w \in \mathcal{S}_{pot} \backslash \mathcal{S}_{bad}} \langle w^{(t)}, v \rangle \langle \bar{w}^{(t)}, \bar{v} \rangle^{2j-2} w^{(t)} \right)$$
$$- b'_{2j} \left(\mathbb{E}_{w^{(t)} \sim \mathcal{P}, w \in \mathcal{S}_{pot} \backslash \mathcal{S}_{bad}} \langle w^{(t)}, v \rangle \langle \bar{w}^{(t)}, \bar{v} \rangle^{2j-2} \langle w^{(t)}, \bar{v} \rangle \right) \bar{v}.$$

Dynamic of 0th order tensor. Recall that this substage runs for $T_3 \leq \frac{d \log^{1.01} d}{\eta}$ iterations. We first focus on the update of the 0th order term $\Delta^{(t)}$. Let κ_2 denote $e^{\text{poly}(\kappa_1)}$. We show the following claim.

Claim A.6 (Dynamic of 0th order tensor I) *In the setting of Lemma 11, suppose that Proposition 10 holds. Let δ be any value in the range $\left[\frac{\text{poly}(\kappa_2)c_t}{d^2}, \frac{1}{\kappa d}\right]$. When $\Delta^{(t)} \geq \delta$, for any iteration $t \in [T_2 + 1, T_3]$, we have that*

$$\mathop{\mathbb{E}}_{w \sim \mathcal{P}^{(t)}} \|\nabla_w L_\infty(\mathcal{P}^{(t)})\|_2^2 \geq \Omega\left(d^2 \delta^4 / \kappa\right).$$

Proof Let us denote $\delta' = \min\{\delta, \frac{\max\{C_1, C_2\}}{10\kappa d}\}$. We shall see that when $\Delta^{(t)} \geq \delta$, then for every i with $\beta_i^{(t)} + \gamma_i^{(t)} \geq a_i - \frac{\delta'}{4 \max\{C_1, C_2\}}$, we have that

$$-\Delta^{(t)} + C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) \leq -\frac{\delta}{2}$$

Therefore, using equation 59, we have that

$$\mathop{\mathbb{E}}_{w \sim \mathcal{P}^{(t)}, w \notin \mathcal{S}_{pot}} [\nabla_w]_i^2 = \Omega(\beta_i^{(t)} \delta^2).$$

On the other hand, when $\gamma_i^{(t)} \geq a_i - \frac{\delta'}{3 \max\{C_1, C_2\}}$, we have that

$$-\Delta^{(t)} + C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + C_2(a_i - \gamma_i^{(t)}) \leq -\frac{\delta}{6}.$$

This implies that

$$\mathop{\mathbb{E}}_{w \sim \mathcal{P}^{(t)}, w \in \mathcal{S}_{i,pot}} [\nabla_w]_i^2 = \Omega\left(\gamma_i^{(t)} \delta^2\right).$$

In either case, we have that as long as $\beta_i^{(t)} + \gamma_i^{(t)} \geq a_i - \frac{\delta'}{4 \max\{C_1, C_2\}}$, it holds that

$$\mathop{\mathbb{E}}_{w \sim \mathcal{P}^{(t)}, w} [\nabla_w]_i^2 \geq \Omega\left(\frac{d}{\kappa}(\gamma_i^{(t)} + \beta_i^{(t)})\delta^2\delta'\right).$$

Notice that

$$\sum_{i \in [d]} [\beta_i^{(t)} + \gamma_i^{(t)}] 1_{\beta_i^{(t)} + \gamma_i^{(t)} \leq a_i - \frac{\delta'}{4 \max\{C_1, C_2\}}} \leq 1 - d\frac{\delta'}{4 \max\{C_1, C_2\}}.$$

Using $\Delta^{(t)} \geq 0$, we obtain that

$$\sum_{i \in [d]} [\beta_i^{(t)} + \gamma_i^{(t)}] 1_{\beta_i^{(t)} + \gamma_i^{(t)} \geq a_i - \frac{\delta'}{4 \max\{C_1, C_2\}}} \geq d\frac{\delta'}{4 \max\{C_1, C_2\}},$$

which implies that

$$\sum_{i \in [d]} \mathbb{E}_{w \sim \mathcal{P}^{(t)}, w} [\nabla_w]_i^2 \geq \Omega \left(\frac{d}{\kappa} (\delta')^2 \delta^2 \right).$$

∎

Next, we focus on the other side when $\Delta^{(t)}$ is negative. We first show the first lower bound on the neuron mass.

Claim A.7 (Lower bound) *In the setting of Lemma 11, suppose that Proposition 10 holds. Then we have that for any $t \in [T_2 + 1, T_3]$, the following holds:*

$$\beta_i^{(t)} + \gamma_i^{(t)} \geq \frac{1}{\mathrm{poly}(\kappa_1) d}.$$

Proof Initially at $t = 0$, we have that $L_\infty(\mathcal{P}^{(0)}) = O\left(\frac{1}{d}\right)$). Now, for every $\delta \leq \frac{\min\{C_1, 1\}}{100 \kappa d}$, when $\Delta^{(t)} \leq \delta$, we know that as long as $\beta_i^{(t)} + \gamma_i^{(t)} \leq a_i - \frac{2\delta}{C_1}$ and $\beta_i^{(t)} + \gamma_i^{(t)} \geq \frac{1}{\mathrm{poly}(d)}$, we also have that

$$\beta_i^{(t+1)} + \gamma_i^{(t+1)} \geq \beta_i^{(t)} + \gamma_i^{(t)}$$

Thus, when $\beta_i^{(t)} + \gamma_i^{(t)} \leq \frac{a_i}{2}$, it can decrease at next iteration $t + 1$ only when $\delta = \Omega\left(\frac{1}{\kappa d}\right)$, in which case, the total decrement is bounded by $\exp\{-\eta \sum_{t \leq T} |\Delta^{(t)}| 1_{\Delta^{(t)} \geq \delta}\}$. Therefore, taking $\delta = \Theta\left(\frac{1}{\kappa d}\right)$, with the fact that $\beta_i^{(0)} \geq \frac{1}{\kappa d}$, we obtain the result by combining equation 57. ∎

Based on the above claim, we move on to the case when $\Delta^{(t)}$ is negative. We show the following proposition.

Claim A.8 (Dynamic of the 0th order update II) *In the setting of Lemma 11, suppose that Proposition 10 holds. Let δ be any value in the range $\left[\frac{\mathrm{poly}(\kappa_2) c_t}{d^2}, \frac{1}{\kappa d}\right]$. When $\Delta^{(t)} \leq -\delta$, we have*

$$\mathbb{E}_{w \sim \mathcal{P}^{(t)}} \|\nabla_w L_\infty(\mathcal{P}^{(t)})\|_2^2 \geq \Omega \left(\frac{\delta^3 d}{\mathrm{poly}(\kappa_1)} \right).$$

Proof We shall see that when $\Delta^{(t)} \leq -\delta$, then for every i with $\beta_i^{(t)} + \gamma_i^{(t)} \leq a_i + \frac{\delta}{12 \max\{C_1, C_2\}}$, we have that

$$-\Delta^{(t)} + C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) \geq \frac{2}{3} \delta.$$

Therefore, we have that

$$\mathbb{E}_{w \sim \mathcal{P}^{(t)}, w \notin \mathcal{S}_{pot}} [\nabla_w]_i^2 = \Omega(\beta_i^{(t)} \delta^2).$$

On the other hand, $\gamma_i^{(t)} \leq a_i + \frac{\delta'}{12 \max\{C_1, C_2\}}$ as well, this implies that

$$-\Delta^{(t)} + C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + C_2(a_i - \gamma_i^{(t)}) \geq \frac{\delta}{3}.$$

This implies that

$$\mathbb{E}_{w \sim \mathcal{P}^{(t)}, w \in \mathcal{S}_{i,pot}} [\nabla_w]_i^2 = \Omega\left(\gamma_i^{(t)} \delta^2\right).$$

Combining both cases, we have that

$$\mathbb{E}_{w \sim \mathcal{P}^{(t)}, w} [\nabla_w]_i^2 \geq \Omega\left((\gamma_i^{(t)} + \beta_i^{(t)})\delta^2\right).$$

Notice that $\Delta^{(t)} \leq 0$. This implies that

$$\sum_{i \in [d]} 1_{\beta_i^{(t)} + \gamma_i^{(t)} \geq a_i + \frac{\delta'}{12\max\{C_1, C_2\}}} = d - \Omega\left(\frac{\min\left\{\frac{1}{\kappa d}, \delta\right\} d^2}{\kappa}\right).$$

Using the Claim A.7, we have that

$$\sum_{i \in [d]} 1_{\beta_i^{(t)} + \gamma_i^{(t)} \leq a_i + \frac{\delta'}{12\max\{C_1, C_2\}}} [\beta_i^{(t)} + \gamma_i^{(t)}] \geq \frac{\delta d}{\mathrm{poly}(\kappa_1)},$$

which implies that

$$\mathbb{E}_{w \sim \mathcal{P}^{(t)}, w} [\nabla_w]_i^2 \geq \Omega\left(\frac{\delta^2 \min\left\{\frac{1}{\kappa d}, \delta\right\} d}{\mathrm{poly}(\kappa_1)}\right).$$

∎

Claim A.9 *In the setting of Claim A.6 and A.8, for every $T \in [T_2 + 1, T_3]$, the following holds:*

$$\eta \sum_{t=T_2+1}^{T} |\Delta^{(t)}| = O\left(\left(\frac{\eta T \mathrm{poly}(\kappa_1)}{d}\right)^{3/4} + \left(\frac{\eta T \mathrm{poly}(\kappa_1)}{d}\right)^{1/2}\right)$$

Furthermore, we have that:

$$\eta \sum_{t=T_2+1}^{T_3} |\Delta^{(t)}| \leq (\log d)^{0.8}. \tag{56}$$

Proof To prove the above equation, we consider two scenarios. Using Claim A.6, for every $\delta \in \left[\frac{1}{d^{1.5}}, \frac{1}{\kappa d}\right]$, we have:

$$\eta \sum_{t=T_2+1}^{T} |\Delta^{(t)}| 1_{\Delta^{(t)} \geq \delta} = O\left(\frac{\kappa^3}{d^3 \delta^3}\right). \tag{57}$$

Using Claim A.8, we have:

$$\eta \sum_{t=T_2+1}^{T} |\Delta^{(t)}| 1_{\Delta^{(t)} < -\delta} = O\left(\frac{\mathrm{poly}(\kappa_1)}{d^3 \delta^3}\right). \tag{58}$$

Combined together, using the fact that $T_3 \leq \frac{d \log^{1.01} d}{\eta}$, we obtain equation (56). ∎

41

The masses of potential neurons and bad neurons. We now focus on the update of $\gamma_i^{(t)}$ and $\beta_i^{(t)}$. We first prove the following claim on the dynamic of the potential, good, and bad neurons. Based on the claim, we can obtain the update of $\gamma_i^{(t)}$ and $\beta_i^{(t)}$. Let κ_2 denote $e^{\text{poly}(\kappa_1)}$. For any $i \in [d]$, let $\hat{\gamma}_i^{(t)} = \mathbb{E}_{w^{(t)} \sim \mathcal{P}^{(t)}, w \in \mathcal{S}_{i,good}} {w_i^{(t)}}^2$.

Claim A.10 *In the setting of Lemma 11, suppose Proposition 10 holds. There exists fixed constants $C_1, C_2 > 0$ such that for any $t \in [T_2 + 1, T_3]$ and any $i \in [d]$, the update of $\hat{\gamma}_i^{(t)}, \beta_i^{(t)}$ satisfies that*

$$\hat{\gamma}_i^{(t+1)} = \left(1 - \eta \Delta^{(t)} + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + \eta C_2(a_i - \gamma_i^{(t)}) \pm \eta \frac{\text{poly}(\kappa_2) c_t}{d^2}\right) \hat{\gamma}_i^{(t)},$$

$$\beta_i^{(t+1)} = \left(1 - \eta \Delta^{(t)} + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) \pm \eta \frac{\text{poly}(\kappa_2) c_t}{d^2}\right) \beta_i^{(t)}.$$

Moreover, when $\gamma_i^{(t)} \geq \frac{1}{\text{poly}(d)}$, we have that

$$\gamma_i^{(t+1)} = \left(1 - \eta \Delta^{(t)} + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + \eta C_2(a_i - \gamma_i^{(t)}) \pm \eta \frac{\text{poly}(\kappa_2) c_t}{d^2}\right) \gamma_i^{(t)}.$$

The above claim implies that the update between the potential neurons and those not in the potential set differs by a multiplicative factor of $a_i - \gamma_i^{(t)}$. Intuitively, this gap allows us to show that the mass of potential neurons will converge and reduce the value of $a_i - \gamma_i^{(t)}$. On the other hand, the mass of bad neurons $\beta_i^{(t+1)}$ will remain polynomially small throughout the update, since its increment only scales with $\text{poly}(\kappa_2) c_t / d^2$ every iteration. We now describe the proof of the above proposition, which is based on a simple claim that bounds the gradient from irrelevant neurons in equation (59).

Proof We first show the following claim. For every $v \in \mathcal{S}_g$, every $i \in [d]$:

$$\left|\left[\nabla_{v,2j,sig} + \nabla_{v,2j,\neg pot} + \nabla_{v,2j,bad} + \nabla_{v,2j,pot \backslash bad} - \nabla_{2j,v,sd}\right]_i\right| \leq \left(\frac{c_t \text{poly}(\kappa_2)}{d^2}\right) |v_i| \quad (59)$$

To see that the above claim is true, for $v \notin \mathcal{S}_{pot}$, we can bound $\nabla_{v,2j,\neg pot}$ as in Lemma 13. For $v \in \mathcal{S}_{bad}$, we can bound $\nabla_{v,2j,bad}$ directly using equation (22). For $v \in \mathcal{S}_{i,pot}$, we notice

$$\mathbb{E}_{w \sim \mathcal{P}, w \in \mathcal{S}_{pot}, \|w\|_2 \leq d^6} \|w\|_2^2 \leq \frac{1}{\text{poly}(d)}$$

On the other hand, when $v \in \mathcal{S}_{i,pot}$ and $\|v\|_2 \geq d^6$, we have that $\|\bar{v} - e_i\|_2 \leq \frac{1}{d^4}$ by Eq (21). This implies that for

$$\nabla_{2j,v,sd} := -\left(b_{2j} + b'_{2j}\right)\left(\sum_i (a_i - \gamma_i)\langle e_i, v\rangle\langle e_i, \bar{v}\rangle^{2j-2} e_i\right) + b'_{2j}\left(\sum_i (a_i - \gamma_i)\langle e_i, v\rangle\langle e_i, \bar{v}\rangle^{2j-1}\right) \bar{v}.$$

By plugging in the claim in the beginning of the proof into the gradient update rule, we can prove the update rules for each set of neurons. For every $v \notin \mathcal{S}_{pot}$ and every $i \in [d]$, we have that

$$v_i^{(t+1)} = \left(1 - \frac{\eta}{2}\Delta^{(t)} + \eta \frac{C_1}{2}(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) \pm \eta \frac{\text{poly}(\kappa_2) c_t}{d^2}\right) v_i^{(t)}.$$

For every $v \in \mathcal{S}_{i,pot}$ with $|v_i| \geq d^6$, we have that

$$v_i^{(t+1)} = \left(1 - \frac{\eta}{2}\Delta^{(t)} + \eta\frac{C_1}{2}(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + \eta C_2(a_i - \gamma_i^{(t)}) \pm \eta\frac{\mathrm{poly}(\kappa_2)c_t}{d^2} \right) v_i^{(t)}. \quad (60)$$

For all other coordinates $j \neq i$,

$$v_j^{(t+1)} = \left(1 - \frac{\eta}{2}\Delta^{(t)} + \eta\frac{C_1}{2}(a_j - \beta_j^{(t)} - \gamma_j^{(t)}) \pm \eta\frac{\mathrm{poly}(\kappa_2)c_t}{d^2} \right) v_j^{(t)}.$$

By applying the above results on each set of neurons, we obtain the result of this claim. ∎

A.3.1. PROOF OF THE MAIN LEMMA

We are now ready to prove Lemma 11. Based on the dynamic of 0th order tensor and the update of the 2nd order terms shown above, we prove the following proposition that shows $\beta_i^{(t)} + \gamma_i^{(t)}$ cannot be too far away from a_i for too many iterations.

