How does Gradient Descent Learn Features – A Local Analysis for Regularized Two-Layer Neural Networks

Mo Zhou
Department of Computer Science
Duke University
mozhou@cs.duke.edu

Rong Ge
Department of Computer Science
Duke University
rongge@cs.duke.edu

Abstract

The ability of learning useful features is one of the major advantages of neural networks. Although recent works show that neural network can operate in a neural tangent kernel (NTK) regime that does not allow feature learning, many works also demonstrate the potential for neural networks to go beyond NTK regime and perform feature learning. Recently, a line of work highlighted the feature learning capabilities of the early stages of gradient-based training. In this paper we consider another mechanism for feature learning via gradient descent through a local convergence analysis. We show that once the loss is below a certain threshold, gradient descent with a carefully regularized objective will capture ground-truth directions. Our results demonstrate that feature learning not only happens at the initial gradient steps, but can also occur towards the end of training.

1 Introduction

The ability of learning useful features based on the data has long been considered to be a major advantage of neural networks. However, how gradient-based training algorithms can learn useful features are not well-understood. In particular, the most widely applied analysis for overparametrized neural networks is the neural tangent kernel (NTK) [19, 15, 5]. In this setting, the neurons don't move far from their initialization and the features are determined by the network architecture and random initialization.

While there are empirical and theoretical evidences on the limitation of NTK regime [12, 6], extending the analysis beyond the NTK regime has been challenging. Although for 2-layer networks, an alternative framework for analyzing overparametrized neural networks called mean-field analysis was introduced, earlier analysis (such as [11, 23]) require either infinite or exponentially many neurons. Later works (e.g., [20, 17, 10, 22]) can analyze the training dynamics of *mildly overparametrized networks* with polynomially many neurons with strong assumptions on the ground-truth function.

Recently, a line of work [14, 13, 1, 2, 27, 8, 24, 9] showed that early stages of gradient training (either one/a few steps of gradient descent or a small amount of time of gradient flow) can be useful in feature learning. These works show that after the early stages of gradient training, the first layer in a 2-layer neural network already captures useful features (usually in the form of a low dimensional subspace), and continue training the second layer weights will give performance guarantees that are stronger than any kernel or random feature based models. In this work, we consider the natural follow-up question:

Does feature learning only happen in the early stages of gradient training?

We show that this is not the case by demonstrating feature learning capability for the final stage of gradient training – local convergence. In particular, if the data is generated by a 2-layer teacher network, we prove (see Theorem 1) that if the loss is lower than a certain threshold that depends on the

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

complexity of the ground-truth function, gradient descent can continue to optimize the loss function to arbitrarily low test loss. In this process, the weights of first-layer neurons will all converge in direction to some ground truth directions in the teacher, which is a strong form of feature learning that guarantees the generalization performance. Analyzing the entire training dynamics is still challenging so in our algorithm (see Algorithm 1) we use a convex second stage to "fast-forward" to the local analysis. Our technique for local convergence is similar to the earlier work [28], however we consider a more complicated setting with ReLU activations and allow second-layer weights to be both positive or negative. While this change may seem minor, it requires additional regularization in the form of standard weight-decay and new dual certificate analysis.

2 Preliminary

Teacher-student setup We will consider the teacher-student setup for two-layer network networks with Gaussian input $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Consider the teacher network

$$f_*(\boldsymbol{x}) = \sum_{i=1}^{m^*} a_i^* \sigma(\boldsymbol{w}_i^{*\top} \boldsymbol{x}) + \boldsymbol{w}_0^{*\top} \boldsymbol{x} + b_0^*,$$

where σ is ReLU activation, $\dim(S_*) = r$ and $S_* = \operatorname{span}\{\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_{m^*}^*\}$ is the target subspace. Without loss of generality, we will assume $\|\boldsymbol{w}_i^*\|_2 = 1$ due to the homogeneity of ReLU. We will make the following non-degenerate assumptions on the teacher neurons:

Assumption 1. *Teacher neurons are Δ separated, that is $\angle(\boldsymbol{w}_i^*, \boldsymbol{w}_j^*) \geq \Delta$ for all $i \neq j$, where $\angle(\boldsymbol{w}, \boldsymbol{v}) := \arccos(|\boldsymbol{w}^\top \boldsymbol{v}| / \|\boldsymbol{w}\|_2 \|\boldsymbol{v}\|_2)$.*

Assumption 2. *Matrix $\boldsymbol{H} := \sum_{i=1}^{m^*} a_i^* \boldsymbol{w}_i^* \boldsymbol{w}_i^{*\top}$ is non degenerate, i.e., $\kappa := \lambda_{\min}(\boldsymbol{H}) > 0$.*

The first assumption above simply requires all teacher neurons pointing to different directions. Note that f_* can have both neurons $\boldsymbol{w}^*$ and $-\boldsymbol{w}^*$ by using the identity $\sigma(x) - \sigma(-x) = x$. The second assumption roughly says the target network contains low-order (second-order) information, which is related with the notion of information exponent [7]. Many previous works also rely on this or similar assumption to show neural networks can learn features to perform better than kernels [13, 2, 8].

We will use the following student network with extra linear term (can also view as skip connection):

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{m} a_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x}) + \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}, \tag{1}$$

where $\boldsymbol{a} = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$, $\boldsymbol{W} = (\boldsymbol{w}_1 \cdots \boldsymbol{w}_m)^\top \in \mathbb{R}^{m \times d}$ and $\boldsymbol{\theta} = (\boldsymbol{a}, \boldsymbol{W}, \alpha, \boldsymbol{\beta})$.

Preprocessing data Give any $(\boldsymbol{x}, y)$ with $y = f_*(\boldsymbol{x})$, denote $\alpha_* = \mathbb{E}_{\boldsymbol{x}}[y]$ and $\boldsymbol{\beta}_* = \mathbb{E}_{\boldsymbol{x}}[y\boldsymbol{x}]$, let

$$\widetilde{f}_*(\boldsymbol{x}) = \widetilde{y} = y - \alpha_* - \boldsymbol{\beta}_*^\top \boldsymbol{x}.$$

This preprocessing process essentially removes the 0-th and 1-st order term in the Hermite expansion of σ. See Section I for a brief introduction of Hermite polynomials.

Loss and algorithm Denote the regularized loss function under Gaussian input $\boldsymbol{x} \in \mathbb{R}^d$ as

$$L_\lambda(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x} \sim N(0, \boldsymbol{I}_d)}[(f(\boldsymbol{x}; \boldsymbol{\theta}) - \widetilde{y})^2] + \frac{\lambda}{2} \|\boldsymbol{a}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{W}\|_2^2. \tag{2}$$

We will also use L to denote the square loss for simplicity.

Our algorithm shown in Algorithm 1 is roughly gradient descent (GD) following a given schedule of weight decay λ_t and step size η_t. We will use symmetric initialization that $a_i = -a_{i+m/2}$, $\boldsymbol{w}_i = \boldsymbol{w}_{i+m/2}$ with $a_i \sim \operatorname{Unif}\{-1/\sqrt{m}, 1/\sqrt{m}\}$, $\boldsymbol{w}_i \sim \operatorname{Unif}((1/\sqrt{m})\mathbb{S}^{d-1})$, $\alpha = 0$, $\boldsymbol{\beta} = \boldsymbol{0}$.

Due to the difficulty in analyzing gradient descent training beyond early and final stage, we choose to only train the norms in Stage 2 as a tractable way to reach the local convergence regime.

Notation Denote $[n]$ as the set $\{1, 2, \ldots, n\}$. We use $\|\boldsymbol{w}\|_2$ for 2-norm of vector $\boldsymbol{w}$, and $\overline{\boldsymbol{w}} = \boldsymbol{w} / \|\boldsymbol{w}\|_2$ as its normalization. Denote $\|\boldsymbol{A}\|_F$ as Frobenius norm of matrix $\boldsymbol{A}$. Denote $\angle(\boldsymbol{w}, \boldsymbol{v}) = \arccos(|\boldsymbol{w}^\top \boldsymbol{v}| / (\|\boldsymbol{w}\|_2 \|\boldsymbol{v}\|_2)] \in [0, \pi/2]$ as the angle between vectores $\boldsymbol{w}$ and $\boldsymbol{v}$ (up to a sign). We will use O_*, Ω_*, Θ_* to hide $\operatorname{poly}(r, m_*, \Delta, a_{\min}, \|\boldsymbol{a}_*\|_1)$, which is the polynomial dependency on relevant parameters of target f_*, and $\widetilde{O}, \widetilde{\Omega}, \widetilde{\Theta}$ to hide polylog factors.

```
Algorithm 1: Learning regularized 2-layer neural networks
Input: initialization $\boldsymbol{\theta}^{(0)}$, schedule of weight decay $\lambda_t$ and stepsize $\eta_t$
Preprocessing data:
$\alpha_* \leftarrow \mathbb{E}_{\boldsymbol{x}}[y]$, $\boldsymbol{\beta}_* \leftarrow \mathbb{E}_{\boldsymbol{x}}[y\boldsymbol{x}]$
$\widetilde{y} \leftarrow y - \alpha_* - \boldsymbol{\beta}_*^\top \boldsymbol{x}$
Stage 1:    // first step gradient
$\boldsymbol{\theta}^{(1)} \leftarrow \boldsymbol{\theta}^{(0)} - \eta_0 \nabla_{\boldsymbol{\theta}} L_{\lambda_0}(\boldsymbol{\theta}^{(0)})$
Stage 2:    // adjust norm
$a_i^{(1)} \leftarrow a_i^{(1)} \left\|\boldsymbol{w}_i^{(1)}\right\|_2, \left\|\boldsymbol{w}_i^{(1)}\right\|_2 \leftarrow 1$
$\boldsymbol{a}^{(T_2)}, \alpha^{(T_2)}, \boldsymbol{\beta}^{(T_2)} \leftarrow \min_{\boldsymbol{a}} \min_{\alpha,\boldsymbol{\beta}} L(\boldsymbol{\theta}) + \lambda \|\boldsymbol{a}\|_1$
$a_i^{(T_2)} \leftarrow a_i^{(T_2)}/\sqrt{|a_i^{(T_2)}|}, \boldsymbol{w}_i^{(T_2)} \leftarrow \boldsymbol{w}_i^{(T_2)}\sqrt{|a_i^{(T_2)}|}$
Stage 3:    // local convergence
for $k \le K$ do
    for $T_{3,k} \le t \le T_{3,(k+1)}$ do
        $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta_t \nabla_{\boldsymbol{\theta}} L_{\lambda_t}(\boldsymbol{\theta}^{(t)})$
    end
end
Output: $\boldsymbol{\theta}^{(T_{3.K})} = (\boldsymbol{a}^{(T_{3,K})}, \boldsymbol{W}^{(T_{3,K})}, \alpha^{(T_{3,K})}, \boldsymbol{\beta}^{(T_{3,K})})$
```

3 Main Results

In this section, we give our main result that shows training student network using Algorithm 1 can recover the target network within polynomial time. We will focus on the case that $d \ge \Omega_*(1)$ when the complexity of target function is small.

Theorem 1 (Main result)**.** *Under Assumption 1,2, consider Algorithm 1 on loss (2). There exists a schedule of weight decay λ_t and step size $\eta_0 = 1$, $\eta_t = \eta = O_*(1/\max\{m,d\})$ such that given $m \ge m_0 = \widetilde{O}_*(1) \cdot (1/\varepsilon_0)^{O(r)}$ neurons with small enough $\varepsilon_0 = \Theta_*(1)$, with high probability we will recover the target network $L(\boldsymbol{\theta}) \le \varepsilon$ within time $T = O_*(1/\eta)\operatorname{poly}(1/\varepsilon)$.*

Note that our results can be extended to only have access to polynomial number of samples by using standard concentration tools. We omit the sample complexity for simplicity in the current version. As a corollary of the main result we also show that to minimize the loss the weight vectors in student network must converge in direction to the teacher network.

Corollary 2. *Denote angle $\delta_j = \angle(\boldsymbol{w}_j, \boldsymbol{w}_i^*)$ for $j \in \mathcal{T}_i$, where $\mathcal{T}_i := \{j : \angle(\boldsymbol{w}_j, \boldsymbol{w}_i^*) \le \angle(\boldsymbol{w}_j, \boldsymbol{w}_k^*)\ \forall k \ne i\}$ forms a partition of the neurons based on the angle to ground-truth direction. At the end of training the following hold*

(i) Far-away neurons are small: $\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} |a_j| \|\boldsymbol{w}_j\|_2 \delta_j^2 = O_*(\sqrt{\varepsilon})$.

(ii) Most neurons are close to ground-truth direction: $\sum_{j\in\mathcal{T}_i:\delta_j\le\delta_{cls}} a_j \|\boldsymbol{w}_j\|_2 \operatorname{sign}(a_i^*) \ge |a_i^*|/2$, where $\delta_{cls} = O_*(\varepsilon^{1/3})$.

In particular when $\varepsilon \to 0$ or equivalently $\lim_{\lambda\to 0} L_\lambda(\boldsymbol{\theta}) = 0$, the above imply that the student neurons' directions match the ground-truth directions.

Our result improves the previous works that only train the first layer weight with small number of gradient steps at the beginning [13, 8, 1, 2]. In these works, neural networks only learn the target subspace and do random features within it. Intuitively, these random features needs to span the whole space of the target function class to perform well, which means its number (the width) should be on the order of the dimension of target function class. For 2-layer networks random features in the target subspace need $(1/\varepsilon)^{O(r)}$ neurons to achieve desired accuracy ε. In contrast, continue training both layer at the last phase of training allows us to learn not only subspace but also exactly the ground-truth directions. Moreover, we only use $(1/\varepsilon_0)^{O(r)}$ neurons that only depends on the complexity of target network. This highlights the benefit of continue training first layer weights instead of fixing them after first step.

4 Proof Sketch

We analyze the three stages separately. For Stage 1, we use the following lemma to show that the first step of gradient descent identifies the target subspace, and it has student neurons that are close to teacher neurons. The key observation here is similar to [13] that $\boldsymbol{w}_i^{(1)} \approx -2\eta_0 a_i^{(0)} \left(\hat{\sigma}_2^2 \boldsymbol{H}\overline{\boldsymbol{w}}_i\right)$ so that given $\boldsymbol{H}$ is non-degenerate we essentially sample $\boldsymbol{w}_i^{(1)}$ from the target subspace.

Lemma 3 (Stage 1)**.** *Under Assumption 1,2, consider Algorithm 1 on loss (2) with $\lambda_0^{-1} = \eta_0 = 1$ and $m \ge m_0 = \widetilde{O}_*(1) \cdot (1/\varepsilon_0)^{O(r)}$ with any given $\varepsilon_0 = \Theta_*(1)$. After first iteration, we have with probability at least $1 - \delta$*

(i) for every teacher neuron $\boldsymbol{w}_i^*$, there exists at least one student neuron $\boldsymbol{w}_i$ such that $\angle(\boldsymbol{w}_i^*, \boldsymbol{w}_j) \le \varepsilon_0$

(ii) $\Omega_*(\frac{\kappa}{\sqrt{md}}) \le \left\|\boldsymbol{w}_i^{(1)}\right\|_2 \le O_*(\frac{1}{\sqrt{md}})$, $|a_i^{(1)}| \le O_*(\frac{1}{\sqrt{m}})$ for all $i \in [m_*]$, $\alpha_1 = 0$ and $\boldsymbol{\beta}_1 = \mathbf{0}$.

Given learned features in Stage 1, we now adjust the norms to reach a low loss solution in Stage 2.

Lemma 4 (Stage 2)**.** *Under Assumption 1,2, consider Algorithm 1 with $\lambda_t = \sqrt{\varepsilon_0}$ and $\eta_t = \eta \le O_*(\varepsilon_0/m)$ to be small enough for $t \le T_2$. Given Stage 1 in Lemma 3, we have Stage 2 ends within time $T_2 = O_*(1/\eta\varepsilon_0)$ such that the optimality gap $\zeta_{T_2} = L_\lambda(\boldsymbol{\theta}^{(T_2)}) - \min_{\mu \in \mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu) \le O_*(\varepsilon_0)$.*

After Stage 2 we already in the local convergence regime. The following lemma shows that we could recover the target network within polynomial time using a multi-epoch gradient descent that decreasing the weight decay λ at every epoch. Note that this result only requires the initial optimality gap is small and width $m \ge m_*$.

Lemma 5 (Stage 3)**.** *Under Assumption 1,2, consider Algorithm 1 on loss (2). Given Stage 2 in Lemma 4, if the initial optimality gap $\zeta_{3,0} \le O_*(\lambda_{3,0}^{9/5})$, weight decay λ follows the schedule of initial value $\lambda_{3,0} = O_*(1)$, and k-th epoch $\lambda_{3,k} = \lambda_{3,0}/(k\lambda_{3,0} + 1)$ and stepsize $\eta_{3k} = \eta \le O_*(\lambda_{3,k}^{-6} d^{-1})$ for all $T_{3,k} \le t \le T_{3,k+1}$ in epoch k, then within $K = O_*(\varepsilon^{-1/2})$ epochs and total $T_3 - T_2 = O_*(\varepsilon^{-5/2}\eta^{-1})$ time we recover the ground-truth network $L(\boldsymbol{\theta}) \le \varepsilon$.*

The lemma above relies on the following result that shows the local landscape is benign in the sense that it satisfies a special case of Łojasiewicz property [21].

Lemma 6 (Gradient lower bound)**.** *Suppose the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - \min_{\mu \in \mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu)$. If $\Omega_*(\lambda^2) \le \zeta \le O_*(\lambda^{9/5})$ and $\lambda \le O_*(1)$, we have*

$$\|\nabla_{\boldsymbol{\theta}} L_\lambda\|_F^2 \ge \Omega_*(\zeta^4/\lambda^2).$$

Note that this generalize the previous result in [28] that only focuses on the two-layer networks with positive second layer weights. The key idea is to construct descent direction that has a large correlation with the gradient direction to get a gradient lower bound. However, the appearance of both positive and negative second layer weights introduces more challenges compared to positive second layer weights, mostly due to the cancellation between neurons with similar directions. We introduce the standard weight decay to allow us handle the cancellation between neurons, since reducing their norms simultaneously would decrease the regularization term and keep the square loss term the same. We use a new dual certificate analysis based on [26] and the idea of residual decomposition and average neuron [28] to characterize the structure of solution. We show that there are always neurons close to the teacher neurons and far-away neurons are small. These properties help us to construct descent directions to get gradient lower bound.

5 Conclusion

In this paper we showed that gradient descent converges in a large local region depending on the complexity of the teacher network, and the local convergence allows 2-layer networks to perform a strong notion of feature learning (matching the directions of ground-truth teacher networks). We hope our result gives a better understanding of why gradient-based training is important for feature learning in neural networks. A natural next step is to understand whether the intermediate steps are also important for feature learning. This is a challenging open problem using the current techniques as the dynamics is very complicated without very strong assumptions (and this is also the reason why we need to optimize only the second-layer in Stage 2).

References

[1] Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34:26989–27002, 2021.

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

[3] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[4] P-A Absil, Robert Mahony, and Jochen Trumpf. An extrinsic look at the riemannian hessian. In *International conference on geometric science of information*, pages 361–368. Springer, 2013.

[5] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252, 2019.

[6] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

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

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

[9] Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

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

[11] 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*, pages 3036–3046, 2018.

[12] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

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

[14] Amit Daniely and Eran Malach. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

[15] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[16] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.

[17] Rong Ge, Yunwei Ren, Xiang Wang, and Mo Zhou. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34: 1299–1311, 2021.

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

[19] 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.

[20] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pages 2613–2682. PMLR, 2020.

[21] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[22] Arvind Mahankali, Jeff Z Haochen, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. *arXiv preprint arXiv:2306.16361*, 2023.

[23] 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.

[24] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.

[25] Ryan O'Donnell. Analysis of boolean functions. *arXiv preprint arXiv:2105.10386*, 2021.

[26] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*, 23(1):241–327, 2023.

[27] Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. *arXiv preprint arXiv:2206.01717*, 2022.

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

A Useful facts and proof of Theorem 1

In this section we provide several useful facts and present the proof of Theorem 1.

Claim 1. *Denote $\hat{\alpha} = -(1/\sqrt{2\pi})\sum_{i=1}^{m} a_i \left\|\boldsymbol{w}_i\right\|_2$, $\hat{\boldsymbol{\beta}} = -(1/2)\sum_{i=1}^{m} a_i \boldsymbol{w}_i$.We have square loss as*

$$L(\boldsymbol{\theta}) = |\alpha - \hat{\alpha}|^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 + \mathbb{E}_{\boldsymbol{x}}[(f_{\geq 2}(\boldsymbol{x}) - \widetilde{f}_*(\boldsymbol{x}))^2]$$

where $f_{\geq 2}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i\in[m]} a_i \sigma_{\geq 2}(\boldsymbol{w}_i^\top \boldsymbol{x})$ and $\sigma_{\geq 2}(x) = \sigma(x) - 1/\sqrt{2\pi} - x/2$ is the activation that after removing 0th and 1st order term in Hermite expansion.

Proof. Following Ge et al. [16], we can write the loss $L(\boldsymbol{\theta})$ as a sum of tensor decomposition problem using Hermite expansion as in Section I (recall $\left\|\boldsymbol{w}_i^*\right\|_2 = 1$ and preprocessing procedure removes the 0-th and 1-st order term in the Hermite expansion of σ):

$$\begin{aligned} L(\boldsymbol{\theta}) =& \mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i\in[m]} a_i \left\|\boldsymbol{w}_i\right\|_2 \sum_{k\geq 0} \hat{\sigma}_k h_k(\overline{\boldsymbol{w}}_i^\top \boldsymbol{x}) + \alpha + h_1(\boldsymbol{\beta}^\top \boldsymbol{x}) - \sum_{i\in[m_*]} a_i^* \left\|\boldsymbol{w}_i^*\right\|_2 \sum_{k\geq 2} \hat{\sigma}_k h_k(\boldsymbol{w}_i^{*\top} \boldsymbol{x})\right)^2\right] \\ =& \left|\alpha + \hat{\sigma}_0 \sum_{i\in[m]} a_i \left\|\boldsymbol{w}_i\right\|_2\right|^2 + \left\|\boldsymbol{\beta} + \hat{\sigma}_1 \sum_{i\in[m]} a_i \boldsymbol{w}_i\right\|_2^2 + \sum_{k\geq 2} \hat{\sigma}_k^2 \left\|\sum_{i\in[m]} a_i \left\|\boldsymbol{w}_i\right\|_2 \overline{\boldsymbol{w}}_i^{\otimes k} - \sum_{i\in[m_*]} a_i^* \left\|\boldsymbol{w}_i^*\right\|_2 \boldsymbol{w}_i^{*\otimes k}\right\|_F^2. \end{aligned}$$

Note that $\hat{\sigma}_0 = 1/\sqrt{2\pi}$, $\hat{\sigma}_1 = 1/2$ as in Lemma 37, we get the result. □

Corollary 2. *Denote angle $\delta_j = \angle(\boldsymbol{w}_j, \boldsymbol{w}_i^*)$ for $j \in \mathcal{T}_i$, where $\mathcal{T}_i := \{j : \angle(\boldsymbol{w}_j, \boldsymbol{w}_i^*) \leq \angle(\boldsymbol{w}_j, \boldsymbol{w}_k^*)\ \forall k \neq i\}$ forms a partition of the neurons based on the angle to ground-truth direction. At the end of training the following hold*

(i) Far-away neurons are small: $\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} |a_j| \left\|\boldsymbol{w}_j\right\|_2 \delta_j^2 = O_*(\sqrt{\varepsilon})$.