Proposition 17 (Bounds on $\beta_i^{(t)} + \gamma_i^{(t)}$ during Stage 2.1) *In the setting of Lemma 11, suppose that Proposition 10 holds, then for every $t \in [T_2 + 1, T_3]$, we have that*

$$\eta \sum_{s=T_2+1}^{t} |a_i - \beta_i^{(s)} - \gamma_i^{(s)}| \leq (\log d)^{0.9}. \quad (61)$$

Moreover, for every $\delta \leq \frac{1}{100\kappa d}$, we have:

$$\eta \sum_{s=T_2+1}^{t} \mathbb{1}_{\beta_i^{(s)}+\gamma_i^{(s)} \geq a_i+\delta} |a_i - \beta_i^{(s)} - \gamma_i^{(s)}| = O\left(\frac{\mathrm{poly}(\kappa_2)}{d^3\delta^3} \right). \quad (62)$$

Proof Let us construct an auxiliary function

$$\Phi^{(t)} = C_1 \left(a_i - \beta_i^{(t)} - \gamma_i^{(t)} \right)^2 + C_2 \left(a_i - \gamma_i^{(t)} \right)^2.$$

We consider an update step, then it holds that as long as $\beta_i^{(t)} + \gamma_i^{(t)} \leq \frac{\mathrm{poly}(\kappa_2)}{d}$, using Claim A.10, the update of $\Phi^{(t)}$ is given as:

$$\Phi^{(t+1)} = \Phi^{(t)} - 2\eta \left(C_1^2 \left(a_i - \beta_i^{(t)} - \gamma_i^{(t)} \right)^2 \beta^{(t)} + \left(C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + C_2(a_i - \gamma_i^{(t)}) \right)^2 \gamma^{(t)} \right)$$

$$+ 2\eta\Delta^{(t)} \left(C_1 \left(a_i - \beta_i^{(t)} - \gamma_i^{(t)} \right) (\beta_i^{(t)} + \gamma_i^{(t)}) + C_2 \left(a_i - \gamma_i^{(t)} \right) (\gamma_i^{(t)}) \right) \pm \eta\frac{c_t\,\mathrm{poly}(\kappa_2)}{d^4}.$$

This also implies that

$$\Phi^{(t+1)} \leq \Phi^{(t)} - 2\eta \left(C_1^2 \left(a_i - \beta_i^{(t)} - \gamma_i^{(t)} \right)^2 \beta^{(t)} + \left(C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + C_2(a_i - \gamma_i^{(t)}) \right)^2 \gamma^{(t)} \right)$$

$$+ \eta|\Delta^{(t)}|\frac{\mathrm{poly}(\kappa_2)}{d^2} + \eta\frac{c_t\,\mathrm{poly}(\kappa_2)}{d^4},$$

which implies that for every $t \geq 0$:

$$\Phi^{(t+1)} \leq \Phi^{(t)} + \eta|\Delta^{(t)}|\frac{\mathrm{poly}(\kappa_2)}{d^2} + \eta\frac{c_t\,\mathrm{poly}(\kappa_2)}{d^4}. \tag{63}$$

Hence, consider the case that $\beta_i^{(t)} + \gamma_i^{(t)} = a_i - \rho^{(t)}$ for $\rho^{(t)} \geq 0$, we have that $\gamma_i^{(t)} \leq a_i - \rho^{(t)}$. Hence in addition to Eq (63), we also have (using Claim A.7):

$$\Phi^{(t+1)} \leq \Phi^{(t)} - \eta\Omega\left(\frac{1}{\mathrm{poly}(\kappa_1)d}(\rho^{(t)})^2\right) + \eta|\Delta^{(t)}|\frac{\mathrm{poly}(\kappa_2)}{d^2} + \eta\frac{c_t\,\mathrm{poly}(\kappa_2)}{d^4}.$$

Note that originally $\Phi^{(0)} = O\left(\frac{1}{d^2}\right)$ using the fact that $\beta_i^{(0)} \leq 2a_i$ and $\gamma_i^{(0)} \leq \frac{1}{\mathrm{poly}(d)}$, with Claim A.9, we have that for $T \leq \frac{d}{\eta}\log^{1.01} d$:

$$\eta\sum_{t\leq T}\mathbb{1}_{\beta_i^{(t)}+\gamma_i^{(t)}\leq a_i}(a_i - \beta_i^{(t)} - \gamma_i^{(t)})^2 \leq \frac{1}{d}(\log d)^{0.81}.$$

This implies that

$$\eta\sum_{t\leq T}\mathbb{1}_{\beta_i^{(t)}+\gamma_i^{(t)}\leq a_i}(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) \leq \sqrt{\eta\frac{1}{d}(\log d)^{0.81}\times T} \leq \frac{1}{2}(\log d)^{0.9}.$$

Similarly, we can see that when $\beta_i^{(t)} + \gamma_i^{(t)} = a_i + \rho^{(t)}$ for $\rho^{(t)} \geq 0$, then either $\beta_i^{(t)} \geq \rho^{(t)}/2$ or $\gamma_i^{(t)} \geq a_i - \rho^{(t)}/2$. In either case, we have that

$$\Phi^{(t+1)} \leq \Phi^{(t)} - \eta\Omega\left(\frac{1}{\mathrm{poly}(\kappa_2)}(\rho^{(t)})^3\right) + \eta|\Delta^{(t)}|\frac{\mathrm{poly}(\kappa_2)}{d^2} + \eta\frac{c_t\,\mathrm{poly}(\kappa_2)}{d^4}.$$

Hence we can also show that

$$\eta\sum_{t\leq T}\mathbb{1}_{\beta_i^{(t)}+\gamma_i^{(t)}\geq a_i}(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) \leq \frac{1}{2}(\log d)^{0.9}.$$

Eventually, consider for every $\delta \leq \frac{1}{100\kappa d}$, when $\beta_i^{(t)} + \gamma_i^{(t)} \geq a_i + \delta$ and $|\Delta^{(t)}| \leq \frac{d^2}{\mathrm{poly}(\kappa_2)}\delta^3$, then we also have

$$\Phi^{(t+1)} \leq \Phi^{(t)} - \eta\Omega\left(\frac{1}{\mathrm{poly}(\kappa_2)}\delta^3\right) + \eta\frac{c_t\,\mathrm{poly}(\kappa_2)}{d^4}.$$

Using equation 58, we obtain that when $T \leq T_3$,

$$\eta\sum_{t\leq T}\mathbb{1}_{\beta_i^{(t)}+\gamma_i^{(t)}\geq a_i+\delta}|a_i - \beta_i^{(t)} - \gamma_i^{(t)}| = O\left(\frac{\mathrm{poly}(\kappa_2)}{d^3\delta^3}\right).$$

∎

Based on the above proposition, we are ready to prove the main Lemma of Stage 2.1, which provides a warm start initialization at a certain iteration $T_3 = \Theta(d\log d/\eta)$.

Proof [Proof of Lemma 11] We first define T_3 more precisely. We note that initially, for any $i \in [d]$, $\hat{\gamma}_i^{(0)} \leq 1/\mathrm{poly}(d)$ by construction. Using equation (56) and equation (61), by working on $\hat{\gamma}_i^{(t)}$ and noticing that $\hat{\gamma}_i^{(t)} \leq \gamma_i^{(t)}$, we have that there exists an iteration $T^{(i)} = O(d\kappa \log(\frac{1}{\hat{\gamma}_i^{(0)}})/\eta)$ such that at this iteration, $\gamma_i^{T^{(i)}} \geq \frac{1}{10\kappa d}$. We shall fix T_3 to be the maximum of $T^{(i)}$ over $i \in [d]$, which is on the order of $\Theta(d \log d/\eta)$.

Next, similar to the proof of Proposition 17, we consider the function

$$\Phi^{(t)} = \max_{i \in [d]} \left\{ C_1 \left(a_i - \beta_i^{(t)} - \gamma_i^{(t)} \right)^2 + C_2 \left(a_i - \gamma_i^{(t)} \right)^2 \right\}.$$

Let i be the coordinate that achieves the maximum for the function above. We show that

$$\Phi^{(t+1)} \leq \Phi^{(t)} - 2\eta \left(C_1^2 \left(a_i - \beta_i^{(t)} - \gamma_i^{(t)} \right)^2 \beta^{(t)} + \left(C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + C_2(a_i - \gamma_i^{(t)}) \right)^2 \gamma^{(t)} \right)$$
$$+ \eta |\Delta^{(t)}| \frac{\mathrm{poly}(\kappa_2)}{d^2} + \eta \frac{c_t \, \mathrm{poly}(\kappa_2)}{d^4}.$$

Let $\mu = C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + C_2(a_i - \gamma_i^{(t)})$, $\nu = C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)})$, we have that

$$\Phi^{(t)} = (\mu - \nu)^2 + \nu^2,$$

with $\beta_i^{(t)} = \frac{\mu - \nu}{C_2} - \frac{\nu}{C_1}$. So we have when $\gamma_i^{(t)} \geq \frac{1}{\mathrm{poly}(\kappa_3)d}$,

$$\Phi^{(t+1)} \leq \Phi^{(t)} - 2\eta \nu^2 \left(\frac{\mu - \nu}{C_2} - \frac{\nu}{C_1} \right)^2 - \eta \frac{1}{\mathrm{poly}(\kappa_3)d} \mu^2 + \eta |\Delta^{(t)}| \frac{\mathrm{poly}(\kappa_2)}{d^2} + \eta \frac{c_t \, \mathrm{poly}(\kappa_2)}{d^4}.$$

When $\Phi^{(t)} \geq \delta$, we have that either $\mu^2 \geq \frac{\delta}{100}$, or $\mu^2 \leq \frac{\delta}{100}$ and $\nu^2 \geq \frac{\delta}{2}$. In the first case, we have that

$$\Phi^{(t+1)} \leq \Phi^{(t)} - \eta \frac{1}{\mathrm{poly}(\kappa_3)d} \delta + \eta |\Delta^{(t)}| \frac{\mathrm{poly}(\kappa_2)}{d^2} + \eta \frac{c_t \, \mathrm{poly}(\kappa_2)}{d^4}.$$

In the second case, we have that

$$\Phi^{(t+1)} \leq \Phi^{(t)} - \eta \Omega(\delta^2) + \eta |\Delta^{(t)}| \frac{\mathrm{poly}(\kappa_2)}{d^2} + \eta \frac{c_t \, \mathrm{poly}(\kappa_2)}{d^4}.$$

Combining this equation with the bound in equation (56), we know that for $\delta = \frac{1}{d \log^{0.01} d}$, we have that $\Phi^{(t)} \geq \delta$ can only happen for at most $\frac{d \log^{0.5} d}{\eta}$ many of the iterations within $t \in [T_2 + 1, T_3]$. Combining the above with equation (58), we obtain the desired result. ∎

A.4. Stage 2.2: The Final Substage

In this section, we present the proof of Lemma 12 for the final substage. In the end, we prove the running inductive hypothesis $\mathcal{H}_1$ in Proposition 10.

Proof [Proof of Lemma 12] Suppose the lemma holds at iteration t, then using the condition at iteration t, together with $\Phi^{(0)} = O\left(\frac{1}{d\log^{0.01}d}\right)$, we have that

$$\forall i \in [d] : \gamma_i^{(t)} \geq a_i - \frac{1}{d\log^{0.001}d} \quad \text{and} \quad \beta_i^{(t)} \leq \frac{1}{d\log^{0.001}d} \tag{64}$$

Define the following function

$$\tau_i := C_1(a_i - \beta_i - \gamma_i) + \frac{C_2\gamma_i}{\beta_i + \gamma_i}\left(a_i - \gamma_i\right),$$

Then we have the following

$$\beta_i^{(t+1)} + \gamma_i^{(t+1)} = \left(\beta_i^{(t)} + \gamma_i^{(t)}\right)\left(1 - \eta\Delta^{(t)} + \eta\tau_i^{(t)} \pm \eta\frac{\text{poly}(\kappa_2)c_t}{d^2}\right).$$

Similar to the proof of Lemma 14, we have that as long as $\Delta_+^{(t)} \geq \frac{\text{poly}(\kappa_2)c_t}{d^2}$ and

$$\Delta_+^{(t)} \geq \left(1 - \frac{1}{\text{poly}(\kappa)}\right)\delta_-^{(t)}.$$

Then it must satisfy that $\Delta_+^{(t+1)} \leq \Delta_+^{(t)}\left(1 - \eta\frac{1}{d\,\text{poly}(\kappa)}\right)$. Hence, if the maximizer of Φ is Δ_+, Then it must be the case that

$$\Delta_+^{(t+1)} \leq \Delta_+^{(t)}\left(1 - \eta\frac{1}{d\,\text{poly}(\kappa)}\right). \tag{65}$$

Now, consider another case when $\Delta_-^{(t)} \leq \left(1 - \frac{1}{\text{poly}(\kappa)}\right)\delta_+^{(t)}$, let i be the argmax of $\{\tau_j\}_{j\in[d]}$, then we must have that

$$\beta_i^{(t+1)} + \gamma_i^{(t+1)} = \left(\beta_i^{(t)} + \gamma_i^{(t)}\right)\left(1 - \eta\Delta^{(t)} + \eta\delta_+^{(t)} \pm \eta\frac{\text{poly}(\kappa_2)c_t}{d^2}\right)$$
$$\geq \left(\beta_i^{(t)} + \gamma_i^{(t)}\right)\left(1 + \eta\frac{1}{\text{poly}(\kappa)}\delta_-^{(t)} - \eta\frac{\text{poly}(\kappa_2)c_t}{d^2}\right).$$

Hence as long as $\delta_-^{(t)} \geq \frac{\text{poly}(\kappa_2)c_t}{d^2}$, we have that

$$\beta_i^{(t+1)} + \gamma_i^{(t+1)} \geq \left(\beta_i^{(t)} + \gamma_i^{(t)}\right)\left(1 + \eta\frac{1}{\text{poly}(\kappa)}\delta_-^{(t)}\right). \tag{66}$$

On the other hand, since

$$\gamma_i^{(t+1)} = \gamma_i^{(t)}\left(1 - \eta\Delta^{(t)} + \eta C_1(a_i - \beta_i - \gamma_i) + \eta C_2\left(a_i - \gamma_i\right) \pm \eta\frac{\text{poly}(\kappa_2)c_t}{d^2}\right)$$
$$\geq \gamma_i^{(t)}\left(1 - \eta\Delta^{(t)} + \eta\delta_-^{(t)} - \eta\frac{\text{poly}(\kappa_2)c_t}{d^2}\right)$$
$$\geq \gamma_i^{(t)}\left(1 + \eta\frac{1}{\text{poly}(\kappa)}\delta_-^{(t)}\right). \tag{67}$$

Notice that

$$f(\gamma_i) = \gamma_i \left(a_i - \gamma_i \right)$$

is a decreasing function of γ_i with slop at least $0.5\gamma_i$ when $\gamma_i \geq \frac{a_i}{2}$, which holds true using Eq (64). Combining Eq (66) and Eq (67), we have that if the maximizer of Φ is δ_-, the following is true

$$\delta_-^{(t+1)} \leq \delta_-^{(t+1)} \left(1 - \eta \frac{1}{\text{poly}(\kappa)d} \right). \tag{68}$$

Consider another case when the maximizer is Δ_-. Similar to the proof of Lemma 14, as long as $\Delta_-^{(t)} \geq \frac{\text{poly}(\kappa_2)c_t}{d^2}$, we have that

$$\Delta_-^{(t)} \geq \left(1 - \frac{1}{\text{poly}(\kappa)} \right) \delta_+^{(t)}.$$

Hence, if the maximizer of Φ is Δ_-, then it must be the case that

$$\Delta_-^{(t+1)} \leq \Delta_-^{(t)} \left(1 - \eta \frac{1}{\text{poly}(\kappa)d} \right). \tag{69}$$

Moreover, using the fact that when $\Delta_-^{(t)} \leq \left(1 - \frac{1}{\text{poly}(\kappa)} \right) \delta_+^{(t)}$, let i be the argmax of $\{-\tau_j\}_{j \in [d]}$, then we must have that as long as $\delta_+^{(t)} \geq \frac{\text{poly}(\kappa_2)c_t}{d^2}$, we have that

$$\beta_i^{(t+1)} + \gamma_i^{(t+1)} \leq \left(\beta_i^{(t)} + \gamma_i^{(t)} \right) \left(1 - \eta \frac{1}{\text{poly}(\kappa)} \delta_+^{(t)} \right)$$

Consider two cases.