(ii) Most neurons are close to ground-truth direction: $\sum_{j\in\mathcal{T}_i:\delta_j\leq\delta_{cls}} a_j \left\|\boldsymbol{w}_j\right\|_2 \operatorname{sign}(a_i^*) \geq |a_i^*|/2$, where $\delta_{cls} = O_*(\varepsilon^{1/3})$.

In particular when $\varepsilon \to 0$ or equivalently $\lim_{\lambda\to 0} L_\lambda(\boldsymbol{\theta}) = 0$, the above imply that the student neurons' directions match the ground-truth directions.

Proof. This is a direct corollary from Lemma 20 and Lemma 19. □

Theorem 1 (Main result)**.** *Under Assumption 1,2, consider Algorithm 1 on loss (2). There exists a schedule of weight decay λ_t and step size $\eta_0 = 1$, $\eta_t = \eta = O_*(1/\max\{m,d\})$ such that given $m \geq m_0 = \widetilde{O}_*(1)\cdot(1/\varepsilon_0)^{O(r)}$ neurons with small enough $\varepsilon_0 = \Theta_*(1)$, with high probability we will recover the target network $L(\boldsymbol{\theta}) \leq \varepsilon$ within time $T = O_*(1/\eta)\operatorname{poly}(1/\varepsilon)$.*

Proof. Combine Lemma 3 (Stage 1), Lemma 4 (Stage 2) and Lemma 5 (Stage 3) together and follow the choice of λ_t and η_t we get the result. □

B Stage 1: first gradient step

In this section, we show that after the first gradient update $\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_m\}$ forms a ε_0-net for the target subspace S_*, given $m = (1/\varepsilon_0)^{O(r)}$ neurons. The proof is deferred to Section B.1.

Lemma 3 (Stage 1)**.** *Under Assumption 1,2, consider Algorithm 1 on loss (2) with $\lambda_0^{-1} = \eta_0 = 1$ and $m \geq m_0 = \widetilde{O}_*(1)\cdot(1/\varepsilon_0)^{O(r)}$ with any given $\varepsilon_0 = \Theta_*(1)$. After first iteration, we have with probability at least $1-\delta$*

(i) for every teacher neuron $\boldsymbol{w}_i^*$, there exists at least one student neuron $\boldsymbol{w}_i$ such that $\angle(\boldsymbol{w}_i^*, \boldsymbol{w}_j) \leq \varepsilon_0$

(ii) $\Omega_*(\frac{\kappa}{\sqrt{md}}) \le \left\|\boldsymbol{w}_i^{(1)}\right\|_2 \le O_*(\frac{1}{\sqrt{md}})$, $|a_i^{(1)}| \le O_*(\frac{1}{\sqrt{m}})$ *for all* $i \in [m_*]$, $\alpha_1 = 0$ *and* $\boldsymbol{\beta}_1 = \mathbf{0}$.

The proof relies on the following lemma from [13] that shows after the first step update $\boldsymbol{w}_i$'s are located at positions as if they are sampled within the target subspace S_*.

Lemma 7 (Lemma 4, [13])**.** *Under Assumption 2, we have with high probability in the ℓ_2 norm sense*

$$\boldsymbol{w}_i^{(1)} = -\eta_0 \nabla_{\boldsymbol{w}_i} L(\boldsymbol{a}^{(0)}, \boldsymbol{W}^{(0)}) = -2\eta_0 a_i^{(0)} \left(\hat{\sigma}_2^2 \boldsymbol{H}\overline{\boldsymbol{w}}_i + \widetilde{O}(\frac{\sqrt{r}}{d})\right),$$

where $\hat{\sigma}_k := \mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{x}) h_k(\boldsymbol{x})]$ is the k-th Hermite polynomial coefficient.

B.1 Proofs in Section B

Lemma 3 (Stage 1)**.** *Under Assumption 1,2, consider Algorithm 1 on loss (2) with $\lambda_0^{-1} = \eta_0 = 1$ and $m \ge m_0 = \widetilde{O}_*(1)\cdot(1/\varepsilon_0)^{O(r)}$ with any given $\varepsilon_0 = \Theta_*(1)$. After first iteration, we have with probability at least $1-\delta$*

(i) for every teacher neuron $\boldsymbol{w}_i^*$, there exists at least one student neuron $\boldsymbol{w}_i$ such that $\angle(\boldsymbol{w}_i^*, \boldsymbol{w}_j) \le \varepsilon_0$

(ii) $\Omega_*(\frac{\kappa}{\sqrt{md}}) \le \left\|\boldsymbol{w}_i^{(1)}\right\|_2 \le O_*(\frac{1}{\sqrt{md}})$, $|a_i^{(1)}| \le O_*(\frac{1}{\sqrt{m}})$ *for all* $i \in [m_*]$, $\alpha_1 = 0$ *and* $\boldsymbol{\beta}_1 = \mathbf{0}$.

Proof. We show them one by one.

Part (i) From Lemma 7 and the fact that $\overline{\boldsymbol{w}}_i^{(0)}$ samples uniformly from unit sphere, we know the probability of $\angle(\overline{\boldsymbol{w}}_i^{(1)}, \boldsymbol{w})$ for any given $\boldsymbol{w}$ is at least $\Omega_*(\varepsilon_0^r)$. Applying union bound we get the desired result.

Part (ii) We have

$$\boldsymbol{w}_i^{(1)} = -\eta_0 \nabla_{\boldsymbol{w}_i} L(\boldsymbol{a}^{(0)}, \boldsymbol{W}^{(0)}) = a_i^{(0)} \mathbb{E}_{\boldsymbol{x}}[\widetilde{f}_*(\boldsymbol{x})\sigma'(\boldsymbol{w}_i^\top \boldsymbol{x})\boldsymbol{x}]$$

For the norm bound, using Lemma 7 we know

$$\Theta(\frac{1}{\sqrt{m}}) \left\|\boldsymbol{H}\overline{\boldsymbol{w}}_i^{(0)}\right\|_2 - \widetilde{O}(\frac{\sqrt{r}}{d\sqrt{m}}) \le (1/\eta_0)\left\|\boldsymbol{w}_i^{(1)}\right\|_2 \le \Theta(\frac{1}{\sqrt{m}}) \left\|\boldsymbol{H}\overline{\boldsymbol{w}}_i^{(0)}\right\|_2 + \widetilde{O}(\frac{\sqrt{r}}{d\sqrt{m}}).$$

Since $\boldsymbol{w}_i^{(0)}$ initializes from Gaussian distribution, we know the desired bound hold. Similarly, one can bound $|a_i^{(1)}|$.

Since we use a symmetric initialization, it is easy to see $\alpha, \boldsymbol{\beta}$ remains at 0.

□

C Stage 2: reaching low loss

In Stage 2, we show that given the features learned in Stage 1 one can adjust the norms on top of it to reach low loss that enters the local convergence regime in Stage 3. The proof is deferred to Section C.1.

We first specify the procedure to solve $\min_{\boldsymbol{a}} \min_{\alpha,\boldsymbol{\beta}} L(\boldsymbol{\theta}) + \lambda \|\boldsymbol{a}\|_1$. For $\boldsymbol{a}$ at current point, we first solve the inner optimization problem, which is a linear regression on $\alpha, \boldsymbol{\beta}$. Then given the $\alpha, \boldsymbol{\beta}$, the outer optimization is a convex optimization for $\boldsymbol{a}$, which can also be solved efficiently.

The following lemma shows that after Stage 2 we reach a low loss solution given the first layer features learned after first gradient step.

Lemma 4 (Stage 2)**.** *Under Assumption 1,2, consider Algorithm 1 with $\lambda_t = \sqrt{\varepsilon_0}$ and $\eta_t = \eta \le O_*(\varepsilon_0/m)$ to be small enough for $t \le T_2$. Given Stage 1 in Lemma 3, we have Stage 2 ends within time $T_2 = O_*(1/\eta\varepsilon_0)$ such that the optimality gap $\zeta_{T_2} = L_\lambda(\boldsymbol{\theta}^{(T_2)}) - \min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu) \le O_*(\varepsilon_0)$.*

The lemma below shows that the inner loop ends quickly. In fact the inner optimization problem is strongly convex in $\alpha, \boldsymbol{\beta}$, which can be seen from Claim 1.

Lemma 8 (Descent direction, α and $\boldsymbol{\beta}$)**.** *We have*

$$|\nabla_\alpha L_\lambda|^2 = 4(\alpha - \hat{\alpha})^2, \quad \|\nabla_{\boldsymbol{\beta}} L_\lambda\|_2^2 = 4 \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2.$$

The following lemma relates the gradient of $\boldsymbol{a}$ to the ideal loss that as if $\alpha, \boldsymbol{\beta}$ are perfectly fitted. This allows us to transfer useful properties of $\widetilde{L}_{1,\lambda}$ to $L_{1,\lambda}$.

Lemma 9. *Let the ideal loss $\widetilde{L}_{1,\lambda}(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{a}^\top \sigma_{\geq 2}(\boldsymbol{W}\boldsymbol{x}) - \widetilde{y})^2] + \lambda \|\boldsymbol{a}\|_1$ that perfectly fits $\alpha, \boldsymbol{\beta}$. Given any $\|\boldsymbol{a}\|_1, \|\widetilde{\boldsymbol{a}}\|_1 = O_*(1/\lambda)$ and $\|\boldsymbol{w}_i\|_2 = 1$, we have*

$$\begin{aligned}
&|\langle \nabla_{\boldsymbol{a}} \widetilde{L}_{1,\lambda} - \nabla_{\boldsymbol{a}} L_{1,\lambda}, \boldsymbol{a} - \widetilde{\boldsymbol{a}}\rangle| \leq O_*(1/\lambda)(|\alpha - \hat{\alpha}| + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2) \\
&L_{1,\lambda}(\boldsymbol{a}) \leq \widetilde{L}_{1,\lambda}(\boldsymbol{a}) + |\alpha - \hat{\alpha}|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}\|_2^2 \\
&\|\nabla_{\boldsymbol{a}} L_{1,\lambda}\|_2 = O_*(\sqrt{m})
\end{aligned}$$

C.1 Proofs in Section C

Lemma 4 (Stage 2)**.** *Under Assumption 1,2, consider Algorithm 1 with $\lambda_t = \sqrt{\varepsilon_0}$ and $\eta_t = \eta \leq O_*(\varepsilon_0/m)$ to be small enough for $t \leq T_2$. Given Stage 1 in Lemma 3, we have Stage 2 ends within time $T_2 = O_*(1/\eta\varepsilon_0)$ such that the optimality gap $\zeta_{T_2} = L_\lambda(\boldsymbol{\theta}^{(T_2)}) - \min_{\mu \in \mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu) \leq O_*(\varepsilon_0)$.*

Proof. From Lemma 8 we know the inner loop ends within $O_*(1)$ time and $(\alpha - \hat{\alpha})^2, \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 \leq O_*(\lambda^2\varepsilon_0^2)$ with small enough hidden factor.

Then for the outer loop, we have

$$\begin{aligned}
\left\|\boldsymbol{a}^{(t+1)} - \widetilde{\boldsymbol{a}}_*\right\|_2^2 &= \left\|\boldsymbol{a}^{(t)} - \widetilde{\boldsymbol{a}}_*\right\|_2^2 - 2\eta\langle \nabla_{\boldsymbol{a}} L_{1,\lambda}(\boldsymbol{a}^{(t)}), \boldsymbol{a}^{(t)} - \widetilde{\boldsymbol{a}}_*\rangle + \eta^2 \left\|\nabla_{\boldsymbol{a}} L_{1,\lambda}(\boldsymbol{a}^{(t)})\right\|_2^2 \\
&\overset{\text{(a)}}{\leq} \left\|\boldsymbol{a}^{(t)} - \widetilde{\boldsymbol{a}}_*\right\|_2^2 - 2\eta(\widetilde{L}_{1,\lambda}(\boldsymbol{a}^{(t)}) - \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*)) + \eta\varepsilon_0/4 + \eta^2 O_*(m) \\
&= \left\|\boldsymbol{a}^{(t)} - \widetilde{\boldsymbol{a}}_*\right\|_2^2 - 2\eta(\widetilde{L}_{1,\lambda}(\boldsymbol{a}^{(t)}) - \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*)) + \eta\varepsilon_0/2,
\end{aligned}$$

where (a) we use $\widetilde{L}_{1,\lambda}$ (defined in Lemma 9) is convex in $\boldsymbol{a}$ and Lemma 9.

Iterating the above inequality over all t we have

$$\left\|\boldsymbol{a}^{(T)} - \widetilde{\boldsymbol{a}}_*\right\|_2^2 \leq \left\|\boldsymbol{a}^{(1)} - \widetilde{\boldsymbol{a}}_*\right\|_2^2 - 2\eta \sum_{t \leq T} (\widetilde{L}_{1,\lambda}(\boldsymbol{a}^{(t)}) - \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*)) + \eta T\varepsilon_0/2,$$

which means

$$\min_{t \leq T} \widetilde{L}_{1,\lambda}(\boldsymbol{a}^{(t)}) - \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*) \leq \frac{1}{T} \sum_{t \leq T} (\widetilde{L}_{1,\lambda}(\boldsymbol{a}^{(t)}) - \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*)) \leq \frac{\left\|\boldsymbol{a}^{(1)} - \widetilde{\boldsymbol{a}}_*\right\|_2^2}{\eta T} + \varepsilon_0/2.$$

It is easy to see $\|\widetilde{\boldsymbol{a}}_*\|_1 = O_*(1)$. Thus, when $T \geq O_*(1/\eta\varepsilon_0)$ we know $\widetilde{L}_{1,\lambda}(\boldsymbol{a}^{(T_2)}) - \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*) \leq 3\varepsilon_0/4$. This suggests the optimality gap

$$\begin{aligned}
\zeta_{T_2} =& L_\lambda(\boldsymbol{\theta}^{(T_2)}) - \min_{\mu \in \mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu) \\
=& L_\lambda(\boldsymbol{\theta}^{(T_2)}) - \widetilde{L}_{1,\lambda}(\boldsymbol{\theta}^{(T_2)}) + \widetilde{L}_{1,\lambda}(\boldsymbol{a}^{(T_2)}) - \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*) + \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*) - \min_{\mu \in \mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu).
\end{aligned}$$

For $L_\lambda(\boldsymbol{\theta}^{(T_2)}) - \widetilde{L}_{1,\lambda}(\boldsymbol{\theta}^{(T_2)})$, noting that we have balanced the norm of two layers, so they are the same.

For $\widetilde{L}_{1,\lambda}(\boldsymbol{a}^{(T_2)}) - \widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*)$, we just show above that it is less than $3\varepsilon_0/4$.

For $\widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*) - \min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu)$, we have

$$\widetilde{L}_{1,\lambda}(\widetilde{\boldsymbol{a}}_*) - \min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu) \le \widetilde{L}_{1,\lambda}(\boldsymbol{a}_*) - \min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu) \le \lambda \left\|\boldsymbol{a}_*\right\|_1 - \lambda|\mu_\lambda^*|_1 \le O_*(\lambda^2),$$

where in the last inequality we use Lemma 17 and $\mu_\lambda^* = \arg\min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu)$.

Together with above calculations, we have $\zeta_{T_2} \le O_*(\varepsilon_0)$. □

The following lemmas rely on the loss decomposition in Claim 1 that $\alpha, \boldsymbol{\beta}$ only fits 0-th and 1-st order term in Hermite expansion.

Lemma 8 (Descent direction, α and $\boldsymbol{\beta}$)**.** *We have*

$$|\nabla_\alpha L_\lambda|^2 = 4(\alpha - \hat{\alpha})^2, \quad \|\nabla_{\boldsymbol{\beta}} L_\lambda\|_2^2 = 4\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2.$$

Proof. Recall Claim 1 that

$$L(\boldsymbol{\theta}) = |\alpha - \hat{\alpha}|^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 + \sum_{k\ge 2} \hat{\sigma}_k^2 \left\| \sum_{i\in[m]} a_i \left\|\boldsymbol{w}_i\right\|_2 \overline{\boldsymbol{w}}_i^{\otimes k} - \sum_{i\in[m_*]} a_i^* \left\|\boldsymbol{w}_i^*\right\|_2 \boldsymbol{w}_i^{*\otimes k} \right\|_F^2,$$

where $\hat{\alpha} = (1/\sqrt{2\pi}) \sum_{i=1}^m a_i \left\|\boldsymbol{w}_i\right\|_2$ and $\hat{\boldsymbol{\beta}} = (1/2)\sum_{i=1}^m a_i \boldsymbol{w}_i$.

We have

$$\nabla_\alpha L_\lambda = 2(\alpha - \hat{\alpha}), \quad \nabla_{\boldsymbol{\beta}} L_\lambda = 2(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),$$

which means that

$$|\nabla_\alpha L_\lambda|^2 = 4(\alpha - \hat{\alpha})^2, \quad \|\nabla_{\boldsymbol{\beta}} L_\lambda\|_2^2 = 4\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2.$$

□

Lemma 9. *Let the ideal loss $\widetilde{L}_{1,\lambda}(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{a}^\top \sigma_{\ge 2}(\boldsymbol{W}\boldsymbol{x}) - \widetilde{y})^2] + \lambda \left\|\boldsymbol{a}\right\|_1$ that perfectly fits $\alpha, \boldsymbol{\beta}$. Given any $\left\|\boldsymbol{a}\right\|_1, \left\|\widetilde{\boldsymbol{a}}\right\|_1 = O_*(1/\lambda)$ and $\left\|\boldsymbol{w}_i\right\|_2 = 1$, we have*

$$\begin{aligned}
&|\langle \nabla_{\boldsymbol{a}} \widetilde{L}_{1,\lambda} - \nabla_{\boldsymbol{a}} L_{1,\lambda}, \boldsymbol{a} - \widetilde{\boldsymbol{a}} \rangle| \le O_*(1/\lambda)(|\alpha - \hat{\alpha}| + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2) \\
&L_{1,\lambda}(\boldsymbol{a}) \le \widetilde{L}_{1,\lambda}(\boldsymbol{a}) + |\alpha - \hat{\alpha}|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}\|_2^2 \\
&\left\|\nabla_{\boldsymbol{a}} L_{1,\lambda}\right\|_2 = O_*(\sqrt{m})
\end{aligned}$$

Proof. Using the property of Hermite polynomial in Section I, we have

$$\begin{aligned}
\nabla_{a_i} L_{1,\lambda} &= 2\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x};\boldsymbol{\theta}) - \widetilde{y})\sigma(\boldsymbol{w}_i^\top \boldsymbol{x})] + \lambda a_i \\
&= 2(\alpha - \hat{\alpha}) \left\|\boldsymbol{w}_i\right\|_2 + 2\langle \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}, \boldsymbol{w}_i \rangle + 2\mathbb{E}_{\boldsymbol{x}}[(f_{\ge 2}(\boldsymbol{x};\boldsymbol{\theta}) - \widetilde{y})\sigma(\boldsymbol{w}_i^\top \boldsymbol{x})] + \lambda a_i \\
&= 2(\alpha - \hat{\alpha}) \left\|\boldsymbol{w}_i\right\|_2 + 2\langle \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}, \boldsymbol{w}_i \rangle + \nabla_{\boldsymbol{a}} \widetilde{L}_{1,\lambda},
\end{aligned}$$

where $f_{\ge 2}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i\in[m]} a_i \sigma_{\ge 2}(\boldsymbol{w}_i^\top \boldsymbol{x})$ and $\sigma_{\ge 2}(x) = \sigma(x) - 1/\sqrt{2\pi} - x/2$ is the activation that after removing 0th and 1st order term in Hermite expansion.

We know

$$\begin{aligned}
|\langle \nabla_{\boldsymbol{a}} L_{1,\lambda} - \nabla_{\boldsymbol{a}} \widetilde{L}_{1,\lambda}, \boldsymbol{a} - \widetilde{\boldsymbol{a}} \rangle| =& |2 \sum_{i\in[m]} (a_i - \widetilde{a}_i)((\alpha - \hat{\alpha}) \left\|\boldsymbol{w}_i\right\|_2 + \langle \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}, \boldsymbol{w}_i \rangle)| \\
\le& O_*(1/\lambda)(|\alpha - \hat{\alpha}| + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2).
\end{aligned}$$

The loss bound directly follows from Claim 1.

The gradient norm bound we have

$$\|\nabla_{\boldsymbol{a}} L_{1,\lambda}\|_2^2 \le O(1) \sum_{i\in[m]} (\alpha - \hat{\alpha})^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 + O_*(1) + \lambda|a_i| = O_*(m).$$

□

D Stage 3: local convergence for regularized 2-layer neural networks

In this section we show the local convergence that loss eventually goes to 0 with and recovers teacher neurons' direction.

The results in this section only need the width $m \geq m_*$ as long as its initial loss is small.

Lemma 5 (Stage 3)**.** *Under Assumption 1,2, consider Algorithm 1 on loss (2). Given Stage 2 in Lemma 4, if the initial optimality gap $\zeta_{3,0} \leq O_*(\lambda_{3,0}^{9/5})$, weight decay λ follows the schedule of initial value $\lambda_{3,0} = O_*(1)$, and k-th epoch $\lambda_{3,k} = \lambda_{3,0}/(k\lambda_{3,0}+1)$ and stepsize $\eta_{3k} = \eta \leq O_*(\lambda_{3,k}^{-6}d^{-1})$ for all $T_{3,k} \leq t \leq T_{3,k+1}$ in epoch k, then within $K = O_*(\varepsilon^{-1/2})$ epochs and total $T_3 - T_2 = O_*(\varepsilon^{-5/2}\eta^{-1})$ time we recover the ground-truth network $L(\boldsymbol{\theta}) \leq \varepsilon$.*

The goal of each epoch is to minimize the loss L_λ with a fix λ. The lemma below shows that as long as the initial optimality gap is $O_*(\lambda^{9/5})$, then at the end of each epoch, L_λ could decrease to $O_*(\lambda^2)$. Therefore, using a slow decay of weight decay parameter λ for each epoch we could stay in the local convergence regime for each epoch and eventually recovers the target network.

Lemma 10 (Loss improve within one epoch)**.** *Suppose $|a_i^{(0)}| \leq \left\|\boldsymbol{w}_i^{(0)}\right\|_2$ for all $i \in [m]$. If $\zeta_0 \leq O_*(\lambda^{9/5})$ and $\lambda \leq O_*(1)$ and $\eta \leq O_*(\lambda^{-6}d^{-1})$, then within $O_*(\lambda^{-4}\eta^{-1})$ time the optimality gap becomes $L_\lambda - L_\lambda(\mu_\lambda^*) = O_*(\lambda^2)$.*

The above result relies on the following characterization of local landscape of regularized loss. We show the gradient is large whenever the optimality gap is large. This is the main contribution of this paper, see Section E for detailed proofs.

Lemma 6 (Gradient lower bound)**.** *Suppose the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - \min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu)$. If $\Omega_*(\lambda^2) \leq \zeta \leq O_*(\lambda^{9/5})$ and $\lambda \leq O_*(1)$, we have*

$$\|\nabla_{\boldsymbol{\theta}} L_\lambda\|_F^2 \geq \Omega_*(\zeta^4/\lambda^2).$$

In order to use the above landscape result with standard descent lemma, we also need certain smoothness condition on the loss function. We show below that this regularized loss indeed satisfies certain smoothness condition (though weaker than standard smoothness condition) to allow the convergence analysis.