1. The maximizer is δ_+. Then we must have that for every $i \in [d]$, $\beta_i^{(t)} \leq C\delta_+^{(t)}$, then we must have that $a_i - \gamma^{(t)} \leq C\delta_+^{(t)}$ as well. Hence, it holds that

$$
\begin{aligned}
\gamma_i^{(t+1)} &\leq \gamma_i^{(t)} \left(1 - \eta \frac{1}{\text{poly}(\kappa)} \delta_+^{(t)} + \eta \frac{\beta_i}{\gamma_i + \beta_i} (a_i - \gamma_i^{(t)}) \right) \\
&\leq \gamma_i^{(t)} \left(1 - \eta \frac{1}{\text{poly}(\kappa)} \delta_+^{(t)} + \eta 2\kappa d (C\delta_+^{(t)})^2 \right) \\
&\leq \gamma_i^{(t)} \left(1 - \eta \frac{1}{\text{poly}(\kappa)} \delta_+^{(t)} \right).
\end{aligned}
$$

Hence if the maximizer of Φ is δ_+, it must be the case:

$$\delta_+^{(t+1)} \leq \delta_+^{(t+1)} \left(1 - \eta \frac{1}{\text{poly}(\kappa)d} \right). \tag{70}$$

2. The maximizer is β_+. Then there is a $j \in [d]$ such that $\beta_j^{(t)} \geq C\delta_+^{(t)}$, $\beta_j^{(t)} \geq C\delta_-^{(t)}$ and $\beta_j^{(t)} \geq C|\Delta^{(t)}|$, we have that for this j, it holds: let $S = C_1 \left(\beta_j^{(t)} + \gamma_j^{(t)} - a_j \right)$ and $\rho = (a_j - \gamma_j^{(t)})$, we have: if $\rho \leq \frac{1}{4}\beta_j^{(t)}$, then

$$S \geq C_1 \beta_j^{(t)} - 2C_1 \rho \geq \frac{C_1}{2} \beta_j^{(t)} \geq 2|\Delta^{(t)}|.$$

On the other hand if $\rho > \frac{1}{4}\beta_j^{(t)}$, then using $\delta_-^{(t)}$, we have:

$$-S + \rho \le \delta_-^{(t)}.$$

This implies that

$$S \ge \rho - \delta_-^{(t)} \ge \frac{1}{4}\beta_j^{(t)} - \delta_-^{(t)} \ge |\Delta^{(t)}| + \frac{1}{8}\beta_j.$$

Hence if the maximizer of Φ is β_+, it must be the case:

$$\beta_j^{(t+1)} = \beta_j^{(t)} \left(1 - \eta\Delta^{(t)} - \eta S \pm \eta\frac{\text{poly}(\kappa_2)c_t}{d^2}\right)$$

$$\le \beta_j^{(t)} \left(1 - \eta\frac{1}{8}\beta_j^{(t)}\right). \tag{71}$$

To sum up, the result follows by combining Eq (71), (70), (69), (65) and (68). ∎

A.4.1. PROOF OF THE INDUCTIVE HYPOTHESIS

Verifying the inductive hypothesis $\mathcal{H}_1$ during Stage 2. **Proof** [Proof of Proposition 10] We first verify the inductive hypothesis for $t \le T_3$. The bound for $v \notin \mathcal{S}_{pot}$ follows from Claim A.9 and Proposition 17. We prove the bound for $v \in \mathcal{S}_{pot}$ by tracking the gradient descent dynamic. Following Eq (46), for every neuron v, and every $p \in [d]$, define

$$Q_p^{(t)} := 2\sum_{j \ge 2} \left(B_{1,2j}(a_p - \gamma_p^{(t)})\frac{(v_p^{(t)})^{2j-2}}{\|v^{(t)}\|_2^{2j-2}} - B_{2,2j}\frac{\sum_r (a_r - \gamma_r^{(t)})(v_r^{(t)})^{2j}}{\|v^{(t)}\|_2^{2j}}\right),$$

$$R_p^{(t)} := 2\sum_{j \ge 2} \left(B_{1,2j}(a_p - \gamma_p^{(t)})\frac{(v_p^{(t)})^{2j-2}}{\|v^{(t)}\|_2^{2j-2}}\right).$$

Hence, using Eq (59), we have that for every $i \in [d]$

$$\left[v_i^{(t+1)}\right]^2 = \left[v_i^{(t)}\right]^2 \left(1 - \eta\Delta^{(t)} + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + \eta Q_p^{(t)} \pm \eta\frac{c_t\,\text{poly}(\kappa_1)}{d^2}\right). \tag{72}$$

Hence, we have that for every $i, j \in [d]$,

$$\frac{\left[v_i^{(t+1)}\right]^2}{\left[v_j^{(t+1)}\right]^2} = \frac{\left[v_i^{(t)}\right]^2}{\left[v_j^{(t)}\right]^2} \left(1 + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) - \eta C_1(a_j - \beta_j^{(t)} - \gamma_j^{(t)}) + \eta R_i^{(t)} - \eta R_j^{(t)} \pm \frac{c_t\,\text{poly}(\kappa_1)}{d^2}\right).$$

Now, if $|\bar{v}_i^{(t)}|^2, |\bar{v}_j^{(t)}|^2 \le \frac{c_t}{d}$, we also have that

$$\frac{\left[v_i^{(t+1)}\right]^2}{\left[v_j^{(t+1)}\right]^2} = \frac{\left[v_i^{(t)}\right]^2}{\left[v_j^{(t)}\right]^2} \left(1 + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) - \eta C_1(a_j - \beta_j^{(t)} - \gamma_j^{(t)}) \pm \eta\frac{c_t\,\text{poly}(\kappa_1)}{d^2}\right).$$

This implies that

$$\frac{\left[v_i^{(t+1)}\right]^2}{\left[v_j^{(t+1)}\right]^2} = \frac{\left[v_i^{(0)}\right]^2}{\left[v_j^{(0)}\right]^2} \exp\left\{ \pm 2\eta C_1 \sum_{s \leq t} \left(|a_i - \beta_i^{(t)} - \gamma_i^{(t)}| + |a_j - \beta_j^{(t)} - \gamma_j^{(t)}| \right) \pm \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^2} t \pm \eta^2 t \right\}.$$

Hence using Proposition 17 we show that when $|\bar{v}_i^{(0)}|^2, |\bar{v}_j^{(0)}|^2 \leq \frac{c_0}{d}$, then $|\bar{v}_i^{(t)}|^2, |\bar{v}_j^{(t)}|^2 \leq \frac{c_t}{d}$ as well for every $t \leq T_3$. Now, we need to give an upper bound on the the coordinates of the neurons. For every $v \notin \mathcal{S}_{pot}$, we know that all coordinates $j \in [d]$ satisfies that $|\bar{v}_j^{(0)}|^2 \leq \frac{c_0}{d}$. Hence, by Eq (72), we have that

$$\left[v_i^{(t+1)}\right]^2 \leq \left[v_i^{(t)}\right]^2 \left(1 + \eta|\Delta^{(t)}| + \eta C_1 |a_i - \beta_i^{(t)} - \gamma_i^{(t)}| + \eta \kappa O(\|v^{(t)}\|_4^4) \right),$$
$$\leq \left[v_i^{(t)}\right]^2 \left(1 + \eta|\Delta^{(t)}| + \eta C_1 |a_i - \beta_i^{(t)} - \gamma_i^{(t)}| + \eta \kappa \frac{c_t^2}{d^2} \right).$$

Hence we have that

$$\left[v_i^{(t+1)}\right]^2 \leq \left[v_i^{(0)}\right]^2 \exp\left\{ \eta \sum_{s \leq t} \left(|\Delta^{(s)}| + C_1 |a_i - \beta_i^{(s)} - \gamma_i^{(s)}| \right) + \eta \kappa \frac{c_t^2}{d} t \right\}.$$

Hence using Proposition 17 and Claim A.9, we have proved Eq (20) and Eq (21).

Next, we proceed to the norm of neurons $v \in \mathcal{S}_{i,good}$. For this neuron, using the fact that $|v_i^{(t)}| \geq d^6$ and equation 60, we have that

$$v_i^{(t+1)} = \left(1 - \eta\Delta^{(t)} + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + \eta C_2(a_i - \gamma_i^{(t)}) \pm \eta \frac{\operatorname{poly}(\kappa_2)c_t}{d^2} \right) v_i^{(t)}.$$

Hence, for every t, using Eq (61) we obtain that:

$$\left[v_i^{(t)}\right]^2 \geq \frac{\left[v_i^{(0)}\right]^2}{d} \geq \frac{1}{\lambda_0 \operatorname{poly}(d)}.$$

Notice that for every neuron $v \in \mathcal{S}_g$, we have that $|\bar{v}_i^{(0)}|^2 \leq \frac{c_0}{d}$ can happen for at most $O(\log^{0.01} d)$ many $i \in [d]$. Denote this set as $\mathcal{Q}_v$, we have that $\left[v_i^{(t+1)}\right]^2 / \left[v_i^{(t)}\right]^2$ is at most

$$1 - \eta\Delta^{(t)} + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + \eta C_2(a_i - \gamma_i^{(t)}) + \eta O\left(\sum_{p \in \mathcal{Q}_v} [(\gamma_p^{(t)} - a_p)]^+ \right) + \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}$$

$$\leq 1 - \eta\Delta^{(t)} + \eta C_1(a_i - \beta_i^{(t)} - \gamma_i^{(t)}) + \eta C_2(a_i - \gamma_i^{(t)}) + \eta O\left(\sum_{p \in \mathcal{Q}_v} |a_p - \beta_p^{(t)} - \gamma_p^{(t)}| \right) + \eta \frac{c_t \operatorname{poly}(\kappa_1)}{d^2}.$$

Hence, for every $t \leq T \leq T_3$, using Eq (56) and Eq (61), by working on $\hat{\gamma}_i$ and notice that $\hat{\gamma}_i \leq \gamma_i$, we conclude that for every $i \in [d]$.

$$\eta \sum_{t \leq T} |a_i - \gamma_i^{(t)}| \leq \log\left(\frac{d}{\hat{\gamma}_i^{(0)}} \right).$$

49

Combining the above equation with Eq (61) we have that for every $i \in [d]$:

$$\left[v_i^{(t)}\right]^2 \leq \left[v_0^{(t)}\right]^2 \exp\{\Gamma_i\} \leq \left[v_0^{(t)}\right]^2 \frac{d}{\hat{\gamma}_i^{(0)}}$$

This proves that gradient truncation never happens during this substage. Now, apply Lemma 9, which says that

$$\hat{\gamma}_i^{(0)} \geq \mu(\mathcal{S}_{i,good}) \frac{\lambda_0}{\text{poly}(d)} \geq \text{poly}(d)\mu(\mathcal{S}_{i,bad}). \tag{73}$$

We complete the proof of the first our statements. For the last statement on γ, β, Claim A.8 also proves the upper bound on $\gamma_i^{(t)} + \beta_i^{(t)}$ as in equations (23) and (24). Taking $\delta = \frac{1}{\kappa d}$, we can show that

$$\beta_i^{(t)} + \gamma_i^{(t)} \leq \frac{\text{poly}(\kappa_2)}{d}.$$

Next verify the running inductive hypothesis $\mathcal{H}_2$ for $T_3 \leq t \leq T_4$. Based on Lemma 12, we have the following bounds on the update of each coordinate of each neuron. For every $i \in [d]$, using Eq (59), we have that

$$|v_i^{(t+1)}| = |v_i^{(t)}| \left(1 \pm \eta O\left(|\Delta^{(t)}| + |a_i - \beta_i^{(t)} - \gamma_i^{(t)}| + |a_i - \gamma_i^{(t)}|\right) \pm \frac{c_t \kappa_1}{d^2}\right)$$

Note that by the definition of $\Phi^{(t)}$ at Lemma 12, we have that

$$|a_i - \gamma_i| \leq O\left(\left|C_1(a_i - \gamma_i) + \frac{C_2 \gamma_i}{\beta_i + \gamma_i}(a_i - \gamma_i)\right|\right) + O(|\beta_i|)$$
$$\leq O\left(\beta_+ + \delta_+ + \delta_-\right) = O(\Phi) \tag{74}$$

Hence, we obtain that for $t \in [T_3, T_4]$, with Lemma 12:

$$|v_i^{(t)}| = |v_i^{(T_3)}| \exp\left\{\pm O\left(\log \frac{t}{d}\right)\right\}$$

Hence as long as $T_4 \leq \frac{d^{1+10Q}}{\eta}$, we obtain the running hypothesis $\mathcal{H}_2$ at this substage. ∎

Appendix B. Proof of the Finite-Width Case

We begin by describing the connection between the finite-width dynamic and the infinite-width dynamic. For a vector $w \in \mathcal{S}_g$, let $\tilde{w}^{(0)}$ and $w^{(0)}$ be a neuron with initialization w in the finite-width and infinite-width case, respectively. Let $\mathcal{P}$ denote the infinite neuron distribution and $\tilde{\mathcal{P}}$ denote the finite neuron population with m samples. Our idea is to track the difference between $\tilde{w}^{(t)}$ and $w^{(t)}$, denoted by $\xi_w^{(t)}$, throughout the update. The neuron w denotes a weight vector from the infinite width case that we specify below. Specifically, the truncated gradient descent update of $\tilde{w}$ for the finite-width case and the update of w for the infinite-width case is equal to the following.

$$w^{(t+1)} = w^{(t)} - \eta \cdot \mathbb{1}_{\|w^{(t)}\|^2 \leq 1/\lambda} \nabla_{w^{(t)}} L_\infty(\mathcal{P}), \tag{75}$$

$$\tilde{w}^{(t+1)} = \tilde{w}^{(t)} - \eta \cdot \mathbb{1}_{\|\tilde{w}^{(t)}\|^2 \leq 1/\lambda} \nabla_{\tilde{w}^{(t)}} L(\tilde{\mathcal{P}}) + \eta \Xi_w^{(t)}, \tag{76}$$

where $\Xi_w^{(t)}$ is an extra error term that arises from the sampling error of the empirical loss.

Our main result in this section is that provided with polynomially many neuron samples and training samples, the errors $\xi_w^{(t)}$ and $\Xi_w^{(t)}$ in equation (76) remain polynomially small throughout Algorithm 1. We first state the result for Stage 1.

Lemma 18 (Error propagation of Stage 1) *In the setting of Theorem 1, let $\tilde{\mathcal{P}}^{(0)}$ be a uniform distribution over m i.i.d. samples from $\mathcal{P}$. There exists a fixed value $\Xi \in [0, \lambda_0/\mathrm{poly}(d)]$ such that for every iteration $t \leq T_2$, the average norm of the error is small: $\mathbb{E}_{\tilde{w} \sim \tilde{\mathcal{P}}^{(t)}} \|\xi_w\|_2^2 \leq \mathrm{poly}(d)\Xi$.*

Furthermore, for every $\tilde{w}^{(t)}$ in $\tilde{\mathcal{P}}^{(t)}$, the individual error terms are small: $\|\Xi_w^{(t)}\|_2^2 \leq \Xi$ and $\|\xi_w^{(t)}\|_2^2 \leq \mathrm{poly}(d) \cdot \Xi/\lambda_0$.

The proof of Lemma 18 can be found in Section B.2.3. Next, we consider the error propagation of Stage 2.1. We show that the norm of ξ_w is much smaller than that of w.