Lemma 11 (Smoothness)**.** *Suppose $|a_i| \leq \|\boldsymbol{w}_i\|_2$ and $\left\|\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}]\right\|_2^2 = O_*(d)$ for all $i \in [m]$. If $\eta = O_*(1/d)$, then*

$$L_\lambda(\boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}} L_\lambda) \leq L_\lambda(\boldsymbol{\theta}) - \eta\,\|\nabla_{\boldsymbol{\theta}} L_\lambda\|_F^2 + O_*(\eta^{3/2}d^{3/2})$$

D.1 Proofs in Section D

Lemma 5 (Stage 3)**.** *Under Assumption 1,2, consider Algorithm 1 on loss (2). Given Stage 2 in Lemma 4, if the initial optimality gap $\zeta_{3,0} \leq O_*(\lambda_{3,0}^{9/5})$, weight decay λ follows the schedule of initial value $\lambda_{3,0} = O_*(1)$, and k-th epoch $\lambda_{3,k} = \lambda_{3,0}/(k\lambda_{3,0}+1)$ and stepsize $\eta_{3k} = \eta \leq O_*(\lambda_{3,k}^{-6}d^{-1})$ for all $T_{3,k} \leq t \leq T_{3,k+1}$ in epoch k, then within $K = O_*(\varepsilon^{-1/2})$ epochs and total $T_3 - T_2 = O_*(\varepsilon^{-5/2}\eta^{-1})$ time we recover the ground-truth network $L(\boldsymbol{\theta}) \leq \varepsilon$.*

Proof. Since $|a_i^{(0)}| \leq \left\|\boldsymbol{w}_i^{(0)}\right\|_2$ for all $i \in [m]$ at the beginning of Stage 3, from Lemma 12 we know they will remain hold for all epoch and all time t.

From Lemma 10 we know for epoch k it finishes within $O_*(\lambda_k^{-4}\eta^{-1})$ time and achieves $L_{\lambda_k} - L_{\lambda_k}(\mu_{\lambda_k}^*) = O_*(\lambda_k^2)$. To proceed to next epoch $k+1$, we only need to show the solution at the end of epoch k $\boldsymbol{\theta}^{(k)}$ gives the optimality gap $\zeta = O_*(\lambda_{k+1}^{9/5})$ for the next λ_{k+1}. We have

$$\begin{aligned}
L_{\lambda_{k+1}}(\boldsymbol{\theta}^{(k)}) - L_{\lambda_{k+1}}(\mu^*_{\lambda_{k+1}}) =& L(\boldsymbol{\theta}^{(k)}) - L(\mu^*_{\lambda_{k+1}}) + \frac{\lambda_{k+1}}{2}\left\|\boldsymbol{a}^{(k)}\right\|_2^2 + \frac{\lambda_{k+1}}{2}\left\|\boldsymbol{W}^{(k)}\right\|_F^2 - \lambda_{k+1}|\mu^*_{\lambda_{k+1}}|_1\\
\overset{(a)}{\le}& O_*(\lambda_k^2) + \frac{\lambda_{k+1}}{\lambda_k}\left(\frac{\lambda_k}{2}\left\|\boldsymbol{a}^{(k)}\right\|_2^2 + \frac{\lambda_k}{2}\left\|\boldsymbol{W}^{(k)}\right\|_F^2 - \lambda_k|\mu^*_{\lambda_{k+1}}|_1\right)\\
\overset{(b)}{\le}& O_*(\lambda_k^2) + \frac{\lambda_{k+1}}{\lambda_k}O_*(\lambda_k^2) + \frac{\lambda_{k+1}}{\lambda_k}(L(\mu^*_{\lambda_k}) - L(\boldsymbol{\theta}^{(k)}))\\
\overset{(c)}{\le}& O_*(\lambda_k^2) \le O_*(\lambda_{k+1}^{9/5})
\end{aligned}$$

where (a) due to Lemma 18; (b) the optimality gap at the end of epoch k is $O_*(\lambda_k^2)$; (c) due to Lemma 17. In this way, we can apply Lemma 10 again for epoch $k+1$.

From Lemma 18 we know at the end of epoch k the square loss $L(\boldsymbol{\theta}^{(k)}) = O_*(\lambda_k^2)$. Thus, to reach ε square loss, we need $\lambda_k = O_*(\varepsilon^{1/2})$, which means we need to take $O_*(\varepsilon^{-1/2})$ epoch. Since epoch k it finishes within $O_*(\lambda_k^{-4}\eta^{-1})$ time, we know the total time is at most $O_*(\varepsilon^{-5/2}\eta^{-1})$ time. □

Lemma 10 (Loss improve within one epoch)**.** *Suppose $|a_i^{(0)}| \le \left\|\boldsymbol{w}_i^{(0)}\right\|_2$ for all $i \in [m]$. If $\zeta_0 \le O_*(\lambda^{9/5})$ and $\lambda \le O_*(1)$ and $\eta \le O_*(\lambda^{-6}d^{-1})$, then within $O_*(\lambda^{-4}\eta^{-1})$ time the optimality gap becomes $L_\lambda - L_\lambda(\mu^*_\lambda) = O_*(\lambda^2)$.*

Proof. Since $|a_i^{(0)}| \le \left\|\boldsymbol{w}_i^{(0)}\right\|_2$ for all $i \in [m]$ at the beginning of current epoch, from Lemma 12 we know they will remain hold for all time t. Then combine Lemma 13 and Lemma 11 we know

$$L_\lambda(\boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}} L_\lambda) \le L_\lambda(\boldsymbol{\theta}) - \eta\left\|\nabla_{\boldsymbol{\theta}} L_\lambda\right\|_F^2 + O_*(\eta^{3/2}d^{3/2}).$$

Recall $\zeta_t = L_\lambda(\boldsymbol{\theta}^{(t)}) - L_\lambda(\mu^*_\lambda)$. Using Lemma 6 and consider the time before ζ_t reach $O_*(\lambda^2)$ we have

$$\zeta_{t+1} \le \zeta_t - \eta\Omega_*(\zeta_t^4/\lambda^2) + O_*(\eta^{3/2}d^{3/2}) \le \zeta_t - \Omega_*(\eta\zeta_t^4/\lambda^2),$$

where we use $\eta = O_*(\lambda^{-6}d^{-1})$ to be small enough.

The above recursion implies that

$$\zeta_t = O_*((t/\lambda^2 + \zeta_0)^{-1/3}).$$

Thus, within $O_*(1/\lambda^4)$ the optimality gap ζ_t reaches $O_*(\lambda^2)$. □

Lemma 12. *If we start at $|a_i^{(0)}| \le \left\|\boldsymbol{w}_i^{(0)}\right\|_2$ and $\eta = O_*(1)$, then we have $|a_i^{(t)}|^2 \le \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2$ for all $i \in [m_*]$ and all time t.*

Proof. Denote $R(\boldsymbol{x}) = f(\boldsymbol{x}) - f_*(\boldsymbol{x})$. Assume $|a_i^{(t)}|^2 - \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2 \le 0$ we have

$$\begin{aligned}
|a_i^{(t+1)}|^2 - \left\|\boldsymbol{w}_i^{(t+1)}\right\|_2^2 =& |a_i^{(t)} - \eta\nabla_{a_i}L_\lambda(\boldsymbol{\theta}^{(t)})|^2 - \left\|\boldsymbol{w}_i^{(t)} - \eta\nabla_{\boldsymbol{w}_i}L_\lambda(\boldsymbol{\theta}^{(t)})\right\|_2^2 \\
=& |a_i^{(t)}|^2 - \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2 + \eta^2|\nabla_{a_i}L_\lambda(\boldsymbol{\theta}^{(t)})|^2 - \eta^2\left\|\nabla_{\boldsymbol{w}_i}L_\lambda(\boldsymbol{\theta}^{(t)})\right\|_2^2 \\
=& |a_i^{(t)}|^2 - \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2 + \eta^2|2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma(\boldsymbol{w}_i^{(t)\top}\boldsymbol{x})] + \lambda a_i^{(t)}|^2 - \eta^2 \Big\| 2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})a_i^{(t)}\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}] + \lambda\boldsymbol{w}_i^{(t} \\
\le& |a_i^{(t)}|^2 - \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2 + \eta^2\left(\left\|\boldsymbol{w}_i^{(t)}\right\|_2^2 |2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})]|^2 - |a_i^{(t)}|^2\left\|2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}]\right\|_2^2\right. \\
&\left. + \lambda^2|a_i^{(t)}|^2 - \lambda^2\left\|\boldsymbol{w}_i^{(t)}\right\|_2^2\right) \\
\overset{\text{(a)}}{\le}& \left(|a_i^{(t)}|^2 - \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2\right)\left(1 + \eta^2\lambda^2 - \eta^2\left\|2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}]\right\|_2^2\right) \\
\overset{\text{(b)}}{\le}& \left(|a_i^{(t)}|^2 - \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2\right)\left(1 - \eta^2\left\|2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}]\right\|_2^2\right) \\
\overset{\text{(c)}}{\le}& 0,
\end{aligned}$$

where (a) due to $|2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})]|^2 \le \left\|2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}]\right\|_2^2$; (b) due to $|a_i^{(t)}|^2 - \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2 \le 0$; (c) we use $\left\|2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})]\boldsymbol{x}\right\|_2^2 = O_*(1)$ from Lemma 13 and $\eta = O_*(1)$.

Therefore, we can see that $|a_i^{(t)}|^2 - \left\|\boldsymbol{w}_i^{(t)}\right\|_2^2 \le 0$ remains for all t. □

Lemma 11 (Smoothness)**.** *Suppose $|a_i| \le \|\boldsymbol{w}_i\|_2$ and $\left\|\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}]\right\|_2^2 = O_*(d)$ for all $i \in [m]$. If $\eta = O_*(1/d)$, then*

$$L_\lambda(\boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}L_\lambda) \le L_\lambda(\boldsymbol{\theta}) - \eta\left\|\nabla_{\boldsymbol{\theta}}L_\lambda\right\|_F^2 + O_*(\eta^{3/2}d^{3/2})$$

Proof. Denote $R_{\boldsymbol{\theta}}(\boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) - f_*(\boldsymbol{x})$ to denote the dependency on $\boldsymbol{\theta}$. For simplicity, we will use $\widetilde{\nabla}_{\boldsymbol{\theta}} = -\eta\nabla_{\boldsymbol{\theta}}L_\lambda$ and same for others. Since $\left\|\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}]\right\|_2^2 = O_*(d)$, we know $|\widetilde{\nabla}_{a_i}| = O_*(\eta\|\boldsymbol{w}_i\|_2 d)$ and $\left\|\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2 = O_*(\eta|a_i|d)$

We have

$$\begin{aligned}
& L_\lambda(\boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}) - L_\lambda(\boldsymbol{\theta}) + \eta\left\|\nabla_{\boldsymbol{\theta}}\right\|_F^2 \\
=& L_\lambda(\boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}) - L_\lambda(\boldsymbol{\theta}) - \langle\nabla_{\boldsymbol{\theta}}, -\eta\nabla_{\boldsymbol{\theta}}\rangle \\
=& \mathbb{E}_{\boldsymbol{x}}[R_{\boldsymbol{\theta}+\widetilde{\nabla}_{\boldsymbol{\theta}}}(\boldsymbol{x})^2] + \frac{\lambda}{2}\left\|\boldsymbol{a} + \widetilde{\nabla}_{\boldsymbol{a}}\right\|_2^2 + \frac{\lambda}{2}\left\|\boldsymbol{W} + \widetilde{\nabla}_{\boldsymbol{W}}\right\|_F^2 - \mathbb{E}_{\boldsymbol{x}}[R_{\boldsymbol{\theta}}(\boldsymbol{x})^2] - \frac{\lambda}{2}\|\boldsymbol{a}\|_2^2 - \frac{\lambda}{2}\|\boldsymbol{W}\|_F^2 \\
& - \sum_{i\in[m]}\mathbb{E}_{\boldsymbol{x}}[R_{\boldsymbol{\theta}}(\boldsymbol{x})\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\widetilde{\nabla}_{a_i}] - \sum_{i\in[m]}\mathbb{E}_{\boldsymbol{x}}[R_{\boldsymbol{\theta}}(\boldsymbol{x})a_i\sigma'(\boldsymbol{w}_i^\top\boldsymbol{x})\boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{w}_i}] - \mathbb{E}_{\boldsymbol{x}}[R_{\boldsymbol{\theta}}(\boldsymbol{x})\widetilde{\nabla}_\alpha] - \mathbb{E}_{\boldsymbol{x}}[R_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{\beta}}] \\
=& \underbrace{\mathbb{E}_{\boldsymbol{x}}[(R_{\boldsymbol{\theta}+\widetilde{\nabla}_{\boldsymbol{\theta}}}(\boldsymbol{x}) - R_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]}_{(I)} \\
& + \underbrace{2\,\mathbb{E}_{\boldsymbol{x}}\left[R_{\boldsymbol{\theta}}(\boldsymbol{x})\left(R_{\boldsymbol{\theta}+\widetilde{\nabla}_{\boldsymbol{\theta}}}(\boldsymbol{x}) - R_{\boldsymbol{\theta}}(\boldsymbol{x}) - \sum_{i\in[m]}\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\widetilde{\nabla}_{a_i} - \sum_{i\in[m]}a_i\sigma'(\boldsymbol{w}_i^\top\boldsymbol{x})\boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{w}_i} - \widetilde{\nabla}_\alpha - \boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{\beta}}\right)\right]}_{(II)}.
\end{aligned}$$

We are going to bound (I) and (II) one by one.

For (I), we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}[(R_{\boldsymbol{\theta}+\widetilde{\nabla}_{\boldsymbol{\theta}}}(\boldsymbol{x}) - R_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
=&\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i\in[m]}(a_i+\widetilde{\nabla}_{a_i})\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x}) - a_i\sigma(\boldsymbol{w}_i^\top\boldsymbol{x}) + \widetilde{\nabla}_\alpha + \boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{\beta}}\right)^2\right] \\
\le& \underbrace{2\,\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i\in[m]}(a_i+\widetilde{\nabla}_{a_i})\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x}) - a_i\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\right)^2\right]}_{(I.i)} + \underbrace{2\,\mathbb{E}_{\boldsymbol{x}}\left[\left(\widetilde{\nabla}_\alpha + \boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{\beta}}\right)^2\right]}_{I.ii}
\end{aligned}
$$

For (I.i), we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i\in[m]}(a_i+\widetilde{\nabla}_{a_i})\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x}) - a_i\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\right)^2\right] \\
\le& 2\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i\in[m]}\widetilde{\nabla}_{a_i}\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x})\right)^2\right] + 2\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i\in[m]}a_i\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x}) - a_i\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\right)^2\right] \\
\le& 2\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i\in[m]}|\widetilde{\nabla}_{a_i}||(\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x}|\right)^2\right] + 2\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i\in[m]}|a_i||\widetilde{\nabla}_{\boldsymbol{w}_i}^\top\boldsymbol{x}|\right)^2\right] \\
\overset{\text{(a)}}{\le}& O(1)\left(\sum_{i\in[m]}|\widetilde{\nabla}_{a_i}|\left\|\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2\right)^2 + O(1)\left(\sum_{i\in[m]}|a_i|\left\|\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2\right)^2 \\
\overset{\text{(b)}}{\le}& O_*(d^2)\left(\sum_{i\in[m]}\eta\|\boldsymbol{w}_i\|_2^2 + \eta^2|a_i|\|\boldsymbol{w}_i\|_2 d\right)^2 + O_*(d^2)\left(\sum_{i\in[m]}\eta a_i^2\right)^2 \\
\overset{\text{(c)}}{\le}& O_*(\eta^2 d^2),
\end{aligned}
$$

where (a) we use Lemma 14; (b) recall $|\widetilde{\nabla}_{a_i}| = O_*(\eta\|\boldsymbol{w}_i\|_2)$ and $\left\|\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2 = O_*(\eta|a_i|)$; (c) $\|\boldsymbol{a}\|, \|\boldsymbol{W}\|_F, \sum_{i\in[m]}|a_i|\|\boldsymbol{w}_i\|_2 = O_*(1)$ from Lemma 26 and Lemma 18.

For (I.ii), we have

$$
\mathbb{E}_{\boldsymbol{x}}\left[\left(\widetilde{\nabla}_\alpha + \boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{\beta}}\right)^2\right] \le O(|\widetilde{\nabla}_\alpha|^2 + \left\|\widetilde{\nabla}_{\boldsymbol{\beta}}\right\|_2^2) = O_*(\eta^2),
$$

where we use Lemma 18.

Combine (I.i) and (I.ii) we know (I)=$O_*(\eta^2 d^2)$.

For (II), we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}\left[R_{\boldsymbol{\theta}}(\boldsymbol{x})\left(R_{\boldsymbol{\theta}+\widetilde{\nabla}_{\boldsymbol{\theta}}}(\boldsymbol{x}) - R_{\boldsymbol{\theta}}(\boldsymbol{x}) - \sum_{i\in[m]}\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\widetilde{\nabla}_{a_i} - \sum_{i\in[m]}a_i\sigma'(\boldsymbol{w}_i^\top\boldsymbol{x})\boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{w}_i} - \widetilde{\nabla}_\alpha - \boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{\beta}}\right)\right] \\
=&\mathbb{E}_{\boldsymbol{x}}\left[R_{\boldsymbol{\theta}}(\boldsymbol{x})\left(\sum_{i\in[m]}\underbrace{(a_i+\widetilde{\nabla}_{a_i})\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x}) - a_i\sigma(\boldsymbol{w}_i^\top\boldsymbol{x}) - \sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\widetilde{\nabla}_{a_i} - a_i\sigma'(\boldsymbol{w}_i^\top\boldsymbol{x})\boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{w}_i}}_{I_i(\boldsymbol{x})}\right)\right] \\
\le&\sum_{i\in[m]}\|R_{\boldsymbol{\theta}}\|\,\|I_i\|
\end{aligned}
$$

For $I_i(\boldsymbol{x})$, we have

$$\begin{aligned}
\|I_i\|_2^2 =&\mathbb{E}_{\boldsymbol{x}}\left[\left((a_i+\widetilde{\nabla}_{a_i})\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x})-a_i\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})-\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\widetilde{\nabla}_{a_i}-a_i\sigma'(\boldsymbol{w}_i^\top\boldsymbol{x})\boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{w}_i}\right)^2\right]\\
\le&\mathbb{E}_{\boldsymbol{x}}\left[2\left(\widetilde{\nabla}_{a_i}(\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x})-\sigma(\boldsymbol{w}_i^\top\boldsymbol{x}))\right)^2+2\left(a_i(\sigma((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x})-\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})-\sigma'(\boldsymbol{w}_i^\top\boldsymbol{x})\boldsymbol{x}^\top\widetilde{\nabla}_{\boldsymbol{w}_i})\right)^2\right]\\
\le&2\underbrace{\mathbb{E}_{\boldsymbol{x}}\left[|\widetilde{\nabla}_{a_i}|^2|\widetilde{\nabla}_{\boldsymbol{w}_i}^\top\boldsymbol{x}|^2\right]}_{(II.i)}+2a_i^2\underbrace{\mathbb{E}_{\boldsymbol{x}}\left[|(\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x}|^2(\sigma'((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x})-\sigma'(\boldsymbol{w}_i^\top\boldsymbol{x}))^2\right]}_{(II.ii)}
\end{aligned}$$

For (II.i), recall $|\widetilde{\nabla}_{a_i}|=O_*(\eta\left\|\boldsymbol{w}_i\right\|_2)$ and $\left\|\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2=O_*(\eta|a_i|)$ we have

$$\mathbb{E}_{\boldsymbol{x}}\left[|\widetilde{\nabla}_{a_i}|^2|\widetilde{\nabla}_{\boldsymbol{w}_i}^\top\boldsymbol{x}|^2\right]\le|\widetilde{\nabla}_{a_i}|^2\left\|\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|^2=O_*(\eta^4|a_i|^2\left\|\boldsymbol{w}_i\right\|_2^2d^4).$$

For (II.ii), we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[|\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i}^\top\boldsymbol{x}|^2(\sigma'((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x})-\sigma'(\boldsymbol{w}_i^\top\boldsymbol{x}))^2\right]=&\mathbb{E}_{\boldsymbol{x}}\left[|(\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x}|^2\mathbb{1}_{\mathrm{sign}((\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i})^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_i^\top\boldsymbol{x})}\right]\\
\le&O(\left\|\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2^2\delta^3),
\end{aligned}$$

where $\delta=\angle(\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i},\boldsymbol{w}_i)$ is the angle between $\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i}$ and $\boldsymbol{w}_i$. Since $\left\|\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2=O_*(\eta|a_i|d)=O_*(\eta\left\|\boldsymbol{w}_i\right\|_2d)$, we know $\delta=O(\left\|\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|)$ given $\eta=O_*(1/d)$ to be small enough.

Combine (II.i) and (II.ii) we have

$$\|I_i\|_2^2\le O_*(\eta^4a_i^2\left\|\boldsymbol{w}_i\right\|_2^2d^4)+O(a_i^2\left\|\boldsymbol{w}_i+\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2^2\left\|\widetilde{\nabla}_{\boldsymbol{w}_i}\right\|_2^3)\le O_*(\eta^3a_i^2\left\|\boldsymbol{w}_i\right\|_2^2d^3).$$

This implies

$$(II)\le\sum_{i\in[m]}O_*(\eta^{3/2}a_i\left\|\boldsymbol{w}_i\right\|_2d^{3/2})=O_*(\eta^{3/2}d^{3/2}).$$

Finally, combing (I) and (II) we have

$$L_\lambda(\boldsymbol{\theta}-\eta\nabla_{\boldsymbol{\theta}})-L_\lambda(\boldsymbol{\theta})+\eta\left\|\nabla_{\boldsymbol{\theta}}\right\|_F^2=O_*(\eta^{3/2}d^{3/2}).$$

□

D.2 Technical Lemma

Lemma 13. *We have* $\left\|\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma'(\overline{\boldsymbol{w}}_i^{(t)\top}\boldsymbol{x})\boldsymbol{x}]\right\|_2^2=O_*(d)$

Proof. It is easy to see given $\|R\|=O_*(1)$.

□

Lemma 14 (Lemma D.4 in [28])**.** *Consider $\alpha_i\in\mathbb{R}^d$ for $i\in[n]$. We have*

$$\mathbb{E}_{x\sim N(0,I)}\left[\left(\sum_{i=1}^n|\alpha_i^\top x|\right)^2\right]\le c_0\left(\sum_{i=1}^n\|\alpha_i\|\right)^2,$$

where c_0 is a constant.

E Local landscape of population loss

In this section, we are going to show Lemma 6 that characterizing the population local landscape with a fixed λ by giving the lower bound of gradient. We start by identifying the structure of (approximated) solution of a closely-related problem in Section E.1:

$$\min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu) := L(\mu) + \lambda|\mu|_1 := \mathbb{E}_{\boldsymbol{x},\widetilde{y}}[(f_\mu(\boldsymbol{x}) - \widetilde{y})^2] + \lambda|\mu|_1 = \mathbb{E}_{\boldsymbol{x}}\left[\left(\int_{\boldsymbol{w}} \sigma_{\geq 2}(\boldsymbol{w}^\top\boldsymbol{x})\mathrm{d}\,\mu - \mu_*\right)^2\right] + \lambda|\mu|_1, \tag{3}$$

where $\mathcal{M}(\mathbb{S}^{d-1})$ is the measure space over unit sphere $\mathbb{S}^{d-1}$ and $\sigma_{\geq 2}(x) = \sigma(x) - 1/\sqrt{2\pi} - x/2$ is the activation that after removing 0th and 1st order term in Hermite expansion. Note that when μ represents a finite-wdith network, we have $\mu = \sum_{i\in[m]} a_i \|\boldsymbol{w}_i\|_2 \delta_{\overline{\boldsymbol{w}}_i}$ is a empirical measure over the neurons.