Lemma 19 (Error propagation of Stage 2.1) *In the setting of Theorem 1, let $\tilde{\mathcal{P}}^{(T_2+1)}$ be a uniform distribution over m i.i.d. samples from $\mathcal{P}^{T_2+1}$. There exists a fixed value $\Xi \in \left[0, \frac{1}{\mathrm{poly}_\kappa(d)}\right]$ such that for every iteration $T_2 < t \leq T_3$ and every neuron $\tilde{w}^{(t)}$ in $\tilde{\mathcal{P}}^{(t)}$, the error terms are small:*

$$\|\Xi_w^{(t)}\|_2^2 \leq \Xi, \|\xi_w^{(t)}\|_2^2 \leq \min(\mathrm{poly}_\kappa(d)\Xi, \|\xi_w^{(t)}\|_2^2 \leq \|w^{(t)}\|_2^2/d^{20}).$$

The proof of Lemma 19 involves carefully studying the error term and follows a similar argument to Lemma 18. The details can be found in Section B.3. Finally, we consider the error terms in the final stage. We use a different error analysis. At iteration T_3, let us consider the set

$$\mathcal{S}_{i,singleton} := \left\{ v \in \mathcal{S}_g \mid \|\bar{v} - e_i\|_2 \leq \frac{1}{\mathrm{poly}(d)} \right\}, \text{ for } 1 \leq i \leq d.$$

Let $\mathcal{S}_{singleton} = \cup_{i=1}^d \mathcal{S}_{i,singleton}$. Consider the set

$$\mathcal{S}_{ignore} := \{v \in \mathcal{S}_g \mid v \in \mathcal{S}_{pot}, v \notin \mathcal{S}_{singleton}\},$$

where we recall the definition of $\mathcal{S}_{pot}$ in Proposition 10. We state the error propagation of the final substage as follows.

Lemma 20 (Error propagation of Stage 2.2) *In the setting of Theorem 1, let $\tilde{\mathcal{P}}^{(T_3+1)}$ be a uniform distribution over m i.i.d. samples from $\mathcal{P}^{(T_3+1)}$. There exists a fixed value $\Xi \in [0, 1/\mathrm{poly}_\kappa(d)]$ such that for every iteration $T_3 < t \leq T_4$ and for every $\tilde{w}^{(t)}$, the error term $\|\Xi_w^{(t)}\|_2^2 \leq \Xi$.*

For every neuron w with $w \notin \mathcal{S}_{ignore}$, we have that $\|\xi_w^{(t)}\|_2^2 \leq \|w^{(t)}\|_2^2/\mathrm{poly}(d)$ and

$$\mathop{\mathbb{E}}_{\tilde{w} \sim \tilde{\mathcal{P}}^{(t)}, w \in \mathcal{S}_{ignore}} \|\xi_w^{(t)}\|_2^2 \leq 1/\mathrm{poly}(d).$$

The proof of Lemma 20 can be found in Section B.4. Based on the analysis of error propagation, we are now ready to prove our main result. We prove Theorem 1 as follows.
Proof [Proof of Theorem 1] Let us denote $\tilde{w}^{(t)}$ to be the weight of the neuron $\tilde{w}$ at iteration t, following the update of Algorithm 1. For the next iteration, we have that

$$\tilde{w}^{(t+1)} = \tilde{w}^{(t)} - \eta \cdot \mathbb{1}_{\|\tilde{w}^{(t)}\|_2^2 \leq \frac{1}{2\lambda}} \nabla_{\tilde{w}_i^{(t)}} \hat{L}(\tilde{\mathcal{P}}^{(t)}),$$

51

where $\hat{L}(W)$ denotes the empirical loss.

Recall that we assume that the learning rate $\eta \le \frac{1}{\text{poly}_\kappa(d)}$. Using Claim 2.1, we can see that when N for a sufficiently large polynomial in d, we have that with probability at least $1 - e^{-\log^2 d}$, for every $w \in W$ and every $t \le T_4$: $\left\| \nabla_{\tilde{w}^{(t+1)}} \hat{L}(\tilde{\mathcal{P}}^{(t+1)}) - \nabla_{w^{(t+1)}} L_\infty(\mathcal{P}) \right\|_2 \le \frac{1}{\text{poly}_\kappa(d)}$.

For Stage 1, we can first apply Lemma 18 with $\Xi = \frac{1}{\text{poly}_\kappa(d)}$ and $\tilde{\mathcal{P}}$ being a uniform distribution over W. Using $m = \text{poly}_\kappa(d)$, we can conclude that for every $w \in W$ and $t \le T_2$: $\|\tilde{w}^{(t)} - w^{(t)}\|_2 \le \frac{1}{\text{poly}_\kappa(d)}$.

For Stage 2.1, we can use Lemma 19 with $\Xi = \frac{1}{\text{poly}_\kappa(d)}$ to conclude that for every $t \le T_3$, $\|\tilde{w}^{(t)} - w^{(t)}\|_2 \le \frac{1}{\text{poly}_\kappa(d)}$ as well. For Stage 2.2, we shall use Lemma 20 with $\Xi = \frac{1}{\text{poly}(d)}$ to conclude that for every neuron $w \notin \mathcal{S}_{ignore}$, we have:

$$\|\tilde{w}^{(T_4)} - w^{(T_4)}\|_2^2 \le \frac{1}{\text{poly}(d)} \|w\|_2^2 \text{ and } \mathop{\mathbb{E}}_{\tilde{w}^{(T_4)} \sim \tilde{\mathcal{P}}^{(T_4)}, w \in \mathcal{S}_{ignore}} \|\tilde{w}^{(T_4)} - w^{(T_4)}\|_2^2 \le \frac{1}{\text{poly}(d)}.$$

These statements together give us the following

$$\mathop{\mathbb{E}}_{x \sim \mathcal{N}(0, \text{Id}_{d \times d})} \left(f_{\{\tilde{w}^{(T_4)} | w \in W\}}(x) - f_{\{w^{(T_4)} | w \in W\}}(x) \right)^2 \le \frac{1}{\text{poly}(d)}.$$

Finally, combined with Claim 2.1 and Theorem 4 we complete the proof of Theorem 1. ∎

B.1. Stage 1.1: Analysis of 0th and 2nd Order Tensor Decompositions

In this substage, we consider the error terms in the gradients of the 0th and 2nd order tensor decompositions. Let $\tilde{\mathcal{H}}_0$ is the running hypothesis $\mathcal{H}_0$ in Proposition 7 without the conditionally-symmetric property. For a neuron v, similar to the definition of ∇_v in Claim 3.2, we define the error gradient $\tilde{\nabla}_v$ as $\tilde{\nabla}_v := \nabla_v L_\infty(\tilde{\mathcal{P}}) = \sum_{j \ge 0} \tilde{\nabla}_{2j,v}$. Using the update of equation (75) for the infinite-width case, the gradient of v for the 0th and 2nd order terms over the population loss is given by

$$\nabla_{\le 2,v} = b_0 \left(\mathop{\mathbb{E}}_{\mathcal{P}} \|w\|_2^2 - 1 \right) v + b_2 \left(\mathop{\mathbb{E}}_{\mathcal{P}} ww^\top - A \right) v,$$

where $A = \text{diag}(\{a_i\}_{i \in [d]})$. Using the update of equation (76) for the finite-width case, the gradient of v for the 0th and 2nd order terms over the population loss is given by

$$\tilde{\nabla}_{\le 2,v} = b_0 \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} \|\tilde{w}\|_2^2 - 1 \right) \tilde{v} + b_2 \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} \tilde{w}\tilde{w}^\top - A \right) \tilde{v}.$$

The first order terms of the error term ξ_v for the neuron v is given by

$$\tilde{\nabla}_{0,v,1} := 2b_0 \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} \langle w, \xi_w \rangle \right) v + b_0 \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} \|w\|_2^2 - 1 \right) \xi_v,$$

$$\tilde{\nabla}_{2,v,1} := b_2 \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} ww^\top - A \right) \xi_v + b_2 \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} w\xi_w^\top \right) v + b_2 \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} \xi_w w^\top \right) v. \tag{77}$$

Let $\tilde{\nabla}_{\le 2,v,1}$ denote the sum of $\tilde{\nabla}_{0,v,1}$ and $\tilde{\nabla}_{2,v,1}$. We have the following claim for the error of the 0th and 2nd order terms. We use the notation $(w, \xi_w) \sim \tilde{\mathcal{P}}$ to denote a neuron $\tilde{w} = w + \xi_w$ sampled from $\tilde{\mathcal{P}}$.

Claim B.1 (Error of 0th and 2nd order gradients) *In the setting of Theorem 1, at every iteration $t \leq T_4$, the following is true for the neuron distribution $\tilde{\mathcal{P}} = \tilde{\mathcal{P}}^{(t)}$,*

$$\left(\underset{(w,\xi_w)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \langle w, \xi_w \rangle \right) \underset{(v,\xi_v)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \langle v, \xi_v \rangle \geq 0, \text{ and}$$

$$\underset{(v,\xi_v)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \left(\left\langle \left(\underset{(w,\xi_w)\sim\tilde{\mathcal{P}}}{\mathbb{E}} w\xi_w^\top \right) v, \xi_v \right\rangle \right) + \underset{(v,\xi_v)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \left(\left\langle \left(\underset{(w,\xi_w)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \xi_w w^\top \right) v, \xi_v \right\rangle \right) \geq 0,$$

As a corollary,

$$\underset{(v,\xi_v)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \langle \tilde{\nabla}_{\leq 2,v,1}, \xi_v \rangle \geq b_0 \left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}} \|w\|_2^2 - 1 \right) \underset{\tilde{\mathcal{P}}}{\mathbb{E}}[\|\xi_v\|_2^2] + b_2 \underset{\tilde{\mathcal{P}}}{\mathbb{E}} \left[\xi_v^\top \left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}} ww^\top - A \right) \xi_v \right]. \tag{78}$$

Proof The first inequality is obviously true. Now we consider the second inequality, we have that

$$\langle w\xi_w^\top v, \xi_v \rangle + \langle \xi_w w^\top v, \xi_v \rangle = \xi_v^\top w\xi_w^\top v + \xi_v^\top \xi_w w^\top v$$
$$= \frac{1}{2} \mathrm{Tr} \left(\left(\xi_v v^\top + v\xi_v^\top \right) \left(\xi_w w^\top + w\xi_w^\top \right) \right).$$

This implies that

$$\underset{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \langle w\xi_w^\top v, \xi_v \rangle + \langle \xi_w w^\top v, \xi_v \rangle$$
$$= \frac{1}{2} \underset{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \mathrm{Tr} \left(\left(\xi_v v^\top + v\xi_v^\top \right) \left(\xi_w w^\top + w\xi_w^\top \right) \right)$$
$$= \frac{1}{2} \mathrm{Tr} \left(\underset{(w,\xi_w)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \left(\xi_w w^\top + w\xi_w^\top \right) \right)^2 \geq 0.$$

∎

Next we consider the first order tensor. The first order gradient in the finite-width case for the population loss is

$$\tilde{\nabla}_{1,v} = b_1 \underset{\tilde{\mathcal{P}}}{\mathbb{E}} \left(\langle \tilde{w}, \tilde{v} \rangle \|\tilde{w}\|_2 \bar{\tilde{v}} + \|\tilde{w}\|_2 \|\tilde{v}\|_2 \tilde{w} \right).$$

The 1-order term in the gradient is zero in the infinite-width case of Section A. The first-order expansion of the error is given by:

$$\tilde{\nabla}_{1,v,1} := b_1 \underset{\tilde{\mathcal{P}}}{\mathbb{E}} \langle \bar{w}, \xi_w \rangle \langle w, v \rangle \bar{v} + b_1 \underset{\tilde{\mathcal{P}}}{\mathbb{E}} \langle \bar{v}, \xi_w \rangle \|w\|_2 v$$
$$+ b_1 \underset{\tilde{\mathcal{P}}}{\mathbb{E}} \langle \bar{w}, \xi_w \rangle \|v\|_2 w + b_1 \underset{\tilde{\mathcal{P}}}{\mathbb{E}} \|w\|_2 \|v\|_2 \xi_w$$

We have the following claim for the error in the first order gradients.

Claim B.2 (Error of 1st order gradient) *In the setting of Theorem 1, the following holds for any distribution $\tilde{\mathcal{P}}$ on w, ξ_w (v, ξ_v follows the same distribution).*

$$\langle \tilde{\nabla}_{1,v,1}, \xi_v \rangle = b_1 \left\| \underset{\tilde{\mathcal{P}}}{\mathbb{E}} \|w\|_2 \xi_w + w\langle \bar{w}, \xi_w \rangle \right\|_F^2 \geq 0$$

53

B.2. Stage 1.2: Analysis of Higher Order Tensor Decompositions

In this substage, we consider the error terms of the gradients of the higher order tensor decompositions. Towards showing the error propagation in Lemma 18, our proof outline is as follows.

- We decompose the error of the gradients into individual terms that we analyze one by one.

- In Proposition 21, we provide an upper bound on the average norm of the error. In Proposition 24, we bound the error of the individual terms from the decomposition.

- Finally, we present the proof of Lemma 18 in Section B.2.3.

We begin by writing down the gradient of higher order terms for the population loss..

$$
\tilde{\nabla}_{2j,v} = \left(b_{2j} + b'_{2j}\right) \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} \langle \tilde{w}, \tilde{v} \rangle \langle \tilde{\bar{w}}, \tilde{\bar{v}} \rangle^{2j-2} \tilde{w} - \sum_i a_i \langle e_i, \tilde{v} \rangle \langle e_i, \tilde{\bar{v}} \rangle^{2j-2} e_i \right)
$$
$$
- b'_{2j} \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}} \langle \tilde{w}, \tilde{v} \rangle \langle \tilde{\bar{w}}, \tilde{\bar{v}} \rangle^{2j-2} \langle \tilde{w}, \tilde{\bar{v}} \rangle - \sum_i a_i \langle e_i, \tilde{v} \rangle \langle e_i, \tilde{\bar{v}} \rangle^{2j-1} \right) \tilde{\bar{v}}.
$$

A crucial result is a bound on the average norm of the error. Let us define $\tilde{\Delta} := \left| \mathbb{E}_{w \sim \tilde{\mathcal{P}}} \|w\|_2^2 - 1 \right|$ and $\tilde{\delta} = \left\| \mathbb{E}_{w \sim \tilde{\mathcal{P}}} ww^\top - A \right\|_2$. We have the following result.

Proposition 21 (Average error bound) *In the setting of Lemma 18, suppose the running hypothesis $\tilde{\mathcal{H}}_0$ holds for every $t \in [T_2]$. In addition,*

- *For every neuron w, it holds that $\|\xi_w\|_2 \le \frac{1}{d^{20}} \|w\|_2$;*

- *$\left| \mathbb{E}_{w \sim \tilde{\mathcal{P}}}[w\|w\|_2] \right| \le \frac{1}{d^{40}}$ and $\| \mathbb{E}_{w \sim \tilde{\mathcal{P}}} ww^\top \|_2 \le 1$.*

As long as for every $w \in \mathcal{S}$ (recalling its definition in Def. 6), $\|\xi_w\|_2 \le \frac{1}{\text{poly}(d)}$, then we have

$$
\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t+1)}, w \in \mathcal{S}} \|\xi_w\|_2^2 \le \left(1 + \eta \frac{c_t \, \text{poly}(\kappa)}{d^2} \right) \mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)}, w \in \mathcal{S}} \|\xi_w\|_2^2
$$
$$
+ \eta O\left(\frac{1}{\lambda_0} \right) \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)}, w \notin \mathcal{S}} \|\xi_w\|_2 \right) \left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)}, w \in \mathcal{S}} \|\xi_w\|_2 \right)
$$
$$
+ \eta O\left(\max\left\{ \tilde{\Delta}^{(t)}, \tilde{\delta}^{(t)} \right\} \right) \mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)}, w \in \mathcal{S}} \|\xi_w\|_2^2.
$$

Next we show that the norm of the error in each individual neuron can also be bounded.

Proposition 22 (Individual error bound) *In the setting of Lemma 18, suppose that for every $t \in [T_2]$, we have $\mathbb{E}_{w, \xi_w \sim \tilde{\mathcal{P}}_t} \|\xi_w\|_2^2 \le \Xi$. Then for every $v \in \mathcal{S}_g$ and every $t \in [T_2]$:*

$$
\|\xi_v^{(t)}\|_2^2 \le \frac{\text{poly}(d)}{\lambda_0} \left(\|\xi_v^{(0)}\|_2^2 + \Xi \right).
$$

The proof of Proposition 21 and Proposition 22 is left to Section B.2.3.