We call (3) as the ideal loss because the original problem (2) would become the above (3) when we balance the norms ($\|\boldsymbol{w}_i\|_2 = |a_i|$), perfectly fit α, β and relax the finite-width constraints to allow infinite-width. This is why we slightly abused the notation to use L_λ in both (2) and (3).

Then, in Section E.3 we will use the solution structure to construct descent direction that are positively correlated with gradient and also handle the case when norms are not balanced or α, β are not fitted well.

Notation Denote the optimality gap between the loss at μ and the optimal distribution μ_λ^* as

$$\zeta(\mu) := L_\lambda(\mu) - L_\lambda(\mu_\lambda^*),$$

where μ_λ^* is the optimal measure that minimize (3). For simplicity denote $\widetilde{a}_i = a_i \|\boldsymbol{w}_i\|_2$ so that $|\mu|_1 = \|\widetilde{\boldsymbol{a}}\|_1$ when $\mu = \sum_{i\in[m]} a_i \|\boldsymbol{w}_i\|_2 \delta_{\overline{\boldsymbol{w}}_i}$. Often we use $\zeta_t = \zeta(\mu_t)$ to denote the optimality gap at time t and just ζ for simplicity. We slightly abuse the notation to also use $\zeta = L_\lambda(\theta) - L_\lambda(\mu_\lambda^*)$. Finally denote $\mu^* = \sum_{i\in[m_*]} a_i^* \delta_{\boldsymbol{w}_i^*}$ (assuming $\|\boldsymbol{w}_i^*\|_2 = 1$) so that $f_{\mu^*}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{w}\sim\mu^*}[\sigma_{\geq 2}(\boldsymbol{w}^\top\boldsymbol{x})]$.

We will group the neurons (i.e., partitioning $\mathbb{S}^{d-1}$) based on their distance to the closest teacher neurons: denote $\mathcal{T}_i = \{\boldsymbol{w} : \angle(\boldsymbol{w}, \boldsymbol{w}_i^*) \leq \angle(\boldsymbol{w}, \boldsymbol{w}_j^*)$ for any $j \neq i\}$ so that $\cup_i \mathcal{T}_i = \mathbb{S}^{d-1}$.

We will use O_*, Ω_*, Θ_* to hide $\mathrm{poly}(r, m_*, \Delta, a_{\min}, \|\boldsymbol{a}_*\|_1)$, the polynomial dependency on relevant parameters of target f_*.

E.1 Structure of the ideal loss solution

In this section, we will focus on the structure of approximated solution for the ℓ_1 regularized regression problem (3).

In the rest of this section, we will first introduce the idea of non-degenerate dual certificate and then use it as a tool to characterize the structure of the solutions. The proofs are deferred to Section G.

E.1.1 Non-degenerate dual certificate

We first introduce the notion of non-degenerate dual certificate similar as in [26] but slightly adapted for fit our need. Roughly speaking, η acts as a certificate of the true solution because $|\eta(\boldsymbol{w})| \leq 1$ and $|\eta(\boldsymbol{w})| = 1$ only if $\boldsymbol{w} = \boldsymbol{w}_i^*$ at some ground-truth direction. The non-degenerate means that η decays fast around each $\boldsymbol{w}_i^*$ so that one can simply look at η to identify all the ground-truth directions.

Definition 1 (Non-degenerate dual certificate)**.** *η is called a non-degenerate dual certificate if there exists $p(\boldsymbol{x})$ such that $\eta(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{x})\sigma_{\geq 2}(\boldsymbol{w}^\top\boldsymbol{x})]$ for $\boldsymbol{w} \in \mathbb{S}^{d-1}$ and*

(i) $\eta(\boldsymbol{w}_i^*) = \mathrm{sign}(a_i^*)$ *for* $i = 1, \ldots, m_*$

(ii) $|\eta(\boldsymbol{w})| \leq 1 - \rho_\eta \delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2$ *if* $\boldsymbol{w} \in \mathcal{T}_i$*, where* $\delta(\boldsymbol{w}, \boldsymbol{w}_i^*) = \angle(\boldsymbol{w}, \boldsymbol{w}_i^*)$

We first show that there exist such non-degenerate dual certificate. More discussion and a detailed proof are deferred to Section F.

Lemma 15. *There exists a non-degenerate dual certificate* $\eta = \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{x})\sigma_{\geq 2}(\boldsymbol{w}^\top \boldsymbol{x})]$ *with* $\rho_\eta = \Theta(1)$ *and* $\|p\|_2 \leq \operatorname{poly}(m_*, \Delta)$

The following lemma gives the properties that will be used in the later proofs: the non-degenerate dual certificate η allows us to capture the gap between the current position μ and the target μ^*.

Lemma 16. *Given a non-degenerate dual certificate* η*, then*

(i) For any measure $\mu \in \mathcal{M}(\mathbb{S}^{d-1})$, $|\langle \eta, \mu \rangle| \leq |\mu|_1 - \rho_\eta \sum_{i \in [m_*]} \int_{\mathcal{T}_i} \delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \, \mathrm{d}|\mu|(\boldsymbol{w})$.

(ii) $\langle \eta, \mu^* \rangle = |\mu^*|_1$

(iii) $\langle \eta, \mu - \mu^* \rangle = \langle p, f_\mu - f_{\mu^*} \rangle$*, where* $f_\mu(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{w} \sim \mu}[\sigma_{\geq 2}(\boldsymbol{w}^\top \boldsymbol{x})]$*. Thus,* $|\langle \eta, \mu - \mu^* \rangle| \leq \|p\|_2 \sqrt{L(\mu)}$.

E.1.2 Properties of μ_λ^*

Given the non-degenerate dual certificate η, we now are ready to identify several useful properties of μ_λ^*. The lemma below essentially says that μ_λ^* is similar to μ^* in the sense that most of the norm are concentrated in the ground-truth direction and the square loss is small. The proof relies on comparing μ_λ^* with μ^* using the optimality conditions.

Lemma 17. *We have the following hold*

(i) $|\mu_*|_1 - \lambda \|p\|_2^2 \leq |\mu_\lambda^*|_1 \leq |\mu^*|_1 = \|\boldsymbol{a}^*\|_1$

(ii) $L(\mu_\lambda^*) \leq \lambda^2 \|p\|_2^2 = O_*(\lambda^2)$

(iii) $\sum_{i \in [m_*]} \int_{\mathcal{T}_i} \delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \, \mathrm{d}|\mu_\lambda^*|(\boldsymbol{w}) \leq \lambda \|p\|_2^2 / \rho_\eta = O_*(\lambda)$

E.1.3 Properties of μ with optimality gap ζ

We now characterize the structure of μ when the optimality gap is ζ. The proof mostly relies on comparing μ with μ_λ^* and the structure of μ_λ^* in previous section.

The following lemma shows the square loss is bounded by the optimality gap and norms are always bounded. Note that the conditions are true under Lemma 6.

Lemma 18. *Recall the optimality gap* $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$*. Then, the following holds:*

(i) $L(\mu) \leq 5\lambda^2 \|p\|^2 + 4\zeta = O_*(\lambda^2 + \zeta)$.

(ii) if $\zeta \leq \lambda |\mu^*|_1$ *and* $\lambda \leq |\mu^*|_1 / \|p\|_2^2$*, then* $|\mu|_1 \leq 3|\mu^*|_1 = 3\|\boldsymbol{a}^*\|_1$.

The following two lemma characterize the structure of μ using the fact that the square loss is small in previous lemma. The lemma below says that the total norm of far away neuron is small.

Lemma 19. *Recall the optimality gap* $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$*. Then, we have*

$$\sum_{i \in [m_*]} \int_{\mathcal{T}_i} \delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \, \mathrm{d}|\mu|(\boldsymbol{w}) \leq (\zeta/\lambda + 2\lambda \|p\|_2^2)/\rho_\eta = O_*(\zeta/\lambda + \lambda).$$

In particular, when $\mu = \sum_{i \in [m]} a_i \|\boldsymbol{w}_i\|_2 \, \delta_{\overline{\boldsymbol{w}}_i}$ *represents finite number of neurons, we have*

$$\sum_{i \in [m_*]} \sum_{j \in \mathcal{T}_i} |a_j| \, \|\boldsymbol{w}_j\|_2 \, \delta_j^2 \leq (\zeta/\lambda + 2\lambda \|p\|_2^2)/\rho_\eta = O_*(\zeta/\lambda + \lambda),$$

where $\delta_j = \angle(\boldsymbol{w}_j, \boldsymbol{w}_i^*)$ *for* $j \in \mathcal{T}_i$.

The lemma below shows there are neurons close to the teacher neurons once the gap is small. The proof idea is similar to Section 5.3 in Zhou et al. [28] that use test function to lower bound the loss.

Lemma 20. *Under Lemma 6, if the Hermite coefficient of* σ *decays as* $|\hat{\sigma}_k| = \Theta(k^{-c_\sigma})$ *with some constant* $c_\sigma > 0$*, then the total mass near each target direction is large, i.e.,* $\mu(\mathcal{T}_i(\delta)) \operatorname{sign}(a_i^*) \geq$

$|a_i^*|/2$ for all $i \in [m_*]$ and any $\delta_{close} \geq \widetilde{\Omega}\left((\frac{L(\mu)}{a_{\min}^2})^{1/(4c_\sigma-2)}\right)$ with large enough hidden constant. In particular, for σ is ReLU or absolute function, $\delta_{close} \geq \widetilde{\Omega}\left((\frac{L(\mu)}{a_{\min}^2})^{1/3}\right)$.

As a corollary, if the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$, then $\delta_{close} \geq \widetilde{\Omega}_*\left((\zeta+\lambda^2)^{1/(4c_\sigma-2)}\right)$ and for ReLU or absolute $\delta_{close} \geq \widetilde{\Omega}_*\left((\zeta+\lambda^2)^{1/3}\right)$.

E.1.4 Residual decomposition and average neuron

In this section, we introduce the residual decomposition and average neuron as in [28] that will be used when proving the existence of descent direction.

Denote the decomposition $R(\boldsymbol{x}) = f_\mu(\boldsymbol{x}) - f_{\mu^*}(\boldsymbol{x}) = R_1(\boldsymbol{x}) + R_2(\boldsymbol{x}) + R_3(\boldsymbol{x})$ (this can be directly verified noticing that $\sigma_{\geq 2}(x) = |x|/2 - 1/\sqrt{2\pi}$),

$$\begin{aligned} R_1(\boldsymbol{x}) &= \frac{1}{2}\sum_{i\in[m_*]}\left(\sum_{j\in\mathcal{T}_i} a_j\boldsymbol{w}_j - \boldsymbol{w}_i^*\right)^\top \boldsymbol{x}\,\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x}),\\ R_2(\boldsymbol{x}) &= \frac{1}{2}\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} a_j\boldsymbol{w}_j^\top\boldsymbol{x}(\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x}) - \mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})), \qquad (4)\\ R_3(\boldsymbol{x}) &= \frac{1}{\sqrt{2\pi}}\left(\sum_{i\in[m_*]} a_i^*\left\|\boldsymbol{w}_i^*\right\|_2 - \sum_{i\in[m]} a_i\left\|\boldsymbol{w}_i\right\|_2\right). \end{aligned}$$

In the following we characterize R_1, R_2, R_3 separately. In Lemma 21 we relate R_1 with the average neuron. In Lemma 22 and Lemma 23 we bound R_2 and R_3 respectively.

Lemma 21 (Zhou et al. [28], Lemma 11)**.** $\left\|R_1\right\|_2^2 = \Omega(\Delta^3/m_*^3)\sum_{i\in[m_*]}\left\|\sum_{j\in\mathcal{T}_i} a_j\boldsymbol{w}_j - \boldsymbol{w}_i^*\right\|_2^2$.

Lemma 22. *Under Lemma 6, recall the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$. Then*

$$\left\|R_2\right\|_2^2 = O_*((\zeta/\lambda+\lambda)^{3/2}).$$

Lemma 23. *Under Lemma 6 and recall the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$. If $\hat{\sigma}_0 = 0$ and $\hat{\sigma}_k > 0$ with some $k = \Theta((1/\Delta^2)\log(\zeta/\left\|\boldsymbol{a}_*\right\|_1))$, then*

$$\left\|R_3\right\|_2 = \widetilde{O}_*((\zeta+\lambda^2)^{1/2}/\hat{\sigma}_k + (\zeta/\lambda+\lambda) + \zeta).$$

Now we are ready to bound the difference between average neuron with its corresponding ground-truth neuron.

Lemma 24. *Under Lemma 6, recall the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$. Then for any $i \in [m_*]$, $\zeta = \Omega(\lambda^2)$ and $\zeta, \lambda \leq 1/\mathrm{poly}(m_*, \Delta, \left\|\boldsymbol{a}_*\right\|_1)$*

$$\left\|\sum_{j\in\mathcal{T}_i} a_j\boldsymbol{w}_j - \boldsymbol{w}_i^*\right\|_2 \leq \left(\sum_{i\in[m_*]}\left\|\sum_{j\in\mathcal{T}_i} a_j\boldsymbol{w}_j - \boldsymbol{w}_i^*\right\|_2^2\right)^{1/2} = O_*((\zeta/\lambda)^{3/4}).$$

E.2 From ideal loss solution to real loss solution

In previous section, we consider the ideal loss solution that assumes the norms are perfectly balanced ($|a_i| = \left\|\boldsymbol{w}_i\right\|_2$) and $\alpha, \boldsymbol{\beta}$ are perfectly fitted. However, during the training we are not able to guarantee achieve these exactly but only approximately. This section is devoted to show that the results in previous section still hold though the conditions are only approximately satisfied. Recall that the original loss

$$L_\lambda(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \frac{\lambda}{2}\left\|\boldsymbol{a}\right\|_2^2 + \frac{\lambda}{2}\left\|\boldsymbol{W}\right\|_F^2$$

so that when norm are balanced and $\alpha, \boldsymbol{\beta}$ are perfectly fitted, $L_\lambda(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \lambda \sum_i |a_i| \|\boldsymbol{w}_i\|_2 = L_\lambda(\mu)$.

The lemma below shows that the properties of ideal loss solution in previous section still hold for the solution of original loss, when $\alpha, \boldsymbol{\beta}$ are approximately fitted.

Lemma 25. *Given any $\boldsymbol{\theta} = (\boldsymbol{a}, \boldsymbol{W}, \alpha, \boldsymbol{\beta})$ satisfying $|\alpha - \hat{\alpha}|^2 = O(\zeta)$, $\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 = O(\zeta)$, where $\hat{\alpha} = (1/\sqrt{2\pi}) \sum_{i=1}^m a_i \|\boldsymbol{w}_i\|_2$ and $\hat{\boldsymbol{\beta}} = (1/2) \sum_{i=1}^m a_i \boldsymbol{w}_i$. Let its corresponding balanced version $\boldsymbol{\theta}_{bal} = (\boldsymbol{a}_{bal}, \boldsymbol{W}_{bal}, \alpha_{bal}, \boldsymbol{\beta}_{bal})$ as $a_{bal,i} = \operatorname{sign}(a_i)\sqrt{|a_i| \|\boldsymbol{w}_i\|_2}$, $\boldsymbol{w}_{bal,i} = \overline{\boldsymbol{w}}_i \sqrt{|a_i| \|\boldsymbol{w}_i\|_2}$, $\alpha_{bal} = \hat{\alpha}$ and $\boldsymbol{\beta}_{bal} = \hat{\boldsymbol{\beta}}$. Then, we have*

$$L_\lambda(\boldsymbol{\theta}) - L_\lambda(\boldsymbol{\theta}_{bal}) = |\alpha - \hat{\alpha}|^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 + \frac{\lambda}{2} \sum_{i \in [m]} (|a_i| - \|\boldsymbol{w}_i\|_2)^2 \geq 0.$$

Moreover, let the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*)$, we have results in Lemma 18, Lemma 19, Lemma 20, Lemma 21, Lemma 22, Lemma 23 and Lemma 24 still hold for $L_\lambda(\boldsymbol{\theta})$, with the change of R_3 in (4) as

$$R_3(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}} \left(\sum_{i \in [m_*]} a_i^* \|\boldsymbol{w}_i^*\|_2 - \sum_{i \in [m]} a_i \|\boldsymbol{w}_i\|_2 \right) + \alpha - \hat{\alpha} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{x}.$$

The following lemma shows the norm remains bounded.

Lemma 26. *Under Lemma 6, suppose optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*)$. Then $\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{W}\|_F^2 \leq 3\|\boldsymbol{a}_*\|_1$.*

E.3 Descent direction

In this section, we show that there is a descent direction as long as the optimality gap is small until it reaches $O(\lambda^2)$. We will assume $\zeta = \Omega(\lambda^2)$ in this section for simplicity.

We first recall the lemma appears previously that shows gradient is always large whenever $\alpha, \boldsymbol{\beta}$ are not fitted well.

Lemma 8 (Descent direction, α and $\boldsymbol{\beta}$)**.** *We have*

$$|\nabla_\alpha L_\lambda|^2 = 4(\alpha - \hat{\alpha})^2, \quad \|\nabla_{\boldsymbol{\beta}} L_\lambda\|_2^2 = 4\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2.$$

We then show that if norms are not balanced or norm cancellation happens for neurons with similar direction, then one can always adjust the norm to decrease the loss due to the regularization term.

Lemma 27 (Descent direction, norm balance)**.** *We have*

$$\begin{aligned}\sum_i \sum_{j \in T_i} \left|\langle \nabla_{a_j} L_\lambda, -a_j \rangle + \langle \nabla_{\boldsymbol{w}_j} L_\lambda, \boldsymbol{w}_j \rangle\right| &= \lambda \sum_{i \in [m_*]} \left|a_i^2 - \|\boldsymbol{w}_i\|_2^2\right| \\ &\geq \max\left\{ \lambda \left| \|\boldsymbol{a}\|_2^2 - \|\boldsymbol{W}\|_F^2 \right|, \lambda \sum_{i \in [m_*]} (|a_i| - \|\boldsymbol{w}_i\|_2)^2 \right\}\end{aligned}$$

Lemma 28 (Descent direction, norm cancellation)**.** *Under Lemma 6, suppose the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*)$. For any $\boldsymbol{w}_i^*$, consider δ_{sign} such that $\delta_{close} < \delta_{\mathrm{sign}} = O(\lambda/\zeta^{1/2})$ with small enough hidden constant (δ_{close} defined in Lemma 20), then*

$$\sum_{s \in \{+,-\}} \sum_{j \in T_{i,s}(\delta_{\mathrm{sign}})} \left\langle \nabla_{a_j} L_\lambda, \frac{\operatorname{sign}(a_j)|a_j|}{\sum_{j \in T_{i,s}(\delta_{\mathrm{sign}})} |a_j| \|\boldsymbol{w}_j\|_2} \right\rangle + \left\langle \nabla_{\boldsymbol{w}_j} L_\lambda, \frac{\boldsymbol{w}_j}{\sum_{j \in T_{i,s}(\delta_{\mathrm{sign}})} |a_j| \|\boldsymbol{w}_j\|_2} \right\rangle = \Omega(\lambda).$$

where $T_{i,+}(\delta_{\mathrm{sign}}) = \{j \in T_i : \delta(\boldsymbol{w}_j, \boldsymbol{w}_i^*) \leq \delta_{\mathrm{sign}}, \operatorname{sign}(a_j) = \operatorname{sign}(a_i^*)\}$, $T_{i,-}(\delta_{\mathrm{sign}}) = \{j \in T_i : \delta(\boldsymbol{w}_j, \boldsymbol{w}_i^*) \leq \delta_{\mathrm{sign}}, \operatorname{sign}(a_j) \neq \operatorname{sign}(a_i^*)\}$ are the set of neurons that close to $\boldsymbol{w}_i^*$ with/without same sign of a_i^*.

As a result,

$$\|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2 \geq \lambda^2 \sum_{j \in T_{i,-}(\delta_{\mathrm{sign}})} |a_j| \|\boldsymbol{w}_j\|_2$$

Now given the above lemmas, it suffices to consider the remaining case that $\alpha, \boldsymbol{\beta}$ are well fitted, norms are balanced and no cancellation. In this case, the loss landscape is roughly the same as the ideal loss (3) from Lemma 25. Thus, we could leverage these detailed characterization of the solution (far-away neurons are small and average neuron is close to corresponding ground-truth neuron) to construct descent direction.

Lemma 29 (Descent direction)**.** *Under Lemma 6, suppose the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*)$. Suppose*

(i) norms are (almost) balanced: $|\,\|\boldsymbol{W}\|_F^2 - \|\boldsymbol{a}\|_2^2\,| \le \zeta/\lambda$, $\sum_{i\in[m]}(|a_j| - \|\boldsymbol{w}_j\|_2)^2 = O_*(\zeta^2/\lambda^2)$

(ii) (almost) no norm cancellation: consider all neurons $\boldsymbol{w}_j$ that are δ_{sign}-close w.r.t. teacher neuron $\boldsymbol{w}_i^*$ but has a different sign, i.e., $\operatorname{sign}(a_j) \ne \operatorname{sign}(a_i^*)$ with $\delta_{\text{sign}} = \Theta_*(\lambda/\zeta^{1/2})$, we have $\sum_{j\in T_{i,-}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2 \le \tau = O_*(\zeta^{5/6}/\lambda)$ with small enough hidden constant, where $T_{i,-}(\delta)$ defined in Lemma 28.

(iii) $\alpha, \boldsymbol{\beta}$ are well fitted: $|\alpha - \hat{\alpha}|^2 = O(\zeta)$, $\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 = O(\zeta)$ with small enough hidden constant.

Then, we can construct the following descent direction

$$(\alpha + \alpha_*)\nabla_\alpha L_\lambda + \langle \nabla_{\boldsymbol{\beta}} L_\lambda, \boldsymbol{\beta} + \boldsymbol{\beta}_* \rangle + \sum_{i\in[m_*]} \sum_{j\in\mathcal{T}_i} \langle \nabla_{\boldsymbol{w}_i} L_\lambda, \boldsymbol{w}_j - q_{ij}\boldsymbol{w}_i^* \rangle = \Omega(\zeta),$$

where q_{ij} satisfy the following conditions with $\delta_{close} < \delta_{\text{sign}}$ and $\delta_{close} = O_*(\zeta^{1/3})$: (1) $\sum_{j\in\mathcal{T}_i} a_j q_{ij} = a_i^*$; (2) $q_{ij} \ge 0$; (3) $q_{ij} = 0$ when $\operatorname{sign}(a_j) \ne \operatorname{sign}(a_i^*)$ or $\delta_j > \delta_{close}$. (4) $\sum_{i\in[m_*]} \sum_{j\in\mathcal{T}_i} q_{ij}^2 = O_*(1)$.

E.4 Proof of Lemma 6

Now we are ready to prove the gradient lower bound (Lemma 6) by combining all descent direction lemma in the previous section together.

Lemma 6 (Gradient lower bound)**.** *Suppose the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - \min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu)$. If $\Omega_*(\lambda^2) \le \zeta \le O_*(\lambda^{9/5})$ and $\lambda \le O_*(1)$, we have*

$$\|\nabla_{\boldsymbol{\theta}} L_\lambda\|_F^2 \ge \Omega_*(\zeta^4/\lambda^2).$$

Proof. We check the assumption of Lemma 29 one by one.