B.2.1. DECOMPOSITION OF GRADIENT TERMS

We focus on the leading term that contains the first order term in ξ. We decompose $\tilde{\nabla}_{2j,v}$ into the following terms. In particular, these include terms from $\overline{v + \xi_v} = \bar{v} + \frac{\xi_v}{\|v\|_2} - \frac{\langle \xi_v, \bar{v} \rangle \bar{v}}{\|v\|_2} + O\left(\left(\frac{\|\xi_v\|_2}{\|v\|_2} \right)^2 \right)$.

$$\tilde{\nabla}_{2j,v,1} := (2j-1)\left(b_{2j} + b'_{2j}\right)\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle \xi_w, v \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} w + \underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle w, \xi_v \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} w \right),$$

$$\tilde{\nabla}_{2j,v,2} := \left(b_{2j} + b'_{2j}\right) \underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle w, v \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \xi_w,$$

$$\tilde{\nabla}_{2j,v,3} := -(2j-2)\left(b_{2j} + b'_{2j}\right)\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle \bar{w}, v \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle \xi_w, \bar{w} \rangle w \right),$$

$$\tilde{\nabla}_{2j,v,4} := -(2j-2)\left(b_{2j} + b'_{2j}\right)\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle \xi_v, \bar{v} \rangle w \right),$$

$$\tilde{\nabla}_{2j,v,5} := -(2j)b'_{2j}\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle \xi_w, v \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle w, \bar{v} \rangle \bar{v} \right), \tilde{\nabla}_{2j,v,6} := -(2j)b'_{2j}\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle \xi_v, w \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle w, \bar{v} \rangle \bar{v} \right),$$

$$\tilde{\nabla}_{2j,v,7} := (2j-2)b'_{2j}\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle \xi_w, \bar{w} \rangle \langle \bar{w}, \bar{v} \rangle^{2j-1} \langle w, v \rangle \bar{v} \right),$$

$$\tilde{\nabla}_{2j,v,8} := (2j-1)b'_{2j}\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle \xi_v, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle w, \bar{v} \rangle^2 \bar{v} \right),$$

$$\tilde{\nabla}_{2j,v,9} := -b'_{2j}\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle w, \bar{v} \rangle \xi_v \right), \tilde{\nabla}_{2j,v,10} := b'_{2j}\left(\underset{\tilde{\mathcal{P}}}{\mathbb{E}}\langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle w, \bar{v} \rangle \langle \xi_v, \bar{v} \rangle \bar{v} \right),$$

$$\tilde{\nabla}_{2j,v,11} := -(2j-1)\left(b_{2j} + b'_{2j}\right)\left(\sum_i a_i \langle e_i, \xi_v \rangle \langle e_i, \bar{v} \rangle^{2j-2} e_i \right),$$

$$\tilde{\nabla}_{2j,v,12} := (2j-2)\left(b_{2j} + b'_{2j}\right)\left(\sum_i a_i \langle \xi_v, \bar{v} \rangle \langle e_i, \bar{v} \rangle^{2j-1} e_i \right),$$

$$\tilde{\nabla}_{2j,v,13} := (2j)b'_{2j}\left(\sum_i a_i \langle e_i, \xi_v \rangle \langle e_i, \bar{v} \rangle^{2j-1} \right) \bar{v}, \tilde{\nabla}_{2j,v,14} := -(2j-1)b'_{2j}\left(\sum_i a_i \langle \xi_v, \bar{v} \rangle \langle e_i, \bar{v} \rangle^{2j} \right) \bar{v},$$

$$\tilde{\nabla}_{2j,v,15} := b'_{2j}\left(\sum_i a_i \langle e_i, \bar{v} \rangle^{2j} \right) \xi_v, \tilde{\nabla}_{2j,v,16} := -b'_{2j}\left(\sum_i a_i \langle e_i, \bar{v} \rangle^{2j} \right) \langle \xi_v, \bar{v} \rangle \bar{v}.$$

In addition, we show that the second order terms in ξ that contains $\|\xi_w\|_2^p$ and $\|\xi_v\|_2^q$ for $p + q \geq 3$ are of a lower order compared to the first order terms. Informally, we know that $\|\xi_w\|_2$ and $\|\xi_v\|_2$ are less than λ_0^2. Meanwhile, $\|w\|_2$ and $v\|_2$ are at least $\Omega\left(\frac{1}{d}\right)$, for every $w, v \in \mathcal{S}_g$ by Lemma A.3. Combined together, we show the following result.

Proposition 23 *In the setting of Proposition 21, let (w, ξ_w) be a random sample of $\tilde{\mathcal{P}}$.*

$$\|\tilde{\nabla}_{\leq 2, v} - \tilde{\nabla}_{0,v,1} - \tilde{\nabla}_{2,v,1}\|_2 + \|\tilde{\nabla}_{1,v,1} - \tilde{\nabla}_{1,v}\|_2 + \sum_{j \geq 0} \left\| \tilde{\nabla}_{2j,v} - \sum_p \tilde{\nabla}_{2j,v,p} \right\|_2$$

$$\leq O\left(\frac{1}{d^{10}} \left(\|v\|_2 \sqrt{ \underset{\xi_w \sim \tilde{\mathcal{P}}}{\mathbb{E}} \|\xi_w\|_2^2 } + \|\xi_v\|_2 \right) \right).$$

B.2.2. INDIVIDUAL ERROR NORM BOUND

Based on the decomposition above, we provide several helper claims for bounding the error of the gradient terms. First, for $v \in \mathcal{S}$, we have the following claim.

Claim B.3 *In the setting of Proposition 21, we have that*

$$\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}}} \left(\langle\xi_w, v\rangle\langle\bar{w}, \bar{v}\rangle^{2j-2}\langle w, \xi_v\rangle + \langle w, \xi_v\rangle\langle\bar{w}, \bar{v}\rangle^{2j-2}\langle w, \xi_v\rangle\right) \geq 0, \text{ and}$$

$$\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}}} \langle w, v\rangle\langle\bar{w}, \bar{v}\rangle^{2j-2}\langle\xi_w, \xi_v\rangle \geq 0.$$

This implies that for $p = 1, 2$:

$$\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}}} \langle\xi_v, \tilde{\nabla}_{2j,v,p}\rangle \geq 0.$$

Proof For the first inequality, we know that

$$\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}}} \left(\langle\xi_w, v\rangle\langle\bar{w}, \bar{v}\rangle^{2j-2}\langle w, \xi_v\rangle + \langle w, \xi_v\rangle\langle\bar{w}, \bar{v}\rangle^{2j-2}\langle w, \xi_v\rangle\right)$$

$$= \mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}}} \left(\langle\bar{w}, \bar{v}\rangle^{2j-2}\left(\langle\xi_w, v\rangle\langle w, \xi_v\rangle + \langle w, \xi_v\rangle^2\right)\right)$$

$$= \mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}}} \left(\langle\bar{w}, \bar{v}\rangle^{2j-2}\left(\langle\xi_w, v\rangle\langle w, \xi_v\rangle + \frac{1}{2}\langle w, \xi_v\rangle^2 + \frac{1}{2}\langle v, \xi_w\rangle^2\right)\right)$$

$$= \frac{1}{2}\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}}} \left(\langle\bar{w}, \bar{v}\rangle^{2j-2}(\langle w, \xi_v\rangle + \langle v, \xi_w\rangle)^2\right) \geq 0.$$

The second inequality in the Lemma follows from the fact that $(\langle w, v\rangle)_{w,v}, (\bar{w}, \bar{v}\rangle)_{w,v}, (\langle\xi_w, \xi_v\rangle)_{w,v}$ forms PSD matrices, and the Hadamard product of PSD matrices is PSD. ∎

We also have the following claim, which serves as an upper bound of

$$\tilde{\nabla}_{2j,v,3}, \tilde{\nabla}_{2j,v,4}, \tilde{\nabla}_{2j,v,5}, \tilde{\nabla}_{2j,v,6}, \tilde{\nabla}_{2j,v,7}, \tilde{\nabla}_{2j,v,8}, \tilde{\nabla}_{2j,v,9}, \tilde{\nabla}_{2j,v,10}.$$

Claim B.4 *In the setting of Proposition 21, we have that*

$$\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}},w,v\in\mathcal{S}} \left[\langle\bar{w}, \bar{v}\rangle^2|\langle\xi_w, v\rangle\langle\xi_v, w\rangle|\right] \leq \frac{c_t\kappa}{d^2}\mathbb{E}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\in\mathcal{S}} \|\xi_w\|_2^2, \text{ and}$$

$$\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}},w,v\in\mathcal{S}} \left[\langle\bar{w}, \bar{v}\rangle^2\langle\xi_w, v\rangle^2\right] \leq \frac{c_t\kappa}{d^2}\mathbb{E}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\in\mathcal{S}} \|\xi_w\|_2^2, \text{ and}$$

$$\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}},w,v\in\mathcal{S}} \left[\langle\bar{w}, \bar{v}\rangle^2\langle w, \bar{v}\rangle^2\right] \leq \frac{c_t\kappa}{d^2}.$$

As a corollary, combine the above inequality with Proposition 7, we obtain

$$\sum_{j\geq 2}\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},v\in\mathcal{S}} |\langle\xi_v, \tilde{\nabla}_{2j,v,p}\rangle| = O\left(\frac{c_t\kappa}{d^2}\mathbb{E}_{(v,\xi_v)\sim\tilde{\mathcal{P}},v\in\mathcal{S}} \|\xi_v\|_2^2 + \frac{1}{\lambda_0}\mathbb{E}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\notin\mathcal{S}} \|\xi_w\|_2^2\right),$$

where $p = 3, 4, 5, 6, 7, 8, 9, 10$.

Proof The proof is a direct calculation, using $\langle \bar{w}, \bar{v} \rangle^2 \leq \frac{c_t}{d}$ for $w \in \mathcal{S}$, we have that

$$\underset{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}},w,v\in\mathcal{S}}{\mathbb{E}} \left[\langle \bar{w}, \bar{v} \rangle^2 |\langle \xi_w, v \rangle \langle \xi_v, w \rangle| \right]$$

$$\leq \frac{c_t}{d} \underset{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}},w,v\in\mathcal{S}}{\mathbb{E}} |\langle \xi_w, v \rangle \langle \xi_v, w \rangle|$$

$$\leq \frac{c_t}{d} \underset{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}},w,v\in\mathcal{S}}{\mathbb{E}} |\langle \xi_w, v \rangle \langle \xi_v, w \rangle|$$

$$\leq \frac{c_t}{d} \underset{(v,\xi_v)\sim\tilde{\mathcal{P}},(w,\xi_w)\sim\tilde{\mathcal{P}},w,v\in\mathcal{S}}{\mathbb{E}} \left(\langle \xi_v, w \rangle^2 + \langle \xi_w, v \rangle^2 \right).$$

Now, we can easily calculate that (using the Eq (17))

$$\underset{(v,\xi_v)\sim\tilde{\mathcal{P}},v\in\mathcal{S}}{\mathbb{E}} \langle \xi_w, v \rangle^2 \leq \underset{(v,\xi_v)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \langle \xi_w, v \rangle^2 \leq \frac{2\kappa}{d} \|\xi_w\|_2^2,$$

which completes the proof. For the other two inequalities, we can bound them in the exact same way. ∎

The final claim aims to bound the rest of the terms.

Claim B.5 *In the setting of Proposition 21, we have that*

$$\left(\sum_{i=1}^{d} a_i \langle e_i, \bar{v} \rangle^2 \langle e_i, \xi_v \rangle^2 \right) = O\left(\frac{c_t \kappa}{d^2} \|\xi_v\|_2^2 \right) \text{ and } \left(\sum_{i=1}^{d} a_i \langle e_i, \bar{v} \rangle^2 \right) = O\left(\frac{c_t \kappa}{d^2} \right).$$

As a corollary, combining the above inequality with Proposition 7, we obtain

$$\sum_{j \geq 2} \underset{(v,\xi_v)\sim\tilde{\mathcal{P}},v\in\mathcal{S}}{\mathbb{E}} |\langle \xi_v, \tilde{\nabla}_{2j,v,p} \rangle| = O\left(\frac{c_t \kappa}{d^2} \underset{(v,\xi_v)\sim\tilde{\mathcal{P}},v\in\mathcal{S}}{\mathbb{E}} \|\xi_v\|_2^2 \right),$$

where $p = 11, 12, 13, 14, 15, 16$.

Individual Error Norm Bound for $v \in \mathcal{S}$. Below we also consider the error individually, we will mainly focus on the error term with ξ_v.

Claim B.6 *In the setting of Proposition 21, we have that*

$$\underset{(w,\xi_w)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \langle w, \xi_v \rangle^2 \langle \bar{w}, \bar{v} \rangle^2 \leq O\left(\frac{\kappa c_t}{d^2} \|\xi_v\|_2^2 \right), \text{ and}$$

$$\left| \underset{(w,\xi_w)\sim\tilde{\mathcal{P}}}{\mathbb{E}} \langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^2 \langle \xi_v, \bar{v} \rangle \langle w, \xi_v \rangle \right| \leq O\left(\frac{\kappa c_t}{d^2} \|\xi_v\|_2^2 \right).$$

.

Proof The first inequality is almost trivial. To see the second one, using $\mathbb{E}_{w \in \mathcal{S}, w \sim \tilde{\mathcal{P}}}[\langle \bar{w}, \bar{v} \rangle^2] \leq \frac{c_t}{d}$, we have that

$$
\left| \underset{(w, \xi_w) \sim \tilde{\mathcal{P}}, w \in \mathcal{S}}{\mathbb{E}} \langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^2 \langle \xi_v, \bar{v} \rangle \langle w, \xi_v \rangle \right|
$$

$$
\leq \frac{c_t}{d} \underset{(w, \xi_w) \sim \tilde{\mathcal{P}}}{\mathbb{E}} |\langle w, \bar{v} \rangle \langle w, \xi_v \rangle| \, \|\xi_v\|_2
$$

$$
\leq \frac{c_t}{d} \|\xi_v\|^2 \underset{(w, \xi_w) \sim \tilde{\mathcal{P}}}{\mathbb{E}} \left(\langle w, \bar{v} \rangle^2 + \frac{\langle w, \xi_v \rangle^2}{\|\xi_v\|_2^2} \right)
$$

$$
\leq O\left(\frac{\kappa c_t}{d^2} \|\xi_v\|_2^2 \right).
$$

For $w \notin \mathcal{S}$, we can naively bound $|\langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^2 \langle \xi_v, \bar{v} \rangle \langle w, \xi_v \rangle| \leq \|w\|_2^2 \|\xi_v\|_2^2$. Hence, using Eq (19), we have:

$$
\left| \underset{(w, \xi_w) \sim \tilde{\mathcal{P}}, w \notin \mathcal{S}}{\mathbb{E}} \langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^2 \langle \xi_v, \bar{v} \rangle \langle w, \xi_v \rangle \right| \leq \Lambda \|\xi_v\|_2^2 = \frac{1}{\text{poly}(d)} \|\xi_v\|_2^2.
$$

■

This claim together with Claim B.5 implies that

Claim B.7 (Error bound, $v \in \mathcal{S}$) *In the setting of Proposition 21, for every $v \in \mathcal{S}$, we have that*

$$
\|\xi_v^{(t+1)}\|_2^2 \leq \left(1 + \eta O\left(\max\left\{ \tilde{\Delta}^{(t)}, \tilde{\delta}^{(t)} \right\} \right) + \eta \frac{c_t \, \text{poly}(\kappa)}{d^2} \right) \|\xi_v^{(t)}\|_2^2 + O\left(\frac{d^2}{\lambda_0} \right) \left(\underset{\tilde{\mathcal{P}}^{(t)}}{\mathbb{E}} \|\xi_w\|_2^2 \right).
$$

Individual error bound for all the other neurons. Now we move on to the harder terms, we have the following claim.