For assumption (i) (norm balance) in Lemma 29, whenever $\sum_{i\in[m_*]} \left| a_i^2 - \|\boldsymbol{w}_i\|_2^2 \right| = \Omega_*(\zeta^2/\lambda^2)$, by Lemma 27 we know

$$\sum_i \sum_{j\in T_i} |\langle \nabla_{a_j} L_\lambda, -a_j \rangle + \langle \nabla_{\boldsymbol{w}_j} L_\lambda, \boldsymbol{w}_j \rangle| \ge \Omega_*(\zeta^2/\lambda).$$

With Lemma 26, this implies

$$\sqrt{\|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2} \cdot O(\|\boldsymbol{a}_*\|_1) \ge \sqrt{\|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2}\sqrt{\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{W}\|_F^2} = \Omega_*(\zeta^2/\lambda),$$

which means

$$\|\nabla_{\boldsymbol{\theta}} L_\lambda\|_F^2 \ge \|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2 \ge \Omega_*(\zeta^4/\lambda^2)$$

For assumption (ii) (norm cancellation) in Lemma 29, whenever it does not hold, by Lemma 28 we know

$$\|\nabla_{\boldsymbol{\theta}} L_\lambda\|_F^2 \ge \|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2 \ge \lambda^2 \sum_{j\in T_{i,-}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2 \ge \Omega_*(\zeta^{5/6}\lambda).$$

For assumption (iii) $(\alpha, \boldsymbol{\beta})$ in Lemma 29, whenever it does not hold, by Lemma 8 we know

$$|\nabla_\alpha L_\lambda|^2 = (\alpha - \hat{\alpha})^2 = \Omega(\zeta), \quad \|\nabla_{\boldsymbol{\beta}} L_\lambda\|_2^2 = 4\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 = \Omega(\zeta),$$

which implies

$$\|\nabla_{\boldsymbol{\theta}} L_\lambda\|_F^2 \ge |\nabla_\alpha L_\lambda|^2 + \|\nabla_{\boldsymbol{\beta}} L_\lambda\|_2^2 = \Omega(\zeta).$$

Thus, the remaining case is the one that all assumption (i)-(iii) in Lemma 29 hold and also $\sum_{i\in[m_*]} \left| a_i^2 - \|\boldsymbol{w}_i\|_2^2 \right| = O_*(\zeta^2/\lambda^2)$, we choose

$$q_{ij} = \begin{cases} \frac{a_j a_i^*}{\sum_{j\in T_{i,+}(\delta_{close})} a_j^2} & \text{, if } j \in T_{i,+}(\delta_{close}) \\ 0 & \text{, otherwise} \end{cases}$$

so that condition (1)-(4) all hold: condition (1)-(3) are easy to check, Lemma 36 shows condition (4) holds. Now we know from Lemma 29 that

$$\alpha \nabla_\alpha L_\lambda + \langle \nabla_{\boldsymbol{\beta}} L_\lambda, \boldsymbol{\beta}\rangle + \sum_{i\in[m_*]} \sum_{j\in\mathcal{T}_i} \langle \nabla_{\boldsymbol{w}_i} L_\lambda, \boldsymbol{w}_j - q_{ij}\boldsymbol{w}_i^*\rangle = \Omega(\zeta).$$

Note that

$$\begin{aligned} &(\alpha+\alpha_*)\nabla_\alpha L_\lambda + \langle \nabla_{\boldsymbol{\beta}} L_\lambda, \boldsymbol{\beta}+\boldsymbol{\beta}_*\rangle + \sum_{i\in[m_*]} \sum_{j\in\mathcal{T}_i} \langle \nabla_{\boldsymbol{w}_i} L_\lambda, \boldsymbol{w}_j - q_{ij}\boldsymbol{w}_i^*\rangle \\ \le& \sqrt{|\nabla_\alpha L_\lambda|^2 + \|\nabla_{\boldsymbol{\beta}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2} \sqrt{(\alpha+\alpha_*)^2 + \|\boldsymbol{\beta}+\boldsymbol{\beta}_*\|_2^2 + \sum_{i\in[m_*]} \sum_{j\in\mathcal{T}_i} \|\boldsymbol{w}_j - q_{ij}\boldsymbol{w}_i^*\|_2^2} \end{aligned}$$

and

$$\begin{aligned} |\alpha+\alpha_*| &\le |\hat\alpha| + |\alpha_*| + O(\zeta^{1/2}) \overset{(a)}{\le} O_*(1) \\ \|\boldsymbol{\beta}+\boldsymbol{\beta}_*\|_2 &\le \left\|\hat{\boldsymbol{\beta}}\right\|_2 + \|\boldsymbol{\beta}_*\|_2 + O(\zeta^{1/2}) \overset{(b)}{\le} O_*(1) \\ \sum_{i\in[m_*]} \sum_{j\in\mathcal{T}_i} \|\boldsymbol{w}_j - q_{ij}\boldsymbol{w}_i^*\|_2^2 &\le 2 \sum_{i\in[m_*]} \sum_{j\in\mathcal{T}_i} \|\boldsymbol{w}_j\|_2^2 + q_{ij}^2 \|\boldsymbol{w}_i^*\|_2^2 \overset{(c)}{\le} O_*(1), \end{aligned}$$

where (a)(b) by Lemma 18; (c) we use Lemma 26 and condition (4) on q_{ij}.

Therefore, we get

$$\|\nabla_{\boldsymbol{\theta}} L_\lambda\|_F^2 = |\nabla_\alpha L_\lambda|^2 + \|\nabla_{\boldsymbol{\beta}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2 = \Omega_*(\zeta^2).$$

Combine all cases above, we know

$$\|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2 = \Omega_*(\min\{\zeta^4/\lambda^2, \zeta^{5/6}\lambda, \zeta, \zeta^2\}) = \Omega_*(\zeta^4/\lambda^2),$$

as long as $\zeta = O(\lambda^{9/5}/\operatorname{poly}(r, m_*, \Delta, \|\boldsymbol{a}_*\|_1, a_{\min}))$. □

F Non-degenerate dual certificate

In this section, we show that there indeed exists a non-degenerate dual certificate that satisfies Definition 1 and therefore proving Lemma 15.

Lemma 15. *There exists a non-degenerate dual certificate* $\eta = \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{x})\sigma_{\ge 2}(\boldsymbol{w}^\top \boldsymbol{x})]$ *with* $\rho_\eta = \Theta(1)$ *and* $\|p\|_2 \le \operatorname{poly}(m_*, \Delta)$

Recall that we want to use the dual certificate η to characterize the (approximate) solution for the following regression problem:

$$\min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})} L_\lambda(\mu) = \mathbb{E}_{\boldsymbol{x},\widetilde{y}}[(f_\mu(\boldsymbol{x}) - \widetilde{y})^2] + \lambda|\mu|_1 = \mathbb{E}_{\boldsymbol{x}}\left[\left(\int_{\boldsymbol{w}} \sigma_{\ge 2}(\boldsymbol{w}^\top \boldsymbol{x}) \mathrm{d}\,\mu - \mu_*\right)^2\right] + \lambda|\mu|_1,$$

where $\sigma_{\ge 2}$ is the ReLU activation after removing 0th and 1st order (corresponding to α and β terms in (1)) and $\mu_* = \sum_{i\in[m_*]} a_i^* \delta_{\boldsymbol{w}_i^*}$ is the ground-truth.

We need to first introduce few notations before proceeding to the proof. Denote the kernel $K_{\geq\ell}(\boldsymbol{w},\boldsymbol{u}) = \mathbb{E}_{\boldsymbol{x}\sim N(0,\boldsymbol{I})}[\overline{\sigma_{\geq\ell}}(\overline{\boldsymbol{w}}^\top\boldsymbol{x})\overline{\sigma_{\geq\ell}}(\overline{\boldsymbol{u}}^\top\boldsymbol{x})]$ as the kernel induced by activation $\sigma_{\geq\ell}(x)$, where $\overline{\sigma_{\geq\ell}}(x) = \sum_{k\geq\ell}\hat{\sigma}_k h_k(x)/Z_\sigma$, $Z_\sigma = \|\sigma_{\geq\ell}\|_2 = \sqrt{\sum_{k\geq\ell}\hat{\sigma}_k^2} = \Theta(\ell^{-3/4})$ is the normalizing factor, $h_k(x)$ is the normalized k-th (probabilistic) Hermite polynomial and $\hat{\sigma}_k$ is the corresponding Hermite coefficient. We will specify the value of ℓ later and use K instead of $K_{\geq\ell}$ for simplicity.

We will construct the dual certificate η following the proof strategy in Poon et al. [26] with the form below (the difference is that we now only keep high order terms that are at least ℓ):

$$\eta(\boldsymbol{w}) = \sum_{j\in[m_*]}\alpha_{1,j}K(\boldsymbol{w}_j^*,\boldsymbol{w}) + \sum_{j\in[m_*]}\boldsymbol{\alpha}_{2,j}^\top\nabla_1 K(\boldsymbol{w}_j^*,\boldsymbol{w})$$

such that it satisfies $\eta(\boldsymbol{w}_i^*) = \operatorname{sign}(a_i^*)$ and $\nabla\eta(\boldsymbol{w}_i^*) = 0$ for all $i\in[m_*]$. Here $\boldsymbol{\alpha}_1 = (\alpha_1,\ldots,\alpha_{m_*})^\top \in \mathbb{R}^{m_*}, \boldsymbol{\alpha}_2 = (\boldsymbol{\alpha}_{2,1}^\top,\ldots,\boldsymbol{\alpha}_{2,m_*}^\top)^\top \in \mathbb{R}^{m_* d}$ are the parameters that we are going to solve and ∇_i means the gradient w.r.t. i-th variable.

One can rewrite the above constraints into the matrix form:

$$\boldsymbol{\Upsilon}\begin{pmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{pmatrix} = \boldsymbol{b}, \tag{5}$$

where $\boldsymbol{b} = (\operatorname{sign}(a_1^*),\ldots,\operatorname{sign}(a_{m_*}^*),\boldsymbol{0}_{m^*d}^\top)^\top \in \mathbb{R}^{m_*(d+1)}$, $\boldsymbol{\Upsilon} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\gamma}(\boldsymbol{x})\boldsymbol{\gamma}(\boldsymbol{x})^\top] \in \mathbb{R}^{m_*(d+1)\times m_*(d+1)}$,
$\boldsymbol{\gamma}(\boldsymbol{x}) = (\overline{\sigma_{\geq\ell}}(\boldsymbol{w}_1^{*\top}\boldsymbol{x}),\ldots,\overline{\sigma_{\geq\ell}}(\boldsymbol{w}_{m_*}^{*\top}\boldsymbol{x}),\nabla_{\boldsymbol{w}}\overline{\sigma_{\geq\ell}}(\overline{\boldsymbol{w}}_1^{*\top}\boldsymbol{x})^\top,\ldots,\nabla_{\boldsymbol{w}}\overline{\sigma_{\geq\ell}}(\overline{\boldsymbol{w}}_{m_*}^{*\top}\boldsymbol{x})^\top)^\top \in \mathbb{R}^{m_*(d+1)}$,
$\nabla_{\boldsymbol{w}}\overline{\sigma_{\geq\ell}}(\overline{\boldsymbol{w}}_i^{*\top}\boldsymbol{x}) = \boldsymbol{P}_{\boldsymbol{w}_i^*}\overline{\sigma_{\geq\ell}}'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\boldsymbol{x} \in \mathbb{R}^d$.

Notions on the unit sphere As we could see, the kernel K is invariant under the change of norms, so it suffices to focus on the input on the unit sphere $\mathbb{S}^{d-1}$. On the unite sphere, we could compute the gradient and hessian of a function $f(\boldsymbol{w})$ on the sphere (e.g., Absil et al. [4])

$$\begin{aligned}&\operatorname{grad} f(\boldsymbol{w}) = \boldsymbol{P}_{\boldsymbol{w}}\nabla f(\boldsymbol{w}),\\ &\operatorname{H} f(\boldsymbol{w})[\boldsymbol{z}] = \boldsymbol{P}_{\boldsymbol{w}}(\nabla^2 f(\boldsymbol{w}) - \overline{\boldsymbol{w}}^\top\nabla f(\boldsymbol{w})\boldsymbol{I})\boldsymbol{z} \quad \text{for all tangent vector } \boldsymbol{z} \text{ that } \boldsymbol{z}^\top\boldsymbol{w} = 0,\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{w}} = \boldsymbol{I} - \boldsymbol{w}\boldsymbol{w}^\top$ is the projection matrix.

Then, we could define the derivative as in Poon et al. [26], Absil et al. [3]: for tangent vectors $\boldsymbol{z},\boldsymbol{z}'$

$$\begin{aligned}\operatorname{D}_0 f(\boldsymbol{w}) &:= f(\boldsymbol{w})\\ \operatorname{D}_1 f(\boldsymbol{w})[\boldsymbol{z}] &:= \langle\boldsymbol{z},\operatorname{grad} f(\boldsymbol{w})\rangle = \boldsymbol{z}^\top\boldsymbol{P}_{\boldsymbol{w}}\nabla f(\boldsymbol{w})\\ \operatorname{D}_2 f(\boldsymbol{w})[\boldsymbol{z},\boldsymbol{z}'] &:= \langle\operatorname{H} f(\boldsymbol{w})[\boldsymbol{z}],\boldsymbol{z}'\rangle = \boldsymbol{z}^\top\boldsymbol{P}_{\boldsymbol{w}}(\nabla^2 f(\boldsymbol{w}) - \overline{\boldsymbol{w}}^\top\nabla f(\boldsymbol{w})\boldsymbol{I})\boldsymbol{P}_{\boldsymbol{w}}\boldsymbol{z}',\end{aligned}$$

and their associated norms

$$\begin{aligned}\|\operatorname{D}_1 f(\boldsymbol{w})\|_{\boldsymbol{w}} &:= \sup_{\|\boldsymbol{z}\|_{\boldsymbol{w}}=1}\operatorname{D}_1 f(\boldsymbol{w})[\boldsymbol{z}] = \|\boldsymbol{P}_{\boldsymbol{w}}\nabla f(\boldsymbol{w})\|_2,\\ \|\operatorname{D}_2 f(\boldsymbol{w})\|_{\boldsymbol{w}} &:= \sup_{\|\boldsymbol{z}\|_{\boldsymbol{w}},\|\boldsymbol{z}'\|_{\boldsymbol{w}}=1}\operatorname{D}_2 f(\boldsymbol{w})[\boldsymbol{z},\boldsymbol{z}'] = \|\boldsymbol{P}_{\boldsymbol{w}}\operatorname{H} f(\boldsymbol{w})\boldsymbol{P}_{\boldsymbol{w}}\|_2,\end{aligned}$$

where $\|\boldsymbol{z}\|_{\boldsymbol{w}} = \|\boldsymbol{P}_{\boldsymbol{w}}\boldsymbol{z}\|_2$.

For simplicity, we will use $K^{(ij)}(\boldsymbol{w},\boldsymbol{u})$ to denote $\nabla_1^i\nabla_2^j K(\boldsymbol{w},\boldsymbol{u})$. One can check that this is in fact the same as the one defined Poon et al. [26] under our specific kernel K, $i+j\leq 3$ and $i,j\leq 2$. Let

$$\left\|K^{(ij)}(\boldsymbol{w},\boldsymbol{u})\right\|_{\boldsymbol{w},\boldsymbol{u}} := \sup_{\substack{\left\|\boldsymbol{z}_{\boldsymbol{w}}^{(p)}\right\|_{\boldsymbol{w}}=\left\|\boldsymbol{z}_{\boldsymbol{u}}^{(q)}\right\|_{\boldsymbol{u}}=1,\\ \boldsymbol{w}^\top\boldsymbol{z}_{\boldsymbol{w}}^{(p)}=\boldsymbol{u}^\top\boldsymbol{z}_{\boldsymbol{u}}^{(q)}=0\ \forall p\in[i],q\in[j]}} K^{(ij)}(\boldsymbol{w},\boldsymbol{u})[\boldsymbol{z}_{\boldsymbol{w}}^{(1)},\ldots,\boldsymbol{z}_{\boldsymbol{u}}^{(j)}],$$

where $\boldsymbol{z}_{\boldsymbol{w}}^{(p)}$ applies to the dimension corresponding to $\boldsymbol{w}$ and similarly $\boldsymbol{z}_{\boldsymbol{u}}^{(q)}$ for $\boldsymbol{u}$.

Before solving (5), we first present some useful proprieties of kernel K that will be used later (see Section H for the proofs). The lemma below shows that kernel $K(\boldsymbol{w},\boldsymbol{u})$ is non-degenerate in the sense that it decays at least quadratic at each ground-truth direction ($\boldsymbol{w}\approx\boldsymbol{u}\approx\boldsymbol{w}_i^*$) and contributes almost nothing when $\boldsymbol{w},\boldsymbol{u}$ are away.

Lemma 30 (Non-degeneracy of kernel K)**.** *For $\ell \geq \Theta(\Delta^{-2}\log(m_*\ell/h\Delta))$, kernel $K_{\geq\ell}$ is non-degenerate in the sense that there exists $r = \Theta(\ell^{-1/2}), \rho_1 = \Theta(1), \rho_2 = \Theta(\ell)$ such that following hold:*

(i) $K(\boldsymbol{w},\boldsymbol{u}) \leq 1 - \rho_1$ *for all* $\delta(\boldsymbol{w},\boldsymbol{u}) := \angle(\boldsymbol{w},\boldsymbol{u}) \geq r$.

(ii) $K^{(20)}(\boldsymbol{w},\boldsymbol{u})[\boldsymbol{z},\boldsymbol{z}] \leq -\rho_2 \left\|\boldsymbol{z}\right\|^2$ *for tangent vector* $\boldsymbol{z}$ *that* $\boldsymbol{z}^\top\boldsymbol{w} = 0$ *and* $\delta(\boldsymbol{w},\boldsymbol{u}) \leq r$.

(iii) $\left\|K^{(ij)}(\boldsymbol{w}_1^*,\boldsymbol{w}_k^*)\right\|_{\boldsymbol{w}_i^*,\boldsymbol{w}_k^*} \leq h/m_*^2$ *for* $(i,j) \in \{0,1\}\times\{0,1,2\}$

The following lemma shows that K and its derivatives are bounded.

Lemma 31 (Regularity conditions on kernel K)**.** *Let* $B_{ij} := \sup_{\boldsymbol{w},\boldsymbol{u}} \left\|K^{(ij)}(\boldsymbol{w},\boldsymbol{u})\right\|_{\boldsymbol{w},\boldsymbol{u}}$ *and* $B_0 = B_{00}+B_{10}+1$, $B_2 = B_{20}+B_{21}+1$. *We have* $B_{00} = O(1)$, $B_{10} = O(1)$, $B_{11} = O(\ell)$, $B_{20} = O(\ell)$, $B_{21} = O(\ell^{3/2})$, *and therefore* $B_0 = O(1)$, $B_2 = O(\ell^{3/2})$.

The following lemma from Poon et al. [26] connects the non-degeneracy of kernel K to the dual certificate η that we are interested in.

Lemma 32 (Lemma 2, Poon et al. [26], adapted in our setting)**.** *Let $a \in \{\pm 1\}$. Suppose that for some $\rho > 0$, $B > 0$ and $0 < r \leq B^{-1/2}$ we have: for all $\delta(\boldsymbol{w},\boldsymbol{w}_0)$ and $\boldsymbol{z} \in \mathbb{R}^d$ with $\boldsymbol{z}^\top\boldsymbol{w} = 0$, it holds that $-K^{(02)}(\boldsymbol{w}_0,\boldsymbol{w})[\boldsymbol{z},\boldsymbol{z}] > \rho\left\|\boldsymbol{z}\right\|_2^2$ and $\left\|K^{(02)}(\boldsymbol{w}_0,\boldsymbol{w})\right\|_{\boldsymbol{w}} \leq B$. Let η be a smooth function. If $\eta(\boldsymbol{w}_0) = a$, $\nabla\eta(\boldsymbol{w}_0) = 0$ and $\left\|a\,\mathrm{D}_2\,\eta(\boldsymbol{w}) - K^{(02)}(\boldsymbol{w}_0,\boldsymbol{w})\right\|_{\boldsymbol{w}} \leq \tau$ for all $\delta(\boldsymbol{w},\boldsymbol{w}_0) \leq r$ with $\tau < \rho/2$, then we have $|\eta(\boldsymbol{w})| \leq 1 - ((\rho-2\tau)/2)\delta(\boldsymbol{w},\boldsymbol{w}_0)^2$ for all $\delta(\boldsymbol{w},\boldsymbol{w}_0) \leq r$.*

We now are ready to proof the main result in this section Lemma 15 that shows the non-degenerate dual certificate exists. Roughly speaking, following the same proof as in Poon et al. [26], we can show that $\boldsymbol{\alpha} \approx \mathrm{sign}(\boldsymbol{a}_*)$ and $\boldsymbol{\alpha}_2 \approx \boldsymbol{0}$ and therefore we can transfer the non-degeneracy of kernel K to the dual certificate η with Lemma 32.

Lemma 15. *There exists a non-degenerate dual certificate $\eta = \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{x})\sigma_{\geq 2}(\boldsymbol{w}^\top\boldsymbol{x})]$ with $\rho_\eta = \Theta(1)$ and $\left\|p\right\|_2 \leq \mathrm{poly}(m_*,\Delta)$*

Proof. Note that $\boldsymbol{\Upsilon} = \boldsymbol{S}\boldsymbol{D}\widetilde{\boldsymbol{\Upsilon}}\boldsymbol{D}\boldsymbol{S}$, where

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{I}_{m_*} & & & \\ & \boldsymbol{P}_{\boldsymbol{w}_1^*} & & \\ & & \ddots & \\ & & & \boldsymbol{P}_{\boldsymbol{w}_{m_*}^*} \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{I}_{m_*} & & & \\ & (Z_{\sigma'}/Z_\sigma)\boldsymbol{I}_{m_*} & & \\ & & \ddots & \\ & & & (Z_{\sigma'}/Z_\sigma)\boldsymbol{I}_{m_*} \end{pmatrix}$$

are block diagonal matrices, $\widetilde{\boldsymbol{\Upsilon}} = \mathbb{E}_{\boldsymbol{x}}[\widetilde{\gamma}(\boldsymbol{x})\widetilde{\gamma}(\boldsymbol{x})^\top] \in \mathbb{R}^{m_*(d+1)\times m_*(d+1)}$,

$$\widetilde{\gamma}(\boldsymbol{x}) = (\overline{\sigma_{\geq\ell}}(\boldsymbol{w}_1^{*\top}\boldsymbol{x}),\ldots,\overline{\sigma_{\geq\ell}}(\boldsymbol{w}_{m_*}^{*\top}\boldsymbol{x}),(Z_\sigma/Z_{\sigma'})\nabla_{\boldsymbol{w}}\overline{\sigma_{\geq\ell}}(\boldsymbol{w}_1^{*\top}\boldsymbol{x})^\top,\ldots,(Z_\sigma/Z_{\sigma'})\nabla_{\boldsymbol{w}}\overline{\sigma_{\geq\ell}}(\boldsymbol{w}_{m_*}^{*\top}\boldsymbol{x})^\top)^\top \in \mathbb{R}^{m_*(d+1)},$$

$\nabla_{\boldsymbol{w}}\overline{\sigma_{\geq\ell}}(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) = \overline{\sigma_{\geq\ell}}'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\boldsymbol{x}$ and $Z_{\sigma'} = \sqrt{\sum_{k\geq\ell}\hat{\sigma}_k^2 k} = \Theta(\ell^{-1/4})$ is the normalizing factor so that the diagonal of $\widetilde{\boldsymbol{\Upsilon}}$ are all 1.