Claim B.8 *In the setting of Proposition 21, for every $v \in \mathcal{S}$, we have that for $p = 11, 13$:*

$$
\sum_{j \geq 2} |\langle \xi_v, \nabla_{2j,v,p} + \nabla_{2j,v,p+1} \rangle| = O\left(\frac{\kappa}{d} \|\bar{v}\|_\infty^2 \|\xi_v\|_2^2 \right), \text{ and}
$$

$$
\langle \tilde{\nabla}_{2j,v,15} + \tilde{\nabla}_{2j,v,16}, \xi_v \rangle \geq 0.
$$

Proof We first consider $p = 11$. Let $Q'_{2j,v,11} = -\left(b_{2j} + b'_{2j} \right) \sum_i \left(a_i \langle e_i, \bar{v} \rangle^{2j-2} e_i e_i^\top \right)$. We have that

$$
\tilde{\nabla}_{2j,v,11} + \tilde{\nabla}_{2j,v,12}
$$

$$
= - (2j-1) \left(b_{2j} + b'_{2j} \right) \left(\sum_i a_i \langle e_i, \xi_v \rangle \langle e_i, \bar{v} \rangle^{2j-2} e_i \right)
$$

$$
+ (2j-2) \left(b_{2j} + b'_{2j} \right) \left(\sum_i a_i \langle \xi_v, \bar{v} \rangle \langle e_i, \bar{v} \rangle^{2j-1} e_i \right)
$$

$$
= Q'_{2j,v,11} \xi_v - (2j-1) \left(b_{2j} + b'_{2j} \right) \left(\sum_i a_i \langle e_i, \xi_v \rangle \langle e_i, \bar{v} \rangle^{2j-2} e_i - \sum_i a_i \langle \xi_v, \bar{v} \rangle \langle e_i, \bar{v} \rangle^{2j-1} e_i \right).
$$

Let us assume that $\|\bar{v}\|_\infty = 1 - \delta$ for some value $\delta \geq 0$, then we have that $\|\bar{v} - e_r\|_2^2 = O(\delta)$.

$$
\left| \sum_{i \in [d]} \left(a_i \langle \xi_v, e_i \rangle^2 \langle e_i, \bar{v} \rangle^2 - a_i \langle e_i, \bar{v} \rangle^3 \langle e_i, \xi_v \rangle \langle \xi_v, \bar{v} \rangle \right) \right|
$$

$$
\leq \left| \sum_{i \in [d], i \neq r} \left(a_i \langle \xi_v, e_i \rangle^2 \langle e_i, \bar{v} \rangle^2 - a_i \langle e_i, \bar{v} \rangle^3 \langle e_i, \xi_v \rangle \langle \xi_v, \bar{v} \rangle \right) \right|
$$

$$
+ \frac{\kappa}{d} \left| \langle \xi_v, e_r \rangle^2 \langle e_r, \bar{v} \rangle^2 - \langle e_r, \bar{v} \rangle^3 \langle e_r, \xi_v \rangle \langle \xi_v, \bar{v} \rangle \right|
$$

$$
\leq O\left(\frac{\kappa}{d}(1-\delta)^2 \sqrt{\delta} \|\xi_v\|_2^2 \right).
$$

Using the fact that $b_{2j}, b'_{2j} = \Theta(\frac{1}{j^2})$, we know that

$$
\sum_{j \geq 2} (2j-1) \left(b_{2j} + b'_{2j} \right) \left| \sum_i a_i \langle e_i, \xi_v \rangle \langle e_i, \bar{v} \rangle^{2j-2} \langle e_i, \xi_v \rangle - \sum_i a_i \langle \xi_v, \bar{v} \rangle \langle e_i, \bar{v} \rangle^{2j-1} \langle e_i, \xi_v \rangle \right|
$$

$$
\leq \sum_{j \geq 2} O\left(\frac{1}{j}(1-\delta)^j \sqrt{\delta} \frac{\kappa}{d} \|\xi_v\|_2^2 \right).
$$

Note that $\sum_{j \geq 2} \frac{1}{j}(1-\delta)^j = (1-\delta) \log \frac{1}{\delta}$ we obtain:

$$
\sum_{j \geq 2} \left| \langle \tilde{\nabla}_{2j,v,11} + \tilde{\nabla}_{2j,v,12} - Q'_{2j,v,11} \xi_v, \xi_v \rangle \right| \leq O\left(\frac{\kappa}{d}(1-\delta)^2 \log \frac{1}{\delta} \sqrt{\delta} \|\xi_v\|_2^2 \right) = O\left(\frac{\kappa}{d} \|\xi_v\|_2^2 \right).
$$

Similarly, we can also show that

$$
\sum_{j \geq 2} \|Q'_{2j,v,11}\|_2 = \sum_{j \geq 2} \left(b_{2j} + b'_{2j} \right) \left\| \sum_i \left(a_i \langle e_i, \bar{v} \rangle^{2j-2} e_i e_i^\top \right) \right\|_2
$$

$$
\leq O\left(\frac{\kappa}{d} \|\bar{v}\|_\infty^2 \right),
$$

which completes the proof. On the other hand, for $p = 13$, let $Q'_{2j,v,13} = b'_{2j} \left(\sum_i a_i \langle e_i, \bar{v} \rangle^{2j-1} \bar{v} e_i^\top \right)$ ..
We have that

$$
\tilde{\nabla}_{2j,v,13} + \tilde{\nabla}_{2j,v,14}
$$

$$
= (2j) b'_{2j} \left(\sum_i a_i \langle e_i, \xi_v \rangle \langle e_i, \bar{v} \rangle^{2j-1} \right) \bar{v} - (2j-1) b'_{2j} \left(\sum_i a_i \langle \xi_v, \bar{v} \rangle \langle e_i, \bar{v} \rangle^{2j} \right) \bar{v}
$$

$$
= Q'_{2j,v,13} \xi_v + (2j-1) b'_{2j} \left(\sum_i a_i \langle e_i, \xi_v \rangle \langle e_i, \bar{v} \rangle^{2j-1} - \sum_i a_i \langle \xi_v, \bar{v} \rangle \langle e_i, \bar{v} \rangle^{2j} \right) \bar{v},
$$

We can bound the terms in a similar way. ∎

Using the aforementioned claims, we conclude the proof of the following proposition.

Proposition 24 (Individual error bound) *In the setting of Proposition 21, for every v, we have that*

1. *For $p = 4, 6, 8, 9, 10$, using Claim B.6, we have*

$$\sum_{j \geq 2} |\langle \xi_v, \tilde{\nabla}_{2j,v,p}\rangle| \leq O\left(\frac{c_t \operatorname{poly}(\kappa)}{d^2}\|\xi_v\|_2^2\right)$$

2. *When $\|w\|_2, \|\xi_w\|_2 \leq \frac{1}{\lambda_0}$, for $p = 2, 3, 5, 7$, the following is true*

$$\sum_{j \geq 2} |\langle \xi_v, \tilde{\nabla}_{2j,v,p}\rangle| \leq O\left(\frac{1}{\lambda_0}\right) \left(\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}}} \|\xi_w\|_2\right) \|\xi_v\|_2$$

$$\leq O\left(\frac{d^2}{\lambda_0}\right) \left(\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}}} \|\xi_w\|_2^2\right) + \frac{1}{d^2}\|\xi_v\|_2^2$$

3. *For $p = 1$, similarly we have:*

$$\sum_{j \geq 2} |\langle \xi_v, \tilde{\nabla}_{2j,v,p}\rangle| \leq O\left(\frac{c_t \kappa}{d^2}\|\xi_v\|_2^2 + \frac{1}{\lambda_0}\left(\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}}} \|\xi_w\|_2^2\right)\right)$$

4. *For $p = 15$:*

$$\sum_{j \geq 2} \langle \xi_v, \tilde{\nabla}_{2j,v,p} + \tilde{\nabla}_{2j,v,p+1}\rangle \geq 0$$

5. *For $p = 11, 12, 13, 14$, we have that for $p = 11, 13$, using Claim B.8, we get*

$$\sum_{j \geq 2} |\langle \xi_v, \nabla_{2j,v,p} + \nabla_{2j,v,p+1}\rangle| = O\left(\frac{\kappa}{d}\|\bar{v}\|_\infty^2\|\xi_v\|_2^2\right)$$

B.2.3. PROOF OF ERROR PROPAGATION

Based on the individual error norm bound and the average error norm bound, we are ready to prove the main result of stage 1. We first state the proof of the individual error norm bound.

Proof [Proof of Proposition 22] We consider the error caused by gradient clipping, since we might clip v and $\tilde{v}$ at different time step. We have that at the iteration t when $\|v^{(t)}\|_2^2 \geq \frac{1}{2\lambda_0}$ or $\|v^{(t)} + \xi_v^{(t)}\|_2^2 \geq \frac{1}{2\lambda_0}$, we have that

$$\|v^{(t)} + \xi_v^{(t)}\|_2^2, \|v^{(t)}\|_2^2 \geq \frac{1}{2\lambda_0} - \frac{2}{\sqrt{\lambda_0}}\|\xi_v^{(t)}\|_2.$$

On the other hand, by Eq (48), we have that for this v, if the gradient clipping is not performed, then by the definition of $\mathcal{S}_g$, we have that $\|\bar{v}^{(t)}\|_2 \geq \frac{1}{\log d}$. Therefore,

$$\|v^{(t+1)}\|_2^2 \geq \|v^{(t)}\|_2^2 \left(1 + \eta\Omega\left(\frac{1}{d\log^3 d}\right)\right),$$

which implies that after $t' = O\left(\frac{\sqrt{\lambda_0}d\log^3 d\|\xi_v^{(t)}\|_2}{\eta}\right)$, many iterations, if gradient clipping is not performed, we should have that

$$\|v^{(t+t')} + \xi_v^{(t+t')}\|_2^2, \|v^{(t+t')}\|_2^2 \geq \frac{1}{2\lambda_0}.$$

Since each iteration shall introduce at most $O\left(\eta\frac{1}{\lambda_0}\right)$ amount error, so we have:

$$\|\xi_v^{(t+t')}\|_2 \leq O\left(\frac{d\log^3 d\|\xi_v^{(t)}\|_2}{\sqrt{\lambda_0}}\right).$$

This gives us the final error bound of the individual error when combined with Claim B.5. ∎

Next we state the proof of the average error norm bound.

Proof [Proof of Proposition 21] Using Proposition 24 (together with Eq (78)) and by the definition of Eq (77), we can obtain the desired result. ∎

Based on Proposition 21, Proposition 22, and Proposition 24, we are ready to prove Lemma 18.
Proof [Proof of Lemma 18] Clearly, when $m \geq \frac{\mathrm{poly}(d)}{\mathrm{poly}(\lambda_0)}$, then the running hypothesis $\tilde{\mathcal{H}}_0$ is satisfied for every $t \leq T_2$. To prove this Lemma, we shall maintain using induction that at every iteration $t \in [T_2]$,

$$\mathop{\mathbb{E}}_{w,\xi_w \sim \tilde{\mathcal{P}}^{(t)}} \|\xi_w\|_2^2 \leq \mathrm{poly}(d)\Xi,$$

and for every neuron v, $\|\xi_v^{(t)}\|_2^2 \leq \frac{\mathrm{poly}(d)}{\lambda_0}\Xi$. Suppose this is true for all $t \leq T_0$, then consider $t = T_0 + 1$. We apply Proposition 21, which says that as long as for every $w \in \mathcal{S}$, $\|\xi_w\|_2 \leq \frac{1}{d^3}$, we have that

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t+1)},w\in\mathcal{S}} \|\xi_w\|_2^2 \leq \left(1 + \eta\frac{c_t\,\mathrm{poly}(\kappa)}{d^2}\right) \mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2^2$$
$$+ \eta O\left(\frac{1}{\lambda_0}\right)\left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)},w\notin\mathcal{S}} \|\xi_w\|_2\right)\left(\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2\right)$$
$$+ \eta O\left(\max\left\{|\tilde{\Delta}^{(t)}|, \tilde{\delta}^{(t)}\right\}\right)\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2^2 + \eta \mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2\|\Xi_w\|_2.$$

Combining this with Proposition 22 and $\mathbb{E}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2\|\Xi_w\|_2 \leq \frac{1}{d^2}\mathbb{E}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2 + d^2\Xi$, we have that

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t+1)},w\in\mathcal{S}} \|\xi_w\|_2^2 \leq \left(1 + \eta\frac{c_t\,\mathrm{poly}(\kappa)}{d^2}\right) \mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2^2 + \eta O\left(\frac{\mathrm{poly}(d)}{\lambda_0^2}\right)\mu(\mathcal{S})\,\mathrm{poly}(\kappa_2)\Xi$$
$$+ \eta O\left(\max\left\{|\tilde{\Delta}^{(t)}|, \delta^{(t)}\right\}\right)\mathop{\mathbb{E}}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2^2 + \eta d^2\Xi.$$

Hence, denote $\varepsilon_t = \mathbb{E}_{\tilde{\mathcal{P}}^{(t)},w\in\mathcal{S}} \|\xi_w\|_2^2$, we have that for every $s \leq t$:

$$\varepsilon_{s+1} \leq \varepsilon_s \left(1 + \eta\frac{c_t\,\mathrm{poly}(\kappa)}{d^2} + \eta O\left(\max\left\{|\Delta^{(t)}|, \delta_+^{(t)}, \delta_-^{(t)}\right\}\right)\right) + \eta d^2\Xi.$$

By $m \geq \frac{\text{poly}(d)}{\text{poly}(\lambda_0)}$, a simple Chernoff bound gives us:

$$\max\left\{|\tilde{\Delta}^{(t)}|, \tilde{\delta}^{(t)}\right\} \leq O\left(\max\left\{|\Delta^{(t)}|, \delta_+^{(t)}, \delta_-^{(t)}\right\}\right) + \frac{1}{\text{poly}(d)}.$$

Now, using the update rule of Eq (32) and in Proposition 15, we have that

$$\sum_{t \leq T_2} \left(\eta \frac{c_t \text{ poly}(\kappa)}{d^2} + \eta \max\left\{|\Delta^{(t)}|, \delta_+^{(t)}, \delta_-^{(t)}\right\}\right) \leq \text{poly}(\kappa_1)$$

Note that at iteration 0, $\varepsilon_0 = 0$. This implies that for $t + 1$: $\varepsilon_{t+1} \leq \text{poly}(d)\Xi$ as well. Combine this with Proposition 24 on the individual norm bound we complete the proof. ∎

B.3. Stage 2.1: Analysis After Reducing the Gradient Truncation Parameter

In this section, we prove Lemma 19, which analyzes the error propagation of this substage. We analyze the formula of $\tilde{\nabla}_{2j,v,p}$ in Section B.2 and show the following claim.

Claim B.9 *In the setting of Lemma 19, let $\sigma_{\max} = \max\left\{\|\mathbb{E}_{w \sim \tilde{\mathcal{P}}} ww^\top\|_2, \frac{\kappa}{d}\right\}$. Then we have the following average error norm bound*

$$\sum_{j \geq 2} \sum_p \mathbb{E}_{(v,\xi_v) \sim \tilde{\mathcal{P}}} \langle \tilde{\nabla}_{2j,v,p}, \xi_v \rangle \geq -O\left(\sigma_{\max} \mathbb{E}_{\xi_w \sim \tilde{\mathcal{P}}} \|\xi_w\|_2^2\right).$$

For every individual neuron v and every value $\alpha \geq 1$, the following holds

$$\sum_{j \geq 2} \sum_p |\langle \tilde{\nabla}_{2j,v,p}, \xi_v \rangle| \leq O\left(\alpha\sigma_{\max}\|\xi_v\|_2^2 + \frac{1}{\alpha}\|v\|_2^2 \mathbb{E}_{\xi_w \sim \tilde{\mathcal{P}}} \|\xi_w\|_2^2\right).$$

Proof The proof of this claim is quite straightforward. We have that for $p = 1, 2$, we use Claim B.3, which gives us:

$$\mathbb{E}_{(v,\xi_v) \sim \tilde{\mathcal{P}}} \langle \xi_v, \tilde{\nabla}_{2j,v,p} \rangle \geq 0.$$

For $p = 3$, we use that $|\langle \bar{w}, \bar{v} \rangle| \leq 1$ and

$$\mathbb{E}_{(w,\xi_w),(v,\xi_v) \sim \tilde{\mathcal{P}}} |\langle \xi_w, \bar{w} \rangle||\langle \xi_v, w \rangle||\langle \bar{w}, v \rangle| \leq \frac{1}{2} \mathbb{E}_{(w,\xi_w),(v,\xi_v) \sim \tilde{\mathcal{P}}} \left(|\langle \xi_w, \bar{w} \rangle|^2 \langle w, v \rangle^2 + |\langle \xi_v, w \rangle|^2\right)$$

$$\leq \sigma_{\max} \mathbb{E}_{\xi_w \sim \tilde{\mathcal{P}}} \|\xi_w\|_2^2.$$

For $p = 4$, we use that

$$\mathbb{E}_{(w,\xi_w),(v,\xi_v) \sim \tilde{\mathcal{P}}} |\langle \xi_v, \bar{v} \rangle||\langle \xi_v, w \rangle||\langle \bar{v}, w \rangle| \leq \mathbb{E}_{(w,\xi_w),(v,\xi_v) \sim \tilde{\mathcal{P}}} \|\xi_v\|_2 |\xi_v^\top ww^\top \bar{v}|$$

$$\leq \sigma_{\max} \mathbb{E}_{\xi_w \sim \tilde{\mathcal{P}}} \|\xi_w\|_2^2.$$

For $p = 5$, we use that

$$\mathbb{E}_{(w,\xi_w),(v,\xi_v)\sim\tilde{\mathcal{P}}} |\langle \xi_w, v\rangle||\langle \xi_v, \bar{v}\rangle||\langle \bar{v}, w\rangle| \leq \frac{1}{2} \mathbb{E}_{(w,\xi_w),(v,\xi_v)\sim\tilde{\mathcal{P}}} \left(\langle \xi_w, v\rangle^2 + \langle \xi_v, \bar{v}\rangle^2\langle \bar{v}, w\rangle^2\right)$$

$$\leq \sigma_{\max} \mathbb{E}_{\xi_w\sim\tilde{\mathcal{P}}} \|\xi_w\|_2^2.$$

For $p = 6$, we use that

$$\mathbb{E}_{(w,\xi_w),(v,\xi_v)\sim\tilde{\mathcal{P}}} |\langle \xi_v, w\rangle||\langle \xi_v, \bar{v}\rangle||\langle \bar{v}, w\rangle| \leq \frac{1}{2} \mathbb{E}_{(w,\xi_w),(v,\xi_v)\sim\tilde{\mathcal{P}}} \left(\langle \xi_v, w\rangle^2 + \langle \xi_v, \bar{v}\rangle^2\langle \bar{v}, w\rangle^2\right)$$

$$\leq \sigma_{\max} \mathbb{E}_{\xi_w\sim\tilde{\mathcal{P}}} \|\xi_w\|_2^2.$$

For $p = 7$, we use that

$$\mathbb{E}_{(w,\xi_w),(v,\xi_v)\sim\tilde{\mathcal{P}}} |\langle \xi_w, \bar{w}\rangle||\langle \xi_v, \bar{v}\rangle||\langle v, w\rangle||\langle \bar{w}, \bar{v}\rangle|$$

$$= \mathbb{E}_{(w,\xi_w),(v,\xi_v)\sim\tilde{\mathcal{P}}} |\langle \xi_w, \bar{w}\rangle||\langle \xi_v, \bar{v}\rangle||\langle v, \bar{w}\rangle||\langle w, \bar{v}\rangle|$$

$$\leq \frac{1}{2} \mathbb{E}_{(w,\xi_w),(v,\xi_v)\sim\tilde{\mathcal{P}}} \left(\langle \xi_w, \bar{w}\rangle^2\langle v, \bar{w}\rangle^2 + \langle \xi_v, \bar{v}\rangle^2\langle \bar{v}, w\rangle^2\right)$$

$$\leq \sigma_{\max} \mathbb{E}_{\xi_w\sim\tilde{\mathcal{P}}} \|\xi_w\|_2^2.$$

For $p = 8, 9, 10, 11$, the result can be obtained similarly. For $p = 12, 13$, we use that

$$\sum_i a_i |\langle \xi_v, \bar{v}\rangle||\langle e_i, \bar{v}\rangle||\langle e_i, \xi_v\rangle|$$

$$\leq \frac{\kappa}{d}\|\xi_v\|_2 \left|\xi_v^\top \sum_i e_i e_i^\top \bar{v}\right| \leq \frac{\kappa}{d}\|\xi_v\|_2^2.$$

For $p = 14$, we use that

$$\sum_i a_i |\langle \xi_v, \bar{v}\rangle||\langle e_i, \bar{v}\rangle|^2|\langle \bar{v}, \xi_v\rangle|$$

$$\leq \|\xi_v\|_2^2 \sum_i a_i \langle e_i, \bar{v}\rangle^2 \leq \frac{\kappa}{d}\|\xi_v\|_2^2.$$

Finally, the individual error bound comes from the following simple calculation.

$$\sum_{j\geq 2} |\langle \tilde{\nabla}_{2j,v,p}, \xi_v\rangle| \leq O\left(\frac{\kappa}{d}\|\xi_v\|_2^2 + \mathbb{E}_{w,\xi_w\sim\tilde{\mathcal{P}}} \|w\|_2^2\|\xi_v\|_2^2 + \|w\|_2\|\xi_w\|_2\|v\|_2\|\xi_v\|_2\right)$$

$$\leq O\left(\alpha\sigma_{\max}\|\xi_v\|_2^2 + \frac{1}{\alpha}\|v\|_2^2 \mathbb{E}_{\xi_w\sim\tilde{\mathcal{P}}} \|\xi_w\|_2^2\right).$$

∎

Proof [Proof of Lemma 19] Note that this substage has T_3 many iterations, where T_3 is upper bounded by $\frac{dC(\kappa)\log d}{\eta}$ for some value $C(\kappa) > 0$ that only depends on κ. By by taking $\alpha = 1$ in Claim B.9, the rest of the proof is similar to the proof of Lemma 18. We omit the details. ∎

B.4. Stage 2.2: The Final Substage

We provide the proof of Lemma 20, which analyzes the error propagation in the final substage. Recall that $\mathcal{S}_{i,singleton}$ and $\mathcal{S}_{ignore}$ have been defined in the beginning of this section. At the beginning of Stage 2.2 when $t = T_3 + 1$, we do a modification:

1. If v in $\mathcal{S}_{i,singleton}$ we will just set $\bar{v} = e_i$ and keep the norm not changed.

2. If v in $\mathcal{S}_{ignore}$, then we will just set $v = 0$.

Thus, we can see that $v^{(t)} = 0$ for every $v \in \mathcal{S}_{ignore}$ and for every $t > T_3$. We define a new update for the infinite neuron process at this substage for $v \in \mathcal{S}_{i,singleton}$. We define v_+, v_- such that at every iteration $t \geq T_3$:

$$v_+^{(t)} = -v_-^{(t)} = \langle v^{(t+1)}, e_i \rangle.$$

We will replace v in the infinite neuron process with two neurons v_+, v_-. For the simplicity of notation, we write v_+ simply as v. For the other neurons, the update does not change.

We can see that this new initial state also satisfies the running hypothesis $\mathcal{H}_1$ and the conditional-symmetric property as well. Thus, the update in Claim A.10 still holds. We consider the new infinite neuron process starting from this initial state. We can see that when $v \in \mathcal{S}_{i,singleton}$, then $\bar{v}^{(t)} = e_i$ for every $t \geq T_3 + 1$. Moreover, when $v \in \mathcal{S}_{i,singleton}$, we define $\tilde{v} = v + \xi_v$ where $\langle \xi_v, e_i \rangle = 0$. Thus, we do not consider the scaling difference between the singleton neurons in the infinite-width case and the finite-width case as an error.

By the running hypothesis $\mathcal{H}_1$ in Proposition 10, we have that at iteration T_3,

$$\mathbb{E}_{w \sim \mathcal{P}^{(T_3)}, w \in \mathcal{S}_{ignore}} \|w\|_2^2 \leq \frac{1}{\text{poly}(d)}.$$

Therefore, the following is also true.

$$\mathbb{E}_{w \sim \mathcal{P}^{(T_3+1)}, w \in \mathcal{S}_{ignore}} \|\xi_w\|_2^2 \leq \frac{1}{\text{poly}(d)}.$$

Moreover, throughout the entire process, by Lemma 12, we will always have that

$$\mathbb{E}_{w \sim \tilde{\mathcal{P}}^{(t)}, w \in \mathcal{S}_{ignore}} \|\xi_w\|_2^2 \leq \frac{1}{\text{poly}(d)}.$$

Therefore, we only need to consider the error of $w \notin \mathcal{S}_{ignore}$. We denote the new running hypothesis $\tilde{\mathcal{H}}_1$ as for every $t \leq T_4$:

1. For every $v \in \mathcal{S}_g$, we have that:

$$\|v^{(t)}\|_2^2 \leq \frac{1}{2\lambda_1}.$$

2. For every $v \notin \mathcal{S}_{pot}$ (cf. Lemma 9 for the definition),

$$\|\bar{v}\|_\infty^2, \|v\|_\infty^2 \leq \frac{c_t}{d}.$$

64

3. The mass of the ignore set is small.

$$\mathop{\mathbb{E}}_{w\sim\tilde{\mathcal{P}},w\in\mathcal{S}_{ignore}} \|\xi\|_2^2 \leq \frac{1}{\text{poly}(d)}.$$

We first prove the following claim.

Claim B.10   *In the setting of Lemma 20, suppose that the distribution $\tilde{\mathcal{P}} = \tilde{\mathcal{P}}^{(t)}$ satisfies the running hypothesis $\tilde{\mathcal{H}}_1$. For every $i \in [d]$ and for $v \in \mathcal{S}_{i,singleton}$, the following holds:*

$$\sum_{j\geq 2}\langle \tilde{\nabla}_{2j,v}, \xi_v\rangle = \sum_{j\geq 2}\langle \tilde{\nabla}_{2j,v,1} + \tilde{\nabla}_{2j,v,2}, \xi_v\rangle$$

$$\pm O\left(|\tilde{\Phi}|\|\xi_v\|_2^2 + \frac{c_t\,\text{poly}(\kappa_2)}{d^2}\|\xi_v\|_2^2 + \frac{c_t\,\text{poly}(\kappa_2)}{d^2}\|v\|_2^2\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\notin\mathcal{S}_{pot}}\|\xi_w\|_2^2 + \frac{1}{\text{poly}(d)}\|\xi_v\|_2\right),$$

where Φ is defined as in Lemma 12 with $\tilde{\mathcal{P}}$ instead of $\mathcal{P}$.

Proof   We consider $v \in \mathcal{S}_{i,singleton}$. For these neurons, we have that

$$\langle \tilde{\nabla}_{2j,v,p}, \xi_v\rangle = 0.$$

Let

$$\nabla_{2j,v,w} := \left(b_{2j} + b'_{2j}\right)\left(\langle w, v\rangle\langle \bar{w}, \bar{v}\rangle^{2j-2}w\right) - b'_{2j}\left(\langle w, v\rangle\langle \bar{w}, \bar{v}\rangle^{2j-2}\langle w, \bar{v}\rangle\right)\bar{v} \qquad (79)$$

as the gradient of v involving only a single neuron w. For $p = 5, 6, 7, 8, 10, 13, 14, 16$. Now, for $p = 3$ we have that

$$\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}}}\langle \bar{w}, v\rangle\langle \bar{w}, \bar{v}\rangle^{2j-2}\langle \xi_w, \bar{w}\rangle\langle \bar{w}, \xi_v\rangle$$

$$= \mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\notin\mathcal{S}_{pot}}\langle \bar{w}, v\rangle\langle \bar{w}, \bar{v}\rangle^{2j-2}\langle \xi_w, \bar{w}\rangle\langle \bar{w}, \xi_v\rangle$$

$$+ \sum_{j\in[d]}\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\in\mathcal{S}_{j,singleton}}\langle \bar{w}, v\rangle\langle \bar{w}, \bar{v}\rangle^{2j-2}\langle \xi_w, \bar{w}\rangle\langle \bar{w}, \xi_v\rangle + \sum_{w\in\mathcal{S}_{ignore}}\langle \nabla_{2j,v,w}, \xi_v\rangle,$$

where $\nabla_{2j,v,w}$ is defined in Eq (79). For the first term, using the running hypothesis $\tilde{\mathcal{H}}_1$ that for every $w \notin \mathcal{S}_{pot}$, $\|\bar{w}\|_\infty^2 \leq \frac{c_t}{d}$, we have that $\langle \bar{w}, \bar{v}\rangle^2 \leq \frac{c_t}{d}$. This implies that

$$\sum_{j\geq 2}\left|\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\notin\mathcal{S}_{pot}}\langle \bar{w}, v\rangle\langle \bar{w}, \bar{v}\rangle^{2j-2}\langle \xi_w, \bar{w}\rangle\langle \bar{w}, \xi_v\rangle\right|$$

$$\leq \frac{c_t\,\text{poly}(\kappa_2)}{d^2}\|v\|_2\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\notin\mathcal{S}_{pot}}\|\xi_w\|_2\|\xi_v\|_2$$

$$\leq \frac{c_t\,\text{poly}(\kappa_2)}{d^2}\|\xi_v\|_2^2 + \|v\|_2^2\frac{c_t\,\text{poly}(\kappa_2)}{d^2}\mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}},w\notin\mathcal{S}_{pot}}\|\xi_w\|_2^2.$$

For the second term, we have that when $j = i$, we have for every $w \in \mathcal{S}_{j,singleton}$: $\langle \bar{w}, \xi_v \rangle = 0$. Otherwise, when $j \neq i$, we have that $\langle \bar{w}, \bar{v} \rangle = 0$. Therefore,

$$\sum_{j \in [d]} \mathop{\mathbb{E}}_{(w,\xi_w) \sim \tilde{\mathcal{P}}, w \in \mathcal{S}_{j,singleton}} \langle \bar{w}, v \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle \xi_w, \bar{w} \rangle \langle \bar{w}, \xi_v \rangle = 0.$$

For the third term, we have that

$$\left| \sum_{w \in \mathcal{S}_{ignore}} \langle \nabla_{2j,v,w}, \xi_v \rangle \right| \le \sum_{w \in \mathcal{S}_{ignore}} \|w\|_2^2 \|\xi_v\|_2 \le \frac{1}{\text{poly}(d)} \|\xi_v\|_2.$$

Hence we have that

$$\left| \langle \tilde{\nabla}_{2j,v,p}, \xi_v \rangle \right| \le \frac{c_t \text{poly}(\kappa_2)}{d^2} \|\xi_v\|_2^2 + \frac{c_t \text{poly}(\kappa_2)}{d^2} \mathop{\mathbb{E}}_{(w,\xi_w) \sim \tilde{\mathcal{P}}, w \notin \mathcal{S}_{pot}} \|\xi_w\|_2^2 + \frac{1}{\text{poly}(d)} \|\xi_v\|_2.$$

For $p = 4$, we also have:

$$\left| \mathop{\mathbb{E}}_{(w,\xi_w) \sim \tilde{\mathcal{P}}} \langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle \xi_v, \bar{v} \rangle \langle \bar{w}, \xi_v \rangle \right|$$

$$\le \left| \mathop{\mathbb{E}}_{(w,\xi_w) \sim \tilde{\mathcal{P}}, w \notin \mathcal{S}_{pot}} \langle w, \bar{v} \rangle \langle \bar{w}, \bar{v} \rangle^{2j-2} \langle \xi_v, \bar{v} \rangle \langle \bar{w}, \xi_v \rangle \right| + \frac{1}{\text{poly}(d)} \|\xi_v\|_2$$

$$\le \frac{c_t \text{poly}(\kappa_2)}{d^2} \|\xi_v\|_2^2 + \frac{c_t \text{poly}(\kappa_2)}{d^2} \mathop{\mathbb{E}}_{(w,\xi_w) \sim \tilde{\mathcal{P}}, w \notin \mathcal{S}_{pot}} \|\xi_w\|_2^2 + \frac{1}{\text{poly}(d)} \|\xi_v\|_2.$$

Now for $p = 11, 12$, we also know

$$\langle \tilde{\nabla}_{2j,v,p}, \xi_v \rangle = 0.$$

For $p = 9, 15$, following the same calculation by dividing w into three parts we can easily conclude that

$$\left| \langle \tilde{\nabla}_{2j,v,9} + \tilde{\nabla}_{2j,v,15} \rangle \right|$$

$$\le \left| b'_{2j}(a_i - \tilde{\gamma}_i) \right| \|\xi_v\|_2^2 + \frac{c_t \text{poly}(\kappa_2)}{d^2} \|\xi_v\|_2^2 + \frac{c_t \text{poly}(\kappa_2)}{d^2} \mathop{\mathbb{E}}_{(w,\xi_w) \sim \tilde{\mathcal{P}}, w \notin \mathcal{S}_{pot}} \|\xi_w\|_2^2 + \frac{1}{\text{poly}(d)} \|\xi_v\|_2,$$

where $\tilde{\gamma}_i = \mathbb{E}_{w \sim \tilde{\mathcal{P}}, w \in \mathcal{S}_{i,pot} \setminus \mathcal{S}_{bad}} w_i^2$. Therefore we finish the proof with Eq (74). $\blacksquare$

Next, we consider the error of the neurons not in $\mathcal{S}_{pot}$. We use a direct corollary of Claim B.9 , except that for every $v \notin \mathcal{S}_{pot}$, it holds that $\langle \bar{v}, \bar{w} \rangle^2 \le \frac{c_t}{d}$ instead of 1 for every vector w. We state the result as follows.