Thus, to solve (5), it is sufficient to solve the following: denote $\widetilde{\boldsymbol{K}} = \boldsymbol{D}\widetilde{\boldsymbol{\Upsilon}}\boldsymbol{D}$

$$\widetilde{\boldsymbol{K}}\begin{pmatrix}\widetilde{\boldsymbol{\alpha}}_1\\ \widetilde{\boldsymbol{\alpha}}_2\end{pmatrix} = \boldsymbol{b}, \tag{6}$$

and let $\boldsymbol{\alpha}_1 = \widetilde{\boldsymbol{\alpha}}_1$, $\boldsymbol{\alpha}_{2,i} = (Z_\sigma/Z_{\sigma'})\widetilde{\boldsymbol{\alpha}}_{2,i}$ to get the solution of (5).

In the following, we are going to first show that $\widetilde{\boldsymbol{K}} \approx \boldsymbol{DD}$ because all the off-diagonal terms of $\widetilde{\boldsymbol{\Upsilon}}$ are small due to Lemma 30 (iii). Specifically, we have

$$\left\|\widetilde{\boldsymbol{K}} - \boldsymbol{DD}\right\|_2 = \sup_{\|\boldsymbol{z}\|_2=1} |\boldsymbol{z}^\top(\widetilde{\boldsymbol{K}} - \boldsymbol{DD})\boldsymbol{z}|$$

$$= \sup_{\|\boldsymbol{z}\|_2=1} \left|\sum_{i,j} z_{1,i} K(\boldsymbol{w}_i^*, \boldsymbol{w}_j^*) z_{1,j} + 2(Z_\sigma/Z_{\sigma'}) \sum_{i,j} z_{1,i} \nabla_1 K(\boldsymbol{w}_i^*, \boldsymbol{w}_j^*)^\top \boldsymbol{z}_{2,j} + (Z_\sigma/Z_{\sigma'})^2 \sum_{i,j} \boldsymbol{z}_{1,i}^\top \nabla_1 \nabla_2 K(\boldsymbol{w}_i^*, \boldsymbol{w}_j^*)^\top \boldsymbol{z}_{2,j}\right|$$

$$\leq \sqrt{\sum_{i,j} K(\boldsymbol{w}_i^*, \boldsymbol{w}_j^*)^2 + \Theta(\ell^{-1}) \left\|K^{(10)}(\boldsymbol{w}_i^*, \boldsymbol{w}_j^*)\right\|_{\boldsymbol{w}_i^*}^2 + \Theta(\ell^{-2}) \left\|K^{(11)}(\boldsymbol{w}_i^*, \boldsymbol{w}_j^*)\right\|_{\boldsymbol{w}_i^*, \boldsymbol{w}_j^*}^2} \leq 2h,$$

where $\boldsymbol{z} = (\boldsymbol{z}_1^\top, \boldsymbol{z}_2^\top)^\top$, $\boldsymbol{z}_1 = (\boldsymbol{z}_{1,1}, \ldots, \boldsymbol{z}_{1,m_*})^\top$ and $\boldsymbol{z}_2 = (\boldsymbol{z}_{2,1}^\top, \ldots, \boldsymbol{z}_{2,m_*}^\top)^\top$ has the same block structure as $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ and we use Lemma 30 in the last line.

Note that $\boldsymbol{DD}$ has exactly $m_* d$ eigenvalues of 1 and m_* eigenvalues of 0, and $\widetilde{\boldsymbol{K}}$ also has m_* eigenvalues of 0. By Weyl's inequality, we know $|\gamma_i - 1| \leq 2h$ where $\widetilde{\boldsymbol{K}} = \sum_{i \in [m_* d]} \gamma_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ is its eigendecomposition. Here $\boldsymbol{v}_i^\top \boldsymbol{v}_\perp = 0$ for all $\boldsymbol{v}_\perp \in V_\perp = \operatorname{span}\{(\boldsymbol{0}, \boldsymbol{w}_1^*, \boldsymbol{0}, \ldots, \boldsymbol{0})^\top, \ldots (\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{w}_{m_*}^*)^\top\}$ in the null space of $\boldsymbol{D}$. Since $\boldsymbol{b}^\top \boldsymbol{v}_\perp = 0$ for all $\boldsymbol{v}_\perp \in V_\perp$, we have

$$\begin{pmatrix} \widetilde{\boldsymbol{\alpha}}_1 \\ \widetilde{\boldsymbol{\alpha}}_2 \end{pmatrix} = \widetilde{\boldsymbol{K}}^\dagger \boldsymbol{b} = \sum_{i \in [m_* d]} \gamma_i^{-1} \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{b} = \sum_{i \in [m_* d]} (\gamma_i^{-1} - 1) \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{b} + \sum_{i \in [m_* d]} \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{b} = \sum_{i \in [m_* d]} (\gamma_i^{-1} - 1) \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{b} + \boldsymbol{b}.$$

Therefore,

$$\left\|\begin{pmatrix} \widetilde{\boldsymbol{\alpha}}_1 \\ \widetilde{\boldsymbol{\alpha}}_2 \end{pmatrix} - \boldsymbol{b}\right\|_2 \leq \left\|\sum_{i \in [m_* d]} (\gamma_i^{-1} - 1) \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{b}\right\|_2 \leq \max_i |\gamma_i^{-1} - 1| \sqrt{m_*} = O(h\sqrt{m_*}) =: h'.$$

This implies $\|\boldsymbol{\alpha}_1 - \operatorname{sign}(\boldsymbol{a}_*)\|_\infty = \|\widetilde{\boldsymbol{\alpha}}_1 - \operatorname{sign}(\boldsymbol{a}_*)\|_\infty \leq h'$, $\|\boldsymbol{\alpha}_1\|_\infty = \|\widetilde{\boldsymbol{\alpha}}_1\|_\infty \leq 1 + h'$ and $\|\boldsymbol{\alpha}_2\|_2 = (Z_\sigma/Z_{\sigma'}) \|\widetilde{\boldsymbol{\alpha}}_{2,i}\|_2 \leq \Theta(h'\ell^{-1/2})$.

Now, given the $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, we can show the corresponding η is non-degenerate. Choosing $h = O(m_*^{-1/2})$ and $\ell = \Theta(\Delta^{-2}\log(m_*/\Delta))$ so that the condition in Lemma 30 holds.

Consider $\boldsymbol{w} \in \mathcal{T}_i$, when $\delta(\boldsymbol{w}, \boldsymbol{w}_i^*) \geq r = \Theta(\ell^{-1/2})$, using Lemma 30 and Lemma 31 we have

$$\begin{aligned} |\eta(\boldsymbol{w})| =& \left|\sum_{j \in [m_*]} \alpha_{1,j} K(\boldsymbol{w}_j^*, \boldsymbol{w}) + \sum_{j \in [m_*]} \boldsymbol{\alpha}_{2,j}^\top \nabla_1 K(\boldsymbol{w}_j^*, \boldsymbol{w})\right| \\ \leq& \sum_{j \in [m_*]} |\alpha_{1,j}| |K(\boldsymbol{w}_j^*, \boldsymbol{w})| + \sum_{j \in [m_*]} \|\boldsymbol{\alpha}_{2,j}\|_{\boldsymbol{w}_j^*} \left\|\nabla_1 K(\boldsymbol{w}_j^*, \boldsymbol{w})\right\|_{\boldsymbol{w}_j^*} \\ \leq& (1 + h')(1 - \rho_1 + h) + \Theta(h'\ell^{-1/2})(B_{10} + h) \leq 1 - \rho_1/2 \leq 1 - \Theta(\rho_1)\delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2, \end{aligned}$$

where we choose $h = O(m_*^{-1/2})$ to be small enough.

When $\delta(\boldsymbol{w}, \boldsymbol{w}_i^*) \leq r = \Theta(\ell^{-1/2})$, again using Lemma 30 and Lemma 31 we have

$$\begin{aligned} \left\|a_i^* \operatorname{D}_2 \eta(\boldsymbol{w}) - K^{(02)}(\boldsymbol{w}_i^*, \boldsymbol{w})\right\|_{\boldsymbol{w}} \leq& \left\|\alpha_{1,i} K^{(02)}(\boldsymbol{w}_i^*, \boldsymbol{w}) - K^{(02)}(\boldsymbol{w}_i^*, \boldsymbol{w})\right\|_{\boldsymbol{w}} + \sum_{j \neq i} \left\|\alpha_{1,j} K^{(02)}(\boldsymbol{w}_j^*, \boldsymbol{w})\right\|_{\boldsymbol{w}} \\ &+ \sum_{j \in [m_*]} \|\boldsymbol{\alpha}_{2,j}\|_{\boldsymbol{w}_j^*} \left\|K^{(12)}(\boldsymbol{w}_j^*, \boldsymbol{w})\right\|_{\boldsymbol{w}_j^*, \boldsymbol{w}} \\ \leq& h' B_{02} + (1 + h')h + \Theta(h'\ell^{-1/2})(B_{21} + h) \leq \rho_2/16, \end{aligned}$$

where again due to our choice of small h. Using Lemma 32 we know that $|\eta(\boldsymbol{w})| \leq 1 - (\rho_2/4)\delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2$.

Combine the above two cases, we have $|\eta(\boldsymbol{w})| \leq 1 - \Theta(1)\delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2$ and $\eta(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{x})\sigma(\boldsymbol{w}^\top \boldsymbol{x})]$ with

$$p(\boldsymbol{x}) = \frac{1}{Z_\sigma^2}\left(\sum_{j\in[m_*]} \alpha_{1,j}\sigma_{\geq\ell}(\boldsymbol{w}_j^{*\top}\boldsymbol{x}) + \sum_{j\in[m_*]} \boldsymbol{\alpha}_{2,j}^\top(\boldsymbol{I} - \boldsymbol{w}_i^*\boldsymbol{w}_i^{*\top})\boldsymbol{x}\sigma'_{\geq\ell}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\right).$$

We have $\|p\| = O(\ell^{3/4}m_* + m_* h' \ell^{-1/2}\ell^{5/4}) = \widetilde{O}(\Delta^{-3/2}m_*)$. □

G Proofs in Section E

In this section, we give the omitted proofs in Section E.

G.1 Omitted proofs in Section E.1

We give the proofs for these results that characterize the structure of ideal loss solution.

The following proof follows from the definition of non-degenerate dual certificate η.

Lemma 16. *Given a non-degenerate dual certificate η, then*

(i) For any measure $\mu \in \mathcal{M}(\mathbb{S}^{d-1})$, $|\langle\eta, \mu\rangle| \leq |\mu|_1 - \rho_\eta \sum_{i\in[m_*]}\int_{\mathcal{T}_i}\delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \,\mathrm{d}|\mu|(\boldsymbol{w})$.

(ii) $\langle\eta, \mu^*\rangle = |\mu^*|_1$

(iii) $\langle\eta, \mu - \mu^*\rangle = \langle p, f_\mu - f_{\mu^*}\rangle$, where $f_\mu(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{w}\sim\mu}[\sigma_{\geq 2}(\boldsymbol{w}^\top\boldsymbol{x})]$. Thus, $|\langle\eta, \mu - \mu^*\rangle| \leq \|p\|_2\sqrt{L(\mu)}$.

Proof. We show the results one by one.

Part (i)(ii) We have

$$|\langle\eta, \mu\rangle| \leq \int_{\mathbb{S}^{d-1}} |\eta(\boldsymbol{w})| \,\mathrm{d}|\mu|(\boldsymbol{w}) = \sum_{i\in[m_*]}\int_{\mathcal{T}_i} |\eta(\boldsymbol{w})| \,\mathrm{d}|\mu|(\boldsymbol{w}) \leq |\mu|_1 - \rho_\eta \sum_{i\in[m_*]}\int_{\mathcal{T}_i}\delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \,\mathrm{d}|\mu|(\boldsymbol{w}).$$

where the last inequality follows the property of non-degenerate dual certificate (Definition 1). The second part then follows directly by the definition of μ^*.

Part (iii) We have

$$\begin{aligned}\langle\eta, \mu - \mu^*\rangle = &\int_{\mathbb{S}^{d-1}} \eta(\boldsymbol{w}) \,\mathrm{d}(\mu - \mu^*)(\boldsymbol{w}) = \int_{\mathbb{S}^{d-1}} \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{x})\sigma_{\geq 2}(\boldsymbol{w}^\top\boldsymbol{x})] \,\mathrm{d}(\mu - \mu^*)(\boldsymbol{w})\\ =&\mathbb{E}_{\boldsymbol{x}}\left[p(\boldsymbol{x})\int_{\mathbb{S}^{d-1}}\sigma_{\geq 2}(\boldsymbol{w}^\top\boldsymbol{x}) \,\mathrm{d}(\mu - \mu^*)(\boldsymbol{w})\right]\\ =&\mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{x})(f_\mu(\boldsymbol{x}) - f_{\mu^*}(\boldsymbol{x}))].\end{aligned}$$

Note that $L(\mu) = \|f_\mu - f_{\mu^*}\|_2^2$, this leads to $|\langle\eta, \mu - \mu^*\rangle| \leq \|p\|_2\sqrt{L(\mu)}$. □

Given the above lemma and the optimality of μ_λ^*, we are able to characterize the structure of μ_λ^* as below: norm is bounded, square loss is small and far-away neurons are small.

Lemma 17. *We have the following hold*

(i) $|\mu_*|_1 - \lambda\|p\|_2^2 \leq |\mu_\lambda^*|_1 \leq |\mu^*|_1 = \|\boldsymbol{a}^*\|_1$

(ii) $L(\mu_\lambda^*) \leq \lambda^2\|p\|_2^2 = O_*(\lambda^2)$

(iii) $\sum_{i\in[m_*]}\int_{\mathcal{T}_i}\delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \,\mathrm{d}|\mu_\lambda^*|(\boldsymbol{w}) \leq \lambda\|p\|_2^2/\rho_\eta = O_*(\lambda)$

Proof. We show the results one by one.

Part (i) Due to the optimality of μ_λ^*, we have

$$L(\mu_\lambda^*) + \lambda|\mu_\lambda^*|_1 = L_\lambda(\mu_\lambda^*) \le L_\lambda(\mu^*) = L(\mu^*) + \lambda|\mu^*|_1.$$

Rearranging the terms, we have

$$\lambda|\mu_\lambda^*|_1 - \lambda|\mu^*|_1 \le L(\mu^*) - L(\mu_\lambda^*) = -L(\mu_\lambda^*) \le 0.$$

For the lower bound, with Lemma 16 we have

$$0 \le |\mu_\lambda^*|_1 - |\mu^*|_1 - \langle \eta, \mu_\lambda^* - \mu^* \rangle \le |\mu_\lambda^*|_1 - |\mu^*|_1 + \|p\|_2 \sqrt{L(\mu_\lambda^*)}.$$

Using part (ii) we get the desired lower bound.

Part (ii) We first have the following inequality due to the optimality of μ_λ^* and adding $\lambda\langle \eta, \mu_\lambda^* - \mu^* \rangle$ on both side:

$$L(\mu_\lambda^*) + \underbrace{\lambda(|\mu_\lambda^*|_1 - |\mu^*|_1) - \lambda\langle \eta, \mu_\lambda^* - \mu^* \rangle}_{(I)} \le L(\mu^*) - \lambda\langle \eta, \mu_\lambda^* - \mu^* \rangle.$$

For (I), we have

$$(I) = \lambda(|\mu_\lambda^*|_1 - \langle \eta, \mu_\lambda^* \rangle) + \lambda(\langle \eta, \mu^* \rangle - |\mu^*|_1) \ge 0,$$

where we use Lemma 16 in the last inequality.

Therefore, the above inequality leads to

$$L(\mu_\lambda^*) \le L(\mu^*) - \lambda\langle \eta, \mu_\lambda^* - \mu^* \rangle \le \lambda \|p\|_2 \sqrt{L(\mu_\lambda^*)},$$

where we again use Lemma 16. This further leads to $L(\mu_\lambda^*) \le \lambda^2 \|p\|_2^2$.

Part (iii) Using part (i) we have

$$|\mu_\lambda^*|_1 - |\mu^*|_1 - \langle \eta, \mu_\lambda^* - \mu^* \rangle \le -\langle \eta, \mu_\lambda^* - \mu^* \rangle.$$

With Lemma 16, LHS and RHS become

$$\text{LHS} = |\mu_\lambda^*|_1 - \langle \eta, \mu_\lambda^* \rangle \ge \rho_\eta \sum_{i \in [m_*]} \int_{\mathcal{T}_i} \delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \, \mathrm{d}|\mu_\lambda^*|(\boldsymbol{w})$$

$$\text{RHS} \le \|p\|_2 \sqrt{L(\mu_\lambda^*)}.$$

Then using part (ii) we have the desired result. □

We are now ready to characterize the approximated solution by comparing μ and μ_λ^*.

Lemma 18. *Recall the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$. Then, the following holds:*

(i) $L(\mu) \le 5\lambda^2 \|p\|^2 + 4\zeta = O_*(\lambda^2 + \zeta)$.

(ii) if $\zeta \le \lambda|\mu^*|_1$ and $\lambda \le |\mu^*|_1 / \|p\|_2^2$, then $|\mu|_1 \le 3|\mu^*|_1 = 3\|\boldsymbol{a}^*\|_1$.

Proof. We show the results one by one.

Part (i) By the definition of the optimality gap ζ and adding $-\lambda\langle \eta, \mu - \mu^* \rangle$ on both side, we have

$$L(\mu) + \lambda(|\mu|_1 - |\mu_\lambda^*|_1) - \lambda\langle \eta, \mu - \mu^* \rangle \le L(\mu_\lambda^*) + \zeta - \lambda\langle \eta, \mu - \mu^* \rangle.$$

Note that on LHS,

$$\lambda(|\mu|_1 - |\mu_\lambda^*|_1) - \lambda\langle \eta, \mu - \mu^* \rangle = \lambda(|\mu|_1 - \langle \eta, \mu \rangle) + \lambda(|\mu^*|_1 - |\mu_\lambda^*|_1) \ge 0,$$

where we use Lemma 16 and Lemma 17.

Therefore, with Lemma 16 and Lemma 17 we get

$$L(\mu) \le L(\mu_\lambda^*) + \zeta - \lambda\langle \eta, \mu - \mu^* \rangle \le \lambda^2 \|p\|_2^2 + \zeta + \lambda \|p\|_2 \sqrt{L(\mu)}.$$

Solving the above inequality on $L(\mu)$ gives $L(\mu) \le 5\lambda^2 \|p\|_2^2 + 4\zeta$.

Part (ii) Again from the definition of the optimality gap ζ, we have

$$\lambda|\mu|_1 \le L(\mu^*_\lambda) + \lambda|\mu^*_\lambda|_1 + \zeta - L(\mu) \le \lambda^2 \left\|p\right\|_2^2 + \lambda|\mu^*|_1 + \zeta,$$

where we use Lemma 17. Thus, $|\mu|_1 \le \lambda \left\|p\right\|_2^2 + |\mu^*|_1 + \zeta/\lambda \le 3|\mu^*|_1$. □

The lemma below shows that far-away neurons are still small even for the approximated solution. Intutively, we use the non-degenerate dual certificate to certify the gap between μ and μ^*_λ and give a bound for it.

Lemma 19. *Recall the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu^*_\lambda)$. Then, we have*

$$\sum_{i\in[m_*]} \int_{\mathcal{T}_i} \delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \,\mathrm{d}|\mu|(\boldsymbol{w}) \le (\zeta/\lambda + 2\lambda \left\|p\right\|_2^2)/\rho_\eta = O_*(\zeta/\lambda + \lambda).$$

In particular, when $\mu = \sum_{i\in[m]} a_i \left\|\boldsymbol{w}_i\right\|_2 \delta_{\overline{\boldsymbol{w}}_i}$ represents finite number of neurons, we have

$$\sum_{i\in[m_*]} \sum_{j\in\mathcal{T}_i} |a_j| \left\|\boldsymbol{w}_j\right\|_2 \delta_j^2 \le (\zeta/\lambda + 2\lambda \left\|p\right\|_2^2)/\rho_\eta = O_*(\zeta/\lambda + \lambda),$$

where $\delta_j = \angle(\boldsymbol{w}_j, \boldsymbol{w}_i^*)$ for $j \in \mathcal{T}_i$.

Proof. By the definition of the optimality gap ζ, we have

$$L(\mu) + \lambda|\mu|_1 = L(\mu^*_\lambda) + \lambda|\mu^*_\lambda|_1 + \zeta.$$

Rearranging the terms and adding $-\langle \eta, \mu - \mu^* \rangle$ on both side, we get

$$|\mu|_1 - |\mu^*_\lambda|_1 - \langle \eta, \mu - \mu^* \rangle = \frac{1}{\lambda}(L(\mu^*_\lambda) - L(\mu) + \zeta) - \langle \eta, \mu - \mu^* \rangle.$$

For LHS, with Lemma 16 and Lemma 17 we have

$$\text{LHS} = |\mu|_1 - \langle \eta, \mu \rangle - |\mu^*_\lambda|_1 + |\mu^*|_1 \ge \rho_\eta \sum_{i\in[m_*]} \int_{\mathcal{T}_i} \delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \,\mathrm{d}|\mu|(\boldsymbol{w}).$$

For RHS, with Lemma 16 and Lemma 17 we have

$$\text{RHS} \le \frac{1}{\lambda}(\lambda^2 \left\|p\right\|_2^2 - L(\mu) + \zeta) + \left\|p\right\|_2 \sqrt{L(\mu)} = \frac{\zeta}{\lambda} + \lambda \left\|p\right\|_2^2 - \frac{L(\mu)}{\lambda} + \left\|p\right\|_2 \sqrt{L(\mu)}.$$

When $L(\mu) \ge \lambda^2 \left\|p\right\|_2^2$, we have $\text{RHS} \le \zeta/\lambda + \lambda \left\|p\right\|_2^2$. When $L(\mu) \le \lambda^2 \left\|p\right\|_2^2$, we have $\text{RHS} \le \zeta/\lambda + 2\lambda \left\|p\right\|_2^2$. Thus, in summary $\text{RHS} \le \zeta/\lambda + 2\lambda \left\|p\right\|_2^2$.

Combine the bounds on LHS and RHS we have

$$\rho_\eta \sum_{i\in[m_*]} \int_{\mathcal{T}_i} \delta(\boldsymbol{w}, \boldsymbol{w}_i^*)^2 \,\mathrm{d}|\mu|(\boldsymbol{w}) \le \zeta/\lambda + 2\lambda \left\|p\right\|_2^2.$$

□

The following lemma shows that every teacher neuron must have at least one close-by student neuron within angle $O_*(\zeta^{1/3})$. This generalize and greatly simplify the previous results Lemma 9 in [28]. In particular, we design a new test function using the Hermite expansion to achieve this.

Lemma 20. *Under Lemma 6, if the Hermite coefficient of σ decays as $|\hat{\sigma}_k| = \Theta(k^{-c_\sigma})$ with some constant $c_\sigma > 0$, then the total mass near each target direction is large, i.e., $\mu(\mathcal{T}_i(\delta)) \operatorname{sign}(a_i^*) \ge |a_i^*|/2$ for all $i \in [m_*]$ and any $\delta_{close} \ge \widetilde{\Omega}\left((\frac{L(\mu)}{a_{\min}^2})^{1/(4c_\sigma - 2)}\right)$ with large enough hidden constant. In particular, for σ is ReLU or absolute function, $\delta_{close} \ge \widetilde{\Omega}\left((\frac{L(\mu)}{a_{\min}^2})^{1/3}\right)$.*

As a corollary, if the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu^*_\lambda)$, then $\delta_{close} \ge \widetilde{\Omega}_*\left((\zeta + \lambda^2)^{1/(4c_\sigma - 2)}\right)$ and for ReLU or absolute $\delta_{close} \ge \widetilde{\Omega}_*\left((\zeta + \lambda^2)^{1/3}\right)$.

Proof. Assume towards contradiction that there exists some $i \in [m_*]$ with some $\delta_{close} \geq \widetilde{\Omega}\left((\frac{L(\mu)}{a_{\min}^2})^{1/(4c_\sigma-2)}\right)$ with large enough hidden constant such that $\mu(\mathcal{T}_i(\delta))\operatorname{sign}(a_i^*) \leq |a_i^*|/2$. For simplicity, we will use δ for δ_{close} in the following.

Let $g(x) = \sum_{\ell\leq k<2\ell} \operatorname{sign}(a_i^*)\operatorname{sign}(\hat{\sigma}_k)h_k(\boldsymbol{w}_i^{*\top}\boldsymbol{x})$ be a test function, where $h_k(x)$ is the k-th normalized probabilistic Hermite polynomial and ℓ will be chosen later.

Denote $R(\boldsymbol{x}) = f_\mu(\boldsymbol{x}) - f_{\mu^*}(\boldsymbol{x})$ so that $\|R\|_2^2 = L(\mu)$. We have

$$\begin{aligned}\sqrt{L(\mu)}\,\|g\|_2 \geq& \langle -R, g\rangle\\ =&\mathbb{E}_{\boldsymbol{x}}\left[\left(a_i^*\sigma(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \int_{\mathcal{T}_i(\delta)}\sigma(\boldsymbol{w}^\top\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{w})\right)g(\boldsymbol{x})\right]\\ &+\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{j\neq i}a_j^*\sigma(\boldsymbol{w}_j^{*\top}\boldsymbol{x}) - \int_{\mathbb{S}^{d-1}\setminus\mathcal{T}_i(\delta)}\sigma(\boldsymbol{w}^\top\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{w})\right)g(\boldsymbol{x})\right].\end{aligned}$$

Recall the Hermite expansion of $\sigma(x) = \sum_{k\geq 0}\hat{\sigma}_k h_k(x)$ and its property in Claim 2. For the first term, it becomes

$$\sum_{\ell\leq k<2\ell}\left(|a_i^*||\hat{\sigma}_k| - \int_{\mathcal{T}_i(\delta)}|\hat{\sigma}_k|\operatorname{sign}(a_i^*)(\boldsymbol{w}^\top\boldsymbol{w}_i^*)^k\,\mathrm{d}\mu(\boldsymbol{w})\right) \geq \frac{1}{2}|a_i^*|\sum_{\ell\leq k<2\ell}|\hat{\sigma}_k|.$$

For the second term, it becomes

$$\begin{aligned}&\sum_{\ell\leq k<2\ell}\left(\sum_{j\neq i}a_j^*|\hat{\sigma}_k|\operatorname{sign}(a_i^*)(\boldsymbol{w}_j^{*\top}\boldsymbol{w}_i^*)^k - \int_{\mathbb{S}^{d-1}\setminus\mathcal{T}_i(\delta)}|\hat{\sigma}_k|\operatorname{sign}(a_i^*)(\boldsymbol{w}^\top\boldsymbol{w}_i^*)^k\,\mathrm{d}\mu(\boldsymbol{w})\right)\\ \leq&(\|\boldsymbol{a}^*\|_1 + |\mu|_1)\sum_{\ell\leq k\leq 2\ell}|\hat{\sigma}_k|\max_{\angle(\boldsymbol{w},\boldsymbol{w}_i^*)\geq\delta}(\boldsymbol{w}^\top\boldsymbol{w}_i^*)^k\\ \leq&(\|\boldsymbol{a}^*\|_1 + |\mu|_1)\sum_{\ell\leq k<2\ell}|\hat{\sigma}_k|(1-\delta^2/5)^\ell\\ \leq&4\,\|\boldsymbol{a}^*\|_1(1-\delta^2/5)^\ell\sum_{\ell\leq k<2\ell}|\hat{\sigma}_k| \leq \frac{1}{4}|a_i^*|\sum_{\ell\leq k<2\ell}|\hat{\sigma}_k|,\end{aligned}$$

where (i) in the third line we use $\cos\delta \leq 1-\delta^2/5$ for $\delta\in[0,\pi/2]$ and (ii) in the last line we use Lemma 18 and choose $\ell = \lceil(5/\delta^2)\log(16\,\|\boldsymbol{a}^*\|_1/|a_i^*|)\rceil$.

Thus, given $|\hat{\sigma}_k| = \Theta(k^{-c_\sigma})$ we have

$$\sqrt{L(\mu)}\sqrt{\ell} = \sqrt{L(\mu)}\,\|g\|_2 \geq \frac{1}{4}|a_i^*|\sum_{\ell\leq k<2\ell}|\hat{\sigma}_k| = \frac{1}{4}|a_i^*|\sum_{\ell\leq k<2\ell}\Theta(k^{-c_\sigma}) = |a_i^*|\Theta(\ell^{1-c_\sigma}).$$

With the choice of $\ell = \widetilde{\Theta}(1/\delta^2)$, we have $\delta = \widetilde{O}\left(\left(\frac{L(\mu)}{|a_i^*|^2}\right)^{1/(4c_\sigma-2)}\right)$. Since $\delta \geq \widetilde{\Omega}\left((\frac{L(\mu)}{a_{\min}^2})^{1/(4c_\sigma-2)}\right)$ with a large enough hidden constant, we know this is a contradiction.

As a corollary, with Lemma 18 that $L(\mu) = 4\zeta + 5\lambda^2\|p\|_2^2$, we have $\delta \geq \widetilde{\Omega}\left((\frac{4\zeta+5\lambda^2\|p\|_2^2}{a_{\min}^2})^{1/(4c_\sigma-2)}\right)$.

For the activation σ is ReLU or absolute function, by Lemma 37 we know $c_\sigma = 5/4$, which gives the desired result. □

The lemma below bounds R_2 using the fact that it is spiky (has small non-zero support).

Lemma 22. *Under Lemma 6, recall the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$. Then*

$$\|R_2\|_2^2 = O_*((\zeta/\lambda+\lambda)^{3/2}).$$

Proof. Using the same calculation as in Lemma 12 in Zhou et al. [28], we have

$$\|R_2\|_2^2 \le O(m_*) \sum_{i\in[m_*]} \left(\sum_{j\in\mathcal{T}_i} |a_j| \, \|\boldsymbol{w}_j\|_2\right)^{1/2} \left(\sum_{j\in\mathcal{T}_i} |a_j| \, \|\boldsymbol{w}_j\|_2 \, \delta_j^2\right)^{3/2}$$

With Lemma 18 and Lemma 19, we have $\|R_2\|_2^2 = O(m_*^2 |\mu^*|^{1/2} (\zeta/\lambda + \lambda)^{3/2})$. □

The following lemma bounds R_3. In fact, in the view of expanding the loss as a sum of tensor decomposition problem, R_3 corresponds to the 0-th order term in the expansion. It would become small when high-order terms become small, as shown in the proof below.

Lemma 23. *Under Lemma 6 and recall the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$. If $\hat{\sigma}_0 = 0$ and $\hat{\sigma}_k > 0$ with some $k = \Theta((1/\Delta^2)\log(\zeta/\|\boldsymbol{a}_*\|_1))$, then*

$$\|R_3\|_2 = \widetilde{O}_*((\zeta + \lambda^2)^{1/2}/\hat{\sigma}_k + (\zeta/\lambda + \lambda) + \zeta).$$

Proof. As shown in Ge et al. [16], Li et al. [20], we can write the loss $L(\mu)$ as sum of tensor decomposition problem (recall $\|\boldsymbol{w}_i^*\|_2 = 1$):

$$L(\mu) = \sum_{k\ge 0} \hat{\sigma}_k^2 \left\| \int_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \boldsymbol{w}^{\otimes k} \, \mathrm{d}\mu(\boldsymbol{w}) - \sum_{i\in[m_*]} a_i^* \, \|\boldsymbol{w}_i^*\|_2 \, \boldsymbol{w}_i^{*\otimes k} \right\|_F^2 .$$

Thus, we know for any $k \ge 1$,

$$\left\| \int_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \boldsymbol{w}^{\otimes k} \, \mathrm{d}\mu(\boldsymbol{w}) - \sum_{i\in[m_*]} a_i^* \, \|\boldsymbol{w}_i^*\|_2 \, \boldsymbol{w}_i^{*\otimes k} \right\|_F^2 \le L(\mu)/\hat{\sigma}_k^2.$$

Given any $\boldsymbol{w}_j^*$, we have

$$\begin{aligned}
&\left\| \int_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \boldsymbol{w}^{\otimes k} \, \mathrm{d}\mu(\boldsymbol{w}) - \sum_{i\in[m_*]} a_i^* \, \|\boldsymbol{w}_i^*\|_2 \, \boldsymbol{w}_i^{*\otimes k} \right\|_F \\
\ge& \left| \left\langle \sum_{i\in[m_*]} a_i^* \, \|\boldsymbol{w}_i^*\|_2 \, \boldsymbol{w}_i^{*\otimes k} - \int_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \boldsymbol{w}^{\otimes k} \, \mathrm{d}\mu(\boldsymbol{w}), \boldsymbol{w}_j^{*\otimes k} \right\rangle \right| \qquad (7) \\
\ge& \left| a_j^* \, \|\boldsymbol{w}_j^*\|_2 - \int_{\mathcal{T}_j} \langle \boldsymbol{w}, \boldsymbol{w}_j^* \rangle^k \, \mathrm{d}\mu(\boldsymbol{w}) \right| - \left| \sum_{i\ne j} a_i^* \, \|\boldsymbol{w}_i^*\|_2 \, \langle \boldsymbol{w}_i^*, \boldsymbol{w}_j^* \rangle^k - \int_{\mathbb{S}^{d-1}\setminus\mathcal{T}_j} \langle \boldsymbol{w}, \boldsymbol{w}_j^* \rangle^k \, \mathrm{d}\mu(\boldsymbol{w}) \right| \\
\ge& \left| a_j^* \, \|\boldsymbol{w}_j^*\|_2 - \int_{\mathcal{T}_j} \mathrm{d}\mu(\boldsymbol{w}) \right| - \left| \int_{\mathcal{T}_j} \mathrm{d}\mu(\boldsymbol{w}) - \int_{\mathcal{T}_j} \langle \boldsymbol{w}, \boldsymbol{w}_j^* \rangle^k \, \mathrm{d}\mu(\boldsymbol{w}) \right| - \left| \sum_{i\ne j} a_i^* \, \|\boldsymbol{w}_i^*\|_2 \, \langle \boldsymbol{w}_i^*, \boldsymbol{w}_j^* \rangle^k - \int_{\mathbb{S}^{d-1}\setminus\mathcal{T}_j} \langle \boldsymbol{w}, \boldsymbol{w}_j^* \rangle^k \, \mathrm{d}\mu(\boldsymbol{w})
\end{aligned}$$

For the second term on RHS, we have

$$\begin{aligned}
\left| \int_{\mathcal{T}_j} \mathrm{d}\mu(\boldsymbol{w}) - \int_{\mathcal{T}_j} \langle \boldsymbol{w}, \boldsymbol{w}_j^* \rangle^k \, \mathrm{d}\mu(\boldsymbol{w}) \right| \le& \int_{\mathcal{T}_j} \left(1 - \langle \boldsymbol{w}, \boldsymbol{w}_j^* \rangle^k\right) \mathrm{d}|\mu|(\boldsymbol{w}) \overset{\text{(a)}}{\le} \int_{\mathcal{T}_j} 1 - (1 - \delta(\boldsymbol{w}, \boldsymbol{w}_j^*)^2/2)^k \, \mathrm{d}|\mu|(\boldsymbol{w}) \\
\overset{\text{(b)}}{\le}& \int_{\mathcal{T}_j, \delta(\boldsymbol{w}, \boldsymbol{w}_j^*)^2 \le 1} O(k) \cdot \delta(\boldsymbol{w}, \boldsymbol{w}_j^*)^2 \, \mathrm{d}|\mu|(\boldsymbol{w}) + \int_{\mathcal{T}_j, \delta(\boldsymbol{w}, \boldsymbol{w}_j^*)^2 > 1} \mathrm{d}|\mu|(\boldsymbol{w}) \\
\le& O(k) \int_{\mathcal{T}_j} \delta(\boldsymbol{w}, \boldsymbol{w}_j^*)^2 \, \mathrm{d}|\mu|(\boldsymbol{w}),
\end{aligned}$$

where (a) $\cos\delta \ge 1 - \delta^2/2$ for $\delta \in [0, \pi/2]$; (b) $(1-x)^k \ge 1 - kx$ for $x \in [0, 1]$.

For the third term on RHS, we have

$$\begin{aligned}\left|\sum_{i\neq j} a_i^* \left\|\boldsymbol{w}_i^*\right\|_2 \langle \boldsymbol{w}_i^*, \boldsymbol{w}_j^*\rangle^k - \int_{\mathbb{S}^{d-1}\setminus\mathcal{T}_j} \langle \boldsymbol{w}, \boldsymbol{w}_j^*\rangle^k \,\mathrm{d}\mu(\boldsymbol{w})\right| \leq & (\|\boldsymbol{a}_*\|_1 + |\mu|_1) \max_{\angle(\boldsymbol{w},\boldsymbol{w}_j^*)\geq\Delta/2} (\boldsymbol{w}^\top \boldsymbol{w}_j^*)^k \\ \overset{\text{(a)}}{\leq} & (\|\boldsymbol{a}_*\|_1 + |\mu|_1)(1-\Delta^2/10)^k \\ \overset{\text{(b)}}{\leq} & O(\zeta),\end{aligned}$$

where (a) $\cos\delta \leq 1 - \delta^2/5$ for $\delta \in [0, \pi/2]$; (b) we choose $k = \Theta((1/\Delta^2)\log(\zeta/\left\|\boldsymbol{a}_*\right\|_1))$ and Lemma 18.

Therefore, we have

$$\left\|\int_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \boldsymbol{w}^{\otimes k}\mu(\boldsymbol{w}) - \sum_{i\in[m_*]} a_i^* \left\|\boldsymbol{w}_i^*\right\|_2 \boldsymbol{w}_i^{*\otimes k}\right\|_F \geq \left|a_j^* \left\|\boldsymbol{w}_j^*\right\|_2 - \int_{\mathcal{T}_j} \mu(\boldsymbol{w})\right| - O(k)\int_{\mathcal{T}_j} \delta(\boldsymbol{w}, \boldsymbol{w}_j^*)^2 |\mu|(\boldsymbol{w}) - O(\zeta).$$

This implies that

$$\begin{aligned} m_*\sqrt{L(\mu)}/\hat{\sigma}_k \geq & \sum_{j\in[m_*]} \left|a_j^* \left\|\boldsymbol{w}_j^*\right\|_2 - \int_{\mathcal{T}_j} \mu(\boldsymbol{w})\right| - O(k) \sum_{j\in[m_*]} \int_{\mathcal{T}_j} \delta(\boldsymbol{w}, \boldsymbol{w}_j^*)^2 |\mu|(\boldsymbol{w}) - O(m_*\zeta) \\ \geq & \left|\sum_{i\in[m_*]} a_i^* \left\|\boldsymbol{w}_i^*\right\|_2 - \int_{\mathbb{S}^{d-1}} \mu(\boldsymbol{w})\right| - \widetilde{O}_*(\zeta/\lambda + \lambda) - O(m_*\zeta),\end{aligned}$$

where we use Lemma 19. Rearranging the terms and recalling $L(\mu) = O_*(\zeta + \lambda^2)$ from Lemma 18, we get the bound.

□

The following lemma gives the bound on the average neuron to its corresponding teacher neuron. It follows directly from the residual decomposition and previous lemmas that characterize R_1, R_2, R_3 respectively.

Lemma 24. *Under Lemma 6, recall the optimality gap $\zeta = L_\lambda(\mu) - L_\lambda(\mu_\lambda^*)$. Then for any $i \in [m_*]$, $\zeta = \Omega(\lambda^2)$ and $\zeta, \lambda \leq 1/\operatorname{poly}(m_*, \Delta, \|\boldsymbol{a}_*\|_1)$*

$$\left\|\sum_{j\in\mathcal{T}_i} a_j \boldsymbol{w}_j - \boldsymbol{w}_i^*\right\|_2 \leq \left(\sum_{i\in[m_*]} \left\|\sum_{j\in\mathcal{T}_i} a_j \boldsymbol{w}_j - \boldsymbol{w}_i^*\right\|_2^2\right)^{1/2} = O_*((\zeta/\lambda)^{3/4}).$$

Proof. With the relation of residual decomposition, Lemma 21, Lemma 22 and Lemma 23, we have for any $i \in [m_*]$

$$\begin{aligned}\Omega(\Delta^{3/2}/m_*^{3/2}) \left(\sum_{i\in[m_*]} \left\|\sum_{j\in\mathcal{T}_i} a_j \boldsymbol{w}_j - \boldsymbol{w}_i^*\right\|_2^2\right)^{1/2} & \leq \|R_1\|_2 \leq \|R\|_2 + \|R_2\|_2 + \|R_3\|_2 \\ = & O_*((\zeta+\lambda^2)^{1/2} + (\zeta/\lambda + \lambda)^{3/4}) + \widetilde{O}_*((\zeta+\lambda^2)^{1/2} + (\zeta/\lambda+\lambda) + \zeta).\end{aligned}$$

Rearranging the terms, we get the result. □

G.2 Omitted proofs in Section E.2

In this section, we give the omitted proofs in Section E.2. The key observation used in the proofs is that balancing the norm and setting $\alpha, \boldsymbol{\beta}$ perfectly to their target values only decrease the loss.

Lemma 25. *Given any $\boldsymbol{\theta} = (\boldsymbol{a}, \boldsymbol{W}, \alpha, \boldsymbol{\beta})$ satisfying $|\alpha - \hat{\alpha}|^2 = O(\zeta)$, $\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 = O(\zeta)$, where $\hat{\alpha} = (1/\sqrt{2\pi}) \sum_{i=1}^m a_i \|\boldsymbol{w}_i\|_2$ and $\hat{\boldsymbol{\beta}} = (1/2) \sum_{i=1}^m a_i \boldsymbol{w}_i$. Let its corresponding balanced version $\boldsymbol{\theta}_{bal} = (\boldsymbol{a}_{bal}, \boldsymbol{W}_{bal}, \alpha_{bal}, \boldsymbol{\beta}_{bal})$ as $a_{bal,i} = \operatorname{sign}(a_i)\sqrt{|a_i| \|\boldsymbol{w}_i\|_2}$, $\boldsymbol{w}_{bal,i} = \overline{\boldsymbol{w}}_i \sqrt{|a_i| \|\boldsymbol{w}_i\|_2}$, $\alpha_{bal} = \hat{\alpha}$ and $\boldsymbol{\beta}_{bal} = \hat{\boldsymbol{\beta}}$. Then, we have*

$$L_\lambda(\boldsymbol{\theta}) - L_\lambda(\boldsymbol{\theta}_{bal}) = |\alpha - \hat{\alpha}|^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 + \frac{\lambda}{2} \sum_{i\in[m]} (|a_i| - \|\boldsymbol{w}_i\|_2)^2 \geq 0.$$

Moreover, let the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*)$, we have results in Lemma 18, Lemma 19, Lemma 20, Lemma 21, Lemma 22, Lemma 23 and Lemma 24 still hold for $L_\lambda(\boldsymbol{\theta})$, with the change of R_3 in (4) as

$$R_3(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}} \left(\sum_{i\in[m_*]} a_i^* \|\boldsymbol{w}_i^*\|_2 - \sum_{i\in[m]} a_i \|\boldsymbol{w}_i\|_2 \right) + \alpha - \hat{\alpha} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{x}.$$

Proof. Recall in Claim 1 we have

$$L(\boldsymbol{\theta}) = |\alpha - \hat{\alpha}|^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 + \sum_{k\geq 2} \hat{\sigma}_k^2 \left\| \sum_{i\in[m]} a_i \|\boldsymbol{w}_i\|_2 \overline{\boldsymbol{w}}_i^{\otimes k} - \sum_{i\in[m_*]} a_i^* \|\boldsymbol{w}_i^*\|_2 \boldsymbol{w}_i^{*\otimes k} \right\|_F^2.$$

Note that $|a_i| \|\boldsymbol{w}_i\|_2 = |a_{bal,i}| \|\boldsymbol{w}_{bal,i}\|_2$ so that $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}_{bal}) + |\alpha - \hat{\alpha}|^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2$. We then have

$$\begin{aligned} L_\lambda(\boldsymbol{\theta}) - L_\lambda(\boldsymbol{\theta}_{bal}) =& |\alpha - \hat{\alpha}|^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{a}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{W}\|_2^2 - \frac{\lambda}{2} \|\boldsymbol{a}_{bal}\|_2^2 - \frac{\lambda}{2} \|\boldsymbol{W}_{bal}\|_2^2 \\ =& |\alpha - \hat{\alpha}|^2 + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 + \frac{\lambda}{2} \sum_{i\in[m]} (|a_i| - \|\boldsymbol{w}_i\|_2)^2. \end{aligned}$$

Therefore, we have the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*) \geq L_\lambda(\boldsymbol{\theta}_{bal}) - L_\lambda(\mu_\lambda^*) = \zeta_{bal}$. Note that $\boldsymbol{\theta}_{bal}$ corresponds to a network that has perfect balanced norms and fitted $\alpha, \boldsymbol{\beta}$, thus all results in Lemma 18, Lemma 19, Lemma 20, Lemma 21, Lemma 22, Lemma 23 and Lemma 24 hold for $\boldsymbol{\theta}_{bal}$. Since $\zeta \geq \zeta_{bal}$, $|a_i| \|\boldsymbol{w}_i\|_2 = |a_{bal,i}| \|\boldsymbol{w}_{bal,i}\|_2$ and $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}_{bal}) + O(\zeta)$, we can easily check that all of them also hold for $\boldsymbol{\theta}$. For the bound on R_3, note that

$$\|R_3\|_2 \leq \frac{1}{\sqrt{2\pi}} \left| \sum_{i\in[m_*]} a_i^* \|\boldsymbol{w}_i^*\|_2 - \sum_{i\in[m]} a_i \|\boldsymbol{w}_i\|_2 \right| + |\alpha - \hat{\alpha}| + \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2$$

so that the same bound still hold for R_3. □

Lemma 26. *Under Lemma 6, suppose optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*)$. Then $\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{W}\|_F^2 \leq 3\|\boldsymbol{a}_*\|_1$.*

Proof. We have

$$\frac{\lambda}{2} \|\boldsymbol{a}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{W}\|_F^2 = \zeta + L(\mu_\lambda^*) + \lambda|\mu_\lambda^*|_1 - L(\boldsymbol{\theta}) \leq \zeta + \lambda^2 \|p\|_2^2 + \lambda|\mu_\lambda^*|_1,$$

where we use Lemma 17. Rearranging the terms, we get the result by noting that $|\mu_\lambda^*|_1 \leq \|\boldsymbol{a}_*\|_1$. □

G.3 Omitted proofs in Section E.3

In this section, we give the omitted proofs in Section E.3. We will consider them case by case.

The lemma below says that one can always decrease the loss if norms are not balanced.