Claim B.11 *In the setting of Lemma 20, let $\sigma_{\max} = \max \left\{ \| \mathbb{E}_{w \sim \tilde{\mathcal{P}}} ww^\top \|_2, \frac{\kappa}{d} \right\}$. Let $\tilde{\mathcal{P}} = \tilde{\mathcal{P}}^{(t)}$ denote the distribution of the neurons. For every v such that $\|\bar{v}\|_\infty^2, \|v\|_\infty^2 \le \frac{c_t}{d}$ and any $\alpha \ge 1$, we have:*

$$\sum_{j \ge 2} \sum_p |\langle \tilde{\nabla}_{2j,v,p}, \xi_v \rangle| \le O \left(\alpha \frac{c_t}{d} \sigma_{\max} \|\xi_v\|_2^2 + \frac{1}{\alpha} \frac{c_t}{d} \|v\|_2^2 \mathop{\mathbb{E}}_{\xi_w \sim \tilde{\mathcal{P}}} \|\xi_w\|_2^2 \right).$$

Based on Claim B.10 and B.11, we prove Lemma 20.

Proof [Proof of Lemma 20]Let $\alpha = \sqrt{d}$ in Claim B.11, we show the following result the bound. For every $v \notin \mathcal{S}_{pot}$,

$$\sum_{j \geq 2} \langle \tilde{\nabla}_{2j,v}, \xi_v \rangle \geq -O\left(|\tilde{\Phi}| \|\xi_v\|_2^2 + \frac{c_t \operatorname{poly}(\kappa_2)}{d^{1.5}} \|\xi_v\|_2^2 + \frac{c_t \operatorname{poly}(\kappa_2)}{d^{1.5}} \|v\|_2^2 \mathop{\mathbb{E}}_{(w,\xi_w)\sim\tilde{\mathcal{P}}} \|\xi_w\|_2^2 + \frac{1}{\operatorname{poly}(d)} \|\xi_v\|_2 \right).$$

Together with the individual error bound as in Claim B.9, we can obtain the desired result using a similar proof to Lemma 18. The details are omitted. ∎

Appendix C. Proof of Lower Bound

We follow the proof of Theorem 2 in Allen-Zhu and Li (2019a) for proving the lower bound. We first describe the construction of the hardness distribution $\mathcal{W}$. We first show the following lemma.

Lemma 25 *For a positive integer r, for every $d \geq r^2$ which is a multiple of r, there exists at least $H = d^{\Omega(r)}$ many sets $\mathcal{C}^{(j)} = \{\mathcal{C}_1^{(j)} \in [d], \cdots, \mathcal{C}_{d/r}^{(j)} \in [d]\}$ for $j = 1, \ldots, Q$ such that*

1. For every $1 \leq i \leq d/r$ and $1 \leq j \leq H$, $\mathcal{C}_i^{(j)}$ is a subset of $[d]$ of size r.

2. For every $1 \leq i \neq i' \leq d/r$ and $1 \leq j \leq H$, $\mathcal{C}_i^{(j)} \cap \mathcal{C}_{i'}^{(j)} = \emptyset$.

3. For every $1 \leq i, i' \leq d/r$ and $1 \leq j \neq j' \leq H$, $\mathcal{C}_i^{(j)} \neq \mathcal{C}_{i'}^{(j')}$.

Proof We consider a uniformly at random distribution over the set $\mathcal{C} = \{\mathcal{C}_1, \cdots, \mathcal{C}_{d/r}\}$, where $\mathcal{C}_i$ is a subset of $[d]$ of size r and for every $i \neq i'$, we have that $\mathcal{C}_i \cap \mathcal{C}_{i'} = \emptyset$. Let us sample Q many sets $\{\mathcal{C}^{(j)}\}_{j \in [Q]}$ from it, then using union bound, we have that:

$$\Pr\left[\exists j \neq j', i, i' \text{ such that } \mathcal{C}_i^{(j)} = \mathcal{C}_{i'}^{(j')} \right] \leq Q^2 \left(\frac{d}{r} \right)^2 \left(\frac{r}{d} \right)^r.$$

Hence when $d \geq r^2$, for some $H = d^{O(r)}$, the above probability is smaller than one. This proves the existence of these sets. ∎

Now, we define the distribution $\mathcal{W}$. Recall that the Hadamard transform of dimension r is a unitary matrix in dimension r whose entries are all $\in \{-1/\sqrt{r}, 1/\sqrt{r}\}$.

Definition 26 (The hardness distribution for the lower bound) *For every r that is a power of 2, for every d that is a multiple of r bigger than r^2, we generate $\mathcal{W}$ as:*

1. Pick $\mathcal{C}$ uniformly at random from the set $\{\mathcal{C}^{(j)}\}_{j \in [H]}$ given by Lemma 25.

2. Define $w_i^\star \in \mathbb{R}^d$ with $i = pr + q$, for $p \in \{0, 1, \cdots d/r - 1\}$ and $q \in [r]$ as:

$$w_i^\star = (0^{pr}, h_q^\star, 0^{d-(p+1)r}),$$

where $h_q^\star$ is the i-th column of the Hadamard transform of dimension r.

3. *Sample $b_1, \cdots, b_d$ independent from $[1, 2]$ uniformly at random. Define*

$$a_i = \frac{b_i}{\sum_{j \in [d]} b_j}.$$

The proof of the lower bound relies on the following Lemma.

Lemma 27 (The boolean analysis lemma) *For every even $r \in \mathbb{N}^\star$, let $\mu = (\mu_1, \mu_2, \cdots, \mu_r) \in \mathbb{R}^r$ be sampled from the Gaussian distribution $\mathcal{N}(0, \mathrm{Id}_{r \times r})$. With probability at least $r^{-O(r)}$ over the choice of μ, it holds that:*

$$\lambda_\mu := \left| \underset{\tau \sim Uniform(\{-1,1\}^r)}{\mathbb{E}} \left[\left| \sum_{i \in [r]} \mu_i \tau_i \right| \prod_{i \in [r]} \tau_i \right] \right| \geq r^{-O(r)}$$

To prove this Lemma, we use Lemma $F.2$ and the proof of Corollary 7.1 in Allen-Zhu and Li (2019a), which says the following.

Corollary 28 (Lemma $F.2$ and Corollary 7.1 in Allen-Zhu and Li (2019a)) *For every $\varepsilon > 0$, there exists a value $V_{r,\varepsilon} = (r \log \frac{1}{\varepsilon})^{O(r)}$ and a function $h : \mathbb{R}^r \to [V_{r,\varepsilon}, V_{r,\varepsilon}]$ such that for every $\tau \in \{-1, 1\}^r$, it holds that:*

$$\underset{\mu \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})}{\mathbb{E}} \left[\left| \sum_{i \in [r]} \mu_i \tau_i \right| h(\mu) \right] = \prod_{i \in [r]} \tau_i \pm \varepsilon$$

Using this Corollary, we can prove Lemma 27.
Proof [Proof of Lemma 27] By applying Corollary 28 with $\varepsilon = 0.5$, we have that there exists a value $V = r^{O(r)}$ and a function $h : \mathbb{R}^r \to [V, V]$ such that

$$\underset{\mu \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})}{\mathbb{E}} \left[\mathrm{Abs} \left(\sum_{i \in [r]} \mu_i \tau_i \right) h(\mu) \right] = \prod_{i \in [r]} \tau_i \pm 0.5.$$

Hence we have that by $\tau_i \in \{-1, 1\}$:

$$\underset{\mu \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})}{\mathbb{E}} \left[h(\mu) \, \mathrm{Abs} \left(\sum_{i \in [r]} \mu_i \tau_i \right) \prod_{i \in [r]} \tau_i \right] = 1 \pm 0.5,$$

which means that

$$\underset{\mu \sim \mathcal{N}(0, \mathrm{Id}_{d \times d}); \tau \sim Uniform(\{-1,1\}^r)}{\mathbb{E}} \left[h(\mu) \, \mathrm{Abs} \left(\sum_{i \in [r]} \mu_i \tau_i \right) \prod_{i \in [r]} \tau_i \right] \geq \frac{1}{2}.$$

This immediately implies that

$$\underset{\mu \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})}{\mathbb{E}} \left[|h(\mu)| \left| \underset{\tau \sim Uniform(\{-1,1\}^r)}{\mathbb{E}} \left[\mathrm{Abs} \left(\sum_{i \in [r]} \mu_i \tau_i \right) \prod_{i \in [r]} \tau_i \right] \right| \right] \geq \frac{1}{2}.$$

Using the fact that $|h(\mu)| \leq V$, we have that

$$\mathbb{E}_{\mu \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})}[\lambda_\mu] \geq \frac{1}{2V}$$

Notice that with probability at least $1 - e^{r^2}$ over μ, we have that $\lambda_\mu \leq r^{O(r)}$. Note that $\lambda_\mu \geq 0$ as well. Thus, using Markov's inequality we complete the proof. ∎

Next we can derive the following corollary of Lemma 27. For two vectors x, y with the same dimension, we denote $x \circ y$ as the entry-wise product of x, y.

Corollary 29 *Let $p_1, \cdots, p_r$ be r vectors in $\{-1/\sqrt{r}, 1/\sqrt{r}\}^r$, let $q_1, \cdots, q_r$ be i.i.d. random variable chosen uniformly at random from $[1, 2]$, define $F_\mu(\tau) = \sum_{i \in [r]} q_i |\langle p_i \circ \mu, \tau \rangle|$, we have that with probability at least $r^{-O(r)}$ over $\mu \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})$ and q:*

$$\lambda_\mu^\star := \left| \mathbb{E}_{\tau \sim Uniform(\{-1,1\}^r)} \left[F_\mu(\tau) \prod_{i \in [r]} \tau_i \right] \right| \geq r^{-O(r)}.$$

Finally, we can complete the proof of Theorem 2.
Proof [Proof of Theorem 2] We prove by contradiction. Suppose on the contrary that equation (3) does not hold. Then, there exists ≥ 0.01 fraction of $\{a_i, w^\star\}_{i \in [d]}$ generated from $\mathcal{W}$ such that for some $w^{(R)}$ we have $\mathcal{R}(x) := w_R^\top \phi(x)$, and it holds that

$$\mathbb{E}_{x \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})} (f^\star(x) - \mathcal{R}(x))^2 = o\left(\frac{1}{d}\right) \tag{80}$$

We consider $x = \bar{x} \circ \tau$ where $\bar{x} \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})$ and $\tau \sim Uniform(\{-1, 1\}^d)$. Clearly, $x \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})$ as well. Thus,

$$\mathbb{E}_{x \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})} (f^\star(x) - \mathcal{R}(x))^2 = \mathbb{E}_{\tau \sim Uniform(\{-1,1\}^d); \bar{x} \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})} (f^\star(\bar{x} \circ \tau) - \mathcal{R}(\bar{x} \circ \tau))^2$$

Therefore, by Markov's inequality we have that with probability at least 0.999 over the choice of $\bar{x}$, we have that

$$\mathbb{E}_{\tau \sim Uniform(\{-1,1\}^d)} (f^\star(\bar{x} \circ \tau) - \mathcal{R}(\bar{x} \circ \tau))^2 = O\left(\mathbb{E}_{x \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})} (f^\star(x) - \mathcal{R}(x))^2 \right)$$

Now we perform Boolean Fourier analysis over $f^\star(\bar{x} \circ \tau)$ and $\mathcal{R}(\bar{x} \circ \tau)$. For a function $f : \{-1, 1\}^d \to \mathbb{R}$, we define it's Fourier expansion as:

$$f(\tau) = \sum_{\mathcal{B} \subseteq [d]} \lambda_\mathcal{B} \prod_{j \in \mathcal{B}} \tau_j,$$

where $\lambda_\mathcal{B}$ if the Fourier coefficient of the subset B. Now, define $\lambda_\mathcal{B}^\star$ to be the Fourier coefficients of $f^\star(\bar{x} \circ \tau)$ and $\lambda_\mathcal{B}^\mathcal{R}$ to be the Fourier coefficient of $\mathcal{R}(\bar{x} \circ \tau)$, we can observe that if we sample $\mathcal{C}$ from $\mathcal{W}$ to generate $w^\star$ according to Definition 26, then it holds that for every $\mathcal{B} \subset [d]$ of size r, we have:

$$\mathcal{B} \notin \mathcal{C} \implies \lambda_\mathcal{B}^\star = 0.$$

Moreover, using Corollary 29, we can conclude that w.p. at least 0.999 over $\mathcal{W}$,

$$\sum_{\mathcal{B} \in \mathcal{C}} (\lambda_{\mathcal{B}}^{\star})^2 \geq \frac{r^{-O(r)}}{d}. \tag{81}$$

On the other hand, for every $\varepsilon > 0$, as long as $\mathbb{E}_{\tau \sim Uniform(\{-1,1\}^d)} \left(f^{\star}(\bar{x} \circ \tau) - \mathcal{R}(\bar{x} \circ \tau) \right)^2 \leq \varepsilon$, we have that

$$\sum_{\mathcal{B} \in \mathcal{C}} \left(\lambda_{\mathcal{B}}^{\star} - \lambda_{\mathcal{B}}^{\mathcal{R}} \right)^2 + \sum_{\mathcal{B} \notin \mathcal{C}} \left(\lambda_{\mathcal{B}}^{\mathcal{R}} \right)^2 \leq \varepsilon$$

Let us consider the set $\mathcal{S}_{gd}$ of $\{a_i, w^{\star}\}_{i \in [d]}$ generated from $\mathcal{W}$. We call $\{a_i, w_i^{\star}\}_{i \in [d]} \in \mathcal{S}_{gd}$ if and only if the function $f^{\star}$ defined using $\{a_i, w^{\star}\}_{i \in [d]}$ satisfies Eq (81) and there is a $w^{(R)}$ such that for $\mathcal{R}(x) := w_R^{\top} \phi(x)$ with

$$\mathbb{E}_{x \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})} (f^{\star}(x) - \mathcal{R}(x))^2 = o\left(\frac{1}{d} \right)$$

We already know that there are at least 0.999 fraction $\{a_i, w^{\star}\}_{i \in [d]}$ generated from $\mathcal{W}$ that satisfies Eq (81). By our assumption, there are ≥ 0.01 fraction of $\{a_i, w^{\star}\}_{i \in [d]}$ generated from $\mathcal{W}$ satisfying that for some $w^{(R)}$ such that $\mathcal{R}(x) := w_R^{\top} \phi(x)$, it holds that

$$\mathbb{E}_{x \sim \mathcal{N}(0, \mathrm{Id}_{d \times d})} (f^{\star}(x) - \mathcal{R}(x))^2 = o\left(\frac{1}{d} \right)$$

Thus, we can conclude $|\mathcal{S}_{gd}| \geq 0.005|\mathcal{W}|$. Together with Lemma 25 which shows that $|\mathcal{W}| \geq d^{\Omega(r)}$, we know that $|\mathcal{S}_{gd}| \geq d^{\Omega(r)}$.

Now, we consider a matrix M, whose rows are indexed by each set of $\{a_i, w_i^{\star}\}_{i \in [d]} \in \mathcal{S}_{gd}$ and Eq (80), whose columns are indexed by $\lambda_{\mathcal{B}}^{\star}$ with $|\mathcal{B}| = r$.

We know that this matrix is of size $d^{\Omega(r)} \times d^{\Omega(r)}$. Moreover, for any matrix M' satisfies that

$$\forall i, \|M_i - M_i'\|_2^2 = o\left(\frac{1}{d} \right)$$

where M_i is the i-th row of M. It must holds that $\mathrm{rank}(M') = d^{\Omega(r)}$. We immediately complete the proof by contradiction, following exactly the same argument in the lower bound proof in Allen-Zhu and Li (2019a) while taking r to be a sufficiently large constant. ∎