Lemma 27 (Descent direction, norm balance)**.** *We have*

$$
\begin{aligned}
\sum_i \sum_{j\in T_i} \left|\langle \nabla_{a_j} L_\lambda, -a_j\rangle + \langle \nabla_{\boldsymbol{w}_j} L_\lambda, \boldsymbol{w}_j\rangle\right| =& \lambda \sum_{i\in[m_*]} \left|a_i^2 - \|\boldsymbol{w}_i\|_2^2\right| \\
&\geq \max\left\{\lambda|\,\|\boldsymbol{a}\|_2^2 - \|\boldsymbol{W}\|_F^2\,|, \lambda \sum_{i\in[m_*]} (|a_i| - \|\boldsymbol{w}_i\|_2)^2\right\}
\end{aligned}
$$

Proof. We have

$$
\begin{aligned}
&\sum_{i\in[m]} \left|\langle \nabla_{a_j} L_\lambda, -a_j\rangle + \langle \nabla_{\boldsymbol{w}_j} L_\lambda, \boldsymbol{w}_j\rangle\right| \\
=& \sum_{i\in[m]} \left|-2\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - f_*(\boldsymbol{x}))a_j\sigma(\boldsymbol{w}_j^\top \boldsymbol{x})] - \lambda a_j^2 + 2\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - f_*(\boldsymbol{x}))a_j\sigma(\boldsymbol{w}_j^\top \boldsymbol{x})] + \lambda\|\boldsymbol{w}_i\|_2^2\right| \\
=&\lambda \sum_{i\in[m]} \left|a_i^2 - \|\boldsymbol{w}_i\|_2^2\right|
\end{aligned}
$$

Note that $|a_i| + \|\boldsymbol{w}_i\|_2 \geq ||a_i| - \|\boldsymbol{w}_i\|_2\,|$, we get the result. □

The following lemma shows that one can always decrease the loss if there are close-by neurons that cancels with others. Intuitively, reducing such norm cancellation decrease the regularization term while keep the square loss term, which decreasing the total loss as a whole.

Lemma 28 (Descent direction, norm cancellation)**.** *Under Lemma 6, suppose the optimality gap* $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*)$*. For any* $\boldsymbol{w}_i^*$*, consider* δ_{sign} *such that* $\delta_{close} < \delta_{\text{sign}} = O(\lambda/\zeta^{1/2})$ *with small enough hidden constant (*δ_{close} *defined in Lemma 20), then*

$$
\sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} \left\langle \nabla_{a_j} L_\lambda, \frac{\operatorname{sign}(a_j)|a_j|}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2}\right\rangle + \left\langle \nabla_{\boldsymbol{w}_j} L_\lambda, \frac{\boldsymbol{w}_j}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2}\right\rangle = \Omega(\lambda).
$$

where $T_{i,+}(\delta_{\text{sign}}) = \{j \in T_i : \delta(\boldsymbol{w}_j, \boldsymbol{w}_i^*) \leq \delta_{\text{sign}}, \operatorname{sign}(a_j) = \operatorname{sign}(a_i^*)\}$*,* $T_{i,-}(\delta_{\text{sign}}) = \{j \in T_i : \delta(\boldsymbol{w}_j, \boldsymbol{w}_i^*) \leq \delta_{\text{sign}}, \operatorname{sign}(a_j) \neq \operatorname{sign}(a_i^*)\}$ *are the set of neurons that close to* $\boldsymbol{w}_i^*$ *with/without same sign of* a_i^**.*

As a result,

$$
\|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2 \geq \lambda^2 \sum_{j\in T_{i,-}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2
$$

Proof. Denote $R(\boldsymbol{x}) = f(\boldsymbol{x}) - f_*(\boldsymbol{x})$. We have

$$
\begin{aligned}
&\sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} \left\langle \nabla_{a_j} L_\lambda, \frac{\operatorname{sign}(a_j)|a_j|}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2}\right\rangle + \left\langle \nabla_{\boldsymbol{w}_j} L_\lambda, \frac{\boldsymbol{w}_j}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2}\right\rangle \\
=& \sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} \frac{\operatorname{sign}(a_j)|a_j|\,\|\boldsymbol{w}_j\|_2}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2} \cdot 2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma(\overline{\boldsymbol{w}}_j^\top \boldsymbol{x})] + \frac{\lambda a_j^2}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2} \\
&+ \sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} \frac{\operatorname{sign}(a_j)|a_j|\,\|\boldsymbol{w}_j\|_2}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2} \cdot 2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma(\overline{\boldsymbol{w}}_j^\top \boldsymbol{x})] + \frac{\lambda\|\boldsymbol{w}_j\|_2^2}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2}
\end{aligned}
$$

We split the above into two terms. WLOG, assume $\mathrm{sign}(a_i^*) = 1$. For the first term,

$$\begin{aligned}(I) =&4 \sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} \frac{\mathrm{sign}(a_j)|a_j| \left\|\boldsymbol{w}_j\right\|_2}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \cdot \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma(\overline{\boldsymbol{w}}_j^\top \boldsymbol{x})]\\
=&4 \sum_{j\in T_{i,+}(\delta_{\text{sign}})} \frac{|a_j| \left\|\boldsymbol{w}_j\right\|_2}{\sum_{j\in T_{i,+}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma(\overline{\boldsymbol{w}}_j^\top \boldsymbol{x})]\\
&- 4 \sum_{j\in T_{i,-}(\delta_{\text{sign}})} \frac{|a_j| \left\|\boldsymbol{w}_j\right\|_2}{\sum_{j\in T_{i,-}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})\sigma(\overline{\boldsymbol{w}}_j^\top \boldsymbol{x})]\\
=&4 \sum_{j\in T_{i,+}(\delta_{\text{sign}})} \frac{|a_j| \left\|\boldsymbol{w}_j\right\|_2}{\sum_{j\in T_{i,+}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})(\sigma(\overline{\boldsymbol{w}}_j^\top \boldsymbol{x}) - \sigma(\overline{\boldsymbol{w}}_i^{*\top} \boldsymbol{x}))]\\
&- 4 \sum_{j\in T_{i,-}(\delta_{\text{sign}})} \frac{|a_j| \left\|\boldsymbol{w}_j\right\|_2}{\sum_{j\in T_{i,-}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})(\sigma(\overline{\boldsymbol{w}}_j^\top \boldsymbol{x}) - \sigma(\overline{\boldsymbol{w}}_i^{*\top} \boldsymbol{x}))]\end{aligned}$$

Since $\overline{\boldsymbol{w}}_j$ is δ_{sign}-close to $\boldsymbol{w}_i^*$ and $\left\|R\right\|_2^2 = L(\boldsymbol{\theta})$, we have

$$|(I)| \le O(\delta_{\text{sign}}) \left\|R\right\|_2 = O_*(\delta_{\text{sign}}\zeta^{1/2}),$$

where we use Lemma 25 that $L(\boldsymbol{\theta}) = O_*(\zeta)$.

For the second term, we have

$$(II) =\lambda \sum_{s\in\{+,-\}} \frac{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} a_j^2 + \left\|\boldsymbol{w}_j\right\|_2^2}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \ge 2\lambda + 2\lambda = 4\lambda.$$

Therefore, when $(I) \le 2\lambda$, i.e., $\delta_{\text{sign}} = O_*(\lambda/\zeta^{1/2})$, we have

$$\begin{aligned}&\sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} \left\langle \nabla_{a_j} L_\lambda, \frac{\mathrm{sign}(a_j)|a_j|}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \right\rangle + \left\langle \nabla_{\boldsymbol{w}_j} L_\lambda, \frac{\boldsymbol{w}_j}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \right\rangle\\
\ge&\frac{\lambda}{2} \sum_{s\in\{+,-\}} \frac{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} a_j^2 + \left\|\boldsymbol{w}_j\right\|_2^2}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2}.\end{aligned}$$

We compute a upper bound for LHS. Note that

$$\begin{aligned}&\sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} \left\langle \nabla_{a_j} L_\lambda, \frac{\mathrm{sign}(a_j)|a_j|}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \right\rangle + \left\langle \nabla_{\boldsymbol{w}_j} L_\lambda, \frac{\boldsymbol{w}_j}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2} \right\rangle\\
\le&\sqrt{\sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} (\nabla_{a_j} L_\lambda)^2 + \left\|\nabla_{\boldsymbol{w}_j} L_\lambda\right\|_2^2} \sqrt{\sum_{s\in\{+,-\}} \sum_{j\in T_{i,s}(\delta_{\text{sign}})} \frac{a_j^2 + \left\|\boldsymbol{w}_j\right\|_2^2}{(\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2)^2}}\\
\le&\sqrt{\left\|\nabla_{\boldsymbol{a}} L_\lambda\right\|_2^2 + \left\|\nabla_{\boldsymbol{W}} L_\lambda\right\|_F^2} \sqrt{\sum_{s\in\{+,-\}} \frac{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} a_j^2 + \left\|\boldsymbol{w}_j\right\|_2^2}{(\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2)^2}}\\
\le&\sqrt{\left\|\nabla_{\boldsymbol{a}} L_\lambda\right\|_2^2 + \left\|\nabla_{\boldsymbol{W}} L_\lambda\right\|_F^2} \sqrt{\sum_{s\in\{+,-\}} \frac{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} a_j^2 + \left\|\boldsymbol{w}_j\right\|_2^2}{\sum_{j\in T_{i,s}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2}} \frac{1}{\sqrt{\sum_{j\in T_{i,-}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2}},\end{aligned}$$

where the last line we use Lemma 20: $\sum_{j\in T_{i,-}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2 < \sum_{j\in T_{i,+}(\delta_{\text{sign}})} |a_j| \left\|\boldsymbol{w}_j\right\|_2$ because $\mu(T_i(\delta)) = \sum_{j\in T_i(\delta_{\text{sign}})} a_j \left\|\boldsymbol{w}_j\right\|_2 > 0$.

Combine with the above descent direction, we have

$$\begin{aligned}
&\sqrt{\|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2}\sqrt{\sum_{s\in\{+,-\}} \frac{\sum_{j\in T_{i,s}(\delta_{\mathrm{sign}})} a_j^2 + \|\boldsymbol{w}_j\|_2^2}{\sum_{j\in T_{i,s}(\delta_{\mathrm{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2}}\, \frac{1}{\sqrt{\sum_{j\in T_{i,-}(\delta_{\mathrm{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2}}\\
\geq&\frac{\lambda}{2}\sum_{s\in\{+,-\}} \frac{\sum_{j\in T_{i,s}(\delta_{\mathrm{sign}})} a_j^2 + \|\boldsymbol{w}_j\|_2^2}{\sum_{j\in T_{i,s}(\delta_{\mathrm{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2},
\end{aligned}$$

which implies

$$\|\nabla_{\boldsymbol{a}} L_\lambda\|_2^2 + \|\nabla_{\boldsymbol{W}} L_\lambda\|_F^2 \geq \lambda^2 \sum_{j\in T_{i,-}(\delta_{\mathrm{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2$$

□

The lemma below shows that when all previous cases are not hold, then there is a descent direction that move all close-by neurons towards their corresponding teacher neuron. The proof relies on calculations that generalize Lemma 8 in [28].

Lemma 29 (Descent direction)**.** *Under Lemma 6, suppose the optimality gap $\zeta = L_\lambda(\boldsymbol{\theta}) - L_\lambda(\mu_\lambda^*)$. Suppose*

(i) norms are (almost) balanced: $|\,\|\boldsymbol{W}\|_F^2 - \|\boldsymbol{a}\|_2^2\,| \leq \zeta/\lambda$, $\sum_{i\in[m]}(|a_j| - \|\boldsymbol{w}_j\|_2)^2 = O_*(\zeta^2/\lambda^2)$

(ii) (almost) no norm cancellation: consider all neurons $\boldsymbol{w}_j$ that are δ_{sign}-close w.r.t. teacher neuron $\boldsymbol{w}_i^*$ but has a different sign, i.e., $\mathrm{sign}(a_j) \neq \mathrm{sign}(a_i^*)$ with $\delta_{\mathrm{sign}} = \Theta_*(\lambda/\zeta^{1/2})$, we have $\sum_{j\in T_{i,-}(\delta_{\mathrm{sign}})} |a_j|\,\|\boldsymbol{w}_j\|_2 \leq \tau = O_*(\zeta^{5/6}/\lambda)$ with small enough hidden constant, where $T_{i,-}(\delta)$ defined in Lemma 28.

(iii) $\alpha, \boldsymbol{\beta}$ are well fitted: $|\alpha - \hat{\alpha}|^2 = O(\zeta)$, $\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2^2 = O(\zeta)$ with small enough hidden constant.

Then, we can construct the following descent direction

$$(\alpha + \alpha_*)\nabla_\alpha L_\lambda + \langle \nabla_{\boldsymbol{\beta}} L_\lambda, \boldsymbol{\beta} + \boldsymbol{\beta}_*\rangle + \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \langle \nabla_{\boldsymbol{w}_i} L_\lambda, \boldsymbol{w}_j - q_{ij}\boldsymbol{w}_i^*\rangle = \Omega(\zeta),$$

where q_{ij} satisfy the following conditions with $\delta_{close} < \delta_{\mathrm{sign}}$ and $\delta_{close} = O_*(\zeta^{1/3})$: (1) $\sum_{j\in\mathcal{T}_i} a_j q_{ij} = a_i^*$; (2) $q_{ij} \geq 0$; (3) $q_{ij} = 0$ when $\mathrm{sign}(a_j) \neq \mathrm{sign}(a_i^*)$ or $\delta_j > \delta_{close}$. (4) $\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}^2 = O_*(1)$.

Proof. Recall residual $R(\boldsymbol{x}) = f(\boldsymbol{x}) - f_*(\boldsymbol{x})$. We have

$$
\begin{aligned}
&(\alpha + \alpha_*)\nabla_\alpha L_\lambda + \langle \nabla_{\boldsymbol{\beta}} L_\lambda, \boldsymbol{\beta} + \boldsymbol{\beta}_*\rangle + \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \langle \nabla_{\boldsymbol{w}_i} L_\lambda, \boldsymbol{w}_j - q_{ij}\boldsymbol{w}_i^*\rangle \\
\overset{\text{(a)}}{=}& 2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})(\alpha + \alpha_*)] + 2\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})(\boldsymbol{\beta} + \boldsymbol{\beta}_*)^\top \boldsymbol{x}] + 2\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x}) a_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x})] - 2\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x}) a_j q_{ij} \sigma(\boldsymbol{w}_i^{*\top} \boldsymbol{x} \\
&+ 2\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x}) a_j q_{ij} \boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_i^\top \boldsymbol{x}))] \\
&+ \lambda \sum_{i\in[m]} \|\boldsymbol{w}_j\|_2^2 - \lambda \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}\boldsymbol{w}_j^\top \boldsymbol{w}_i^* \\
\overset{\text{(b)}}{=}& 2\|R\|_2^2 + \lambda\|\boldsymbol{W}\|_F^2 - \lambda \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}\boldsymbol{w}_j^\top \boldsymbol{w}_i^* \\
&+ 2\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x}) a_j q_{ij} \boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_j^\top \boldsymbol{x}))] \\
\overset{\text{(c)}}{\ge}& L_\lambda(\mu_\lambda^*) + \zeta + \frac{\lambda}{2}(\|\boldsymbol{W}\|_F^2 - \|\boldsymbol{a}\|_2^2) - \lambda \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}\|\boldsymbol{w}_j\|_2 \\
&+ 2\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x}) a_j q_{ij} \boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_j^\top \boldsymbol{x}))], \qquad (8)
\end{aligned}
$$

where (a) we plug in the gradient expression and add and minus the term $2\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})a_j q_{ij}\sigma(\boldsymbol{w}_i^{*\top}\boldsymbol{x})]$; (b) rearranging the terms; (c) using $L_\lambda(\boldsymbol{\theta}) = \|R\|_2^2 + (\lambda/2)\|\boldsymbol{W}\|_F^2 + (\lambda/2)\|\boldsymbol{a}\|_2^2 = L_\lambda(\mu_\lambda^*) + \zeta$.

For the first line on RHS of (8), we have

$$
\begin{aligned}
&L_\lambda(\mu_\lambda^*) + \zeta + \frac{\lambda}{2}(\|\boldsymbol{W}\|_F^2 - \|\boldsymbol{a}\|_2^2) - \lambda \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}\|\boldsymbol{w}_j\|_2 \\
\overset{\text{(a)}}{\ge}& \zeta/2 + L(\mu_\lambda^*) + \lambda|\mu_\lambda^*| - \lambda \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}\|\boldsymbol{w}_j\|_2 \\
\overset{\text{(b)}}{\ge}& \zeta/2 + \lambda|\mu_\lambda^*| - \lambda\|\boldsymbol{a}_*\|_1 + \lambda \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}(|a_j| - \|\boldsymbol{w}_j\|_2) \\
\overset{\text{(c)}}{\ge}& \zeta/2 - O_*(\lambda^2) - \lambda\left(\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}^2\right)^{1/2}\left(\sum_{i\in[m]}(|a_j| - \|\boldsymbol{w}_j\|_2)^2\right)^{1/2} \\
\overset{\text{(d)}}{\ge}& \zeta/4,
\end{aligned}
$$

where (a) due to assumption that norms are balanced; (b) we ignore $L(\mu_\lambda^*)$ and add and minus $\lambda\|\boldsymbol{a}_*\|_1$; (c) due to Lemma 17; (d) due to assumption that norms are balanced and the choice of q_{ij}.

In the following, we will lower bound the last term of (8) to show it is much smaller than $\zeta/8$. Recall the residual decomposition (4) that $R(\boldsymbol{x}) = R_1(\boldsymbol{x}) + R_2(\boldsymbol{x}) + R_3(\boldsymbol{x})$, we have

$$
\begin{aligned}
&\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x}) a_j q_{ij} \boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_j^\top\boldsymbol{x}) - \sigma'(\boldsymbol{w}_i^{*\top} \boldsymbol{x}))] \\
=& \underbrace{\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R_1(\boldsymbol{x}) a_j q_{ij} \boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_j^\top \boldsymbol{x}))]}_{(I)} + \underbrace{\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R_2(\boldsymbol{x}) a_j q_{ij} \boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_j^\top \boldsymbol{x}))]}_{(II)} \\
&+ \underbrace{\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} \mathbb{E}_{\boldsymbol{x}}[R_3(\boldsymbol{x}) a_j q_{ij} \boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_j^\top \boldsymbol{x}))]}_{(III)}
\end{aligned}
$$

Bound (I) For (I), recall $R_1(\boldsymbol{x}) = (1/2)\sum_{i\in[m_*]} \boldsymbol{v}_i^\top \boldsymbol{x}\,\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})$, where $\boldsymbol{v}_i = \sum_{j\in\mathcal{T}_i} a_j \boldsymbol{w}_j - \boldsymbol{w}_i^*$ and $(\sum_{i\in[m_*]}\|\boldsymbol{v}_i\|_2^2)^{1/2} = O_*((\zeta/\lambda)^{3/4})$ from Lemma 24 and Lemma 25. We have

$$\begin{aligned}
&\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\mathbb{E}_{\boldsymbol{x}}[R_1(\boldsymbol{x})a_j q_{ij}\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_j^\top\boldsymbol{x}))]\\
\overset{\text{(a)}}{\geq}& -\frac{1}{2}\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\sum_{k\in[m_*]}\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{v}_k^\top\boldsymbol{x}||a_j q_{ij}||\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})}]\\
\overset{\text{(b)}}{=}& -\frac{1}{2}\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\sum_{k\in[m_*]}|a_j q_{ij}|\,\|\boldsymbol{v}_k\|_2\,\mathbb{E}_{\widetilde{\boldsymbol{x}}}[|\overline{\boldsymbol{v}}_k^\top\widetilde{\boldsymbol{x}}||\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_j^\top\widetilde{\boldsymbol{x}})\neq\mathrm{sign}(\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}})}]\\
\overset{\text{(c)}}{\geq}& -\frac{1}{2}\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\sum_{k\in[m_*]}|a_j q_{ij}|\,\|\boldsymbol{v}_k\|_2\,\delta_j\mathbb{E}_{\widetilde{\boldsymbol{x}}}[\|\widetilde{\boldsymbol{x}}\|_2^2\,\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_j^\top\widetilde{\boldsymbol{x}})\neq\mathrm{sign}(\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}})}]\\
\overset{\text{(d)}}{\geq}& -\frac{1}{2}\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\sum_{k\in[m_*]}|a_j q_{ij}|\,\|\boldsymbol{v}_k\|_2\,\Theta(\delta_j^2)\\
\overset{\text{(e)}}{\geq}& -\Theta_*((\zeta/\lambda)^{3/4}\delta_{close}^2)\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}|a_j q_{ij}| = -\Theta_*((\zeta/\lambda)^{3/4}\delta_{close}^2),
\end{aligned}$$

where in (a) we plug in the definition of R_1 and using the fact that $\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})-\sigma'(\boldsymbol{w}_j^\top\boldsymbol{x})) = |\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})}$; (b) $\widetilde{\boldsymbol{x}}$ is a 3-dimensional Gaussian since the expectation only depends on $\boldsymbol{v}_k, \boldsymbol{w}_i^*, \boldsymbol{w}_j$; (c) $|\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}}| \leq \delta_j\,\|\widetilde{\boldsymbol{x}}\|_2$ when $\mathrm{sign}(\boldsymbol{w}_j^\top\widetilde{\boldsymbol{x}}) \neq \mathrm{sign}(\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}})$; (d) a direct calculation bound as Lemma 34; (e) definition of q_{ij}.

Bound (II) For (II), recall

$$R_2(\boldsymbol{x}) = \frac{1}{2}\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}a_j\boldsymbol{w}_j^\top\boldsymbol{x}(\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})-\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})) = \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}a_j|\boldsymbol{w}_j^\top\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x}))}.$$

For each term in (II) with $j \in \mathcal{T}_i$, we can split it into two terms that corresponding to $\mathcal{T}_i$ and other $\mathcal{T}_k$'s

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}[R_2(\boldsymbol{x})a_j q_{ij}\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_j^\top\boldsymbol{x}))]\\
=&\sum_{k\in[m_*]}\sum_{\ell\in\mathcal{T}_k}\mathbb{E}_{\boldsymbol{x}}[a_\ell|\boldsymbol{w}_\ell^\top\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_\ell^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_k^{*\top}\boldsymbol{x})}\cdot a_j q_{ij}\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}) - \sigma'(\boldsymbol{w}_j^\top\boldsymbol{x})]\\
=&\sum_{k\in[m_*]}\sum_{\ell\in\mathcal{T}_k}a_\ell a_j q_{ij}\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_\ell^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_k^{*\top}\boldsymbol{x})}\cdot|\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})}]\\
=&\underbrace{\sum_{\ell\in\mathcal{T}_i}a_\ell a_j q_{ij}\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top\boldsymbol{x}||\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_\ell^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})}\cdot\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})}]}_{(II.i)}\\
&+\underbrace{\sum_{k\neq i}\sum_{\ell\in\mathcal{T}_k}a_\ell a_j q_{ij}\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top\boldsymbol{x}||\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_\ell^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_k^{*\top}\boldsymbol{x})}\cdot\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})}]}_{(II.ii)}. \qquad (9)
\end{aligned}$$

For (II.i), we further split neurons into $\mathcal{T}_i(\delta_{\mathrm{sign}})$ and others:

$$\begin{aligned}
(II.i) =& \sum_{\ell\in\mathcal{T}_i(\delta_{\mathrm{sign}})}a_\ell a_j q_{ij}\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top\boldsymbol{x}||\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_\ell^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})}\cdot\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})}]\\
&+\sum_{\ell\in\mathcal{T}_i\setminus\mathcal{T}_i(\delta_{\mathrm{sign}})}a_\ell a_j q_{ij}\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top\boldsymbol{x}||\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_\ell^\top\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})}\cdot\mathbb{1}_{\mathrm{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\neq\mathrm{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})}]
\end{aligned}\qquad (10)$$

Consider the first line of (10), from the choice of q_{ij} we know $a_j q_{ij} a_i^* \geq 0$. For $\ell \in \mathcal{T}_{i,+}(\delta_{\text{sign}})$, we know $\text{sign}(a_\ell) = \text{sign}(a_i^*)$, which implies $a_\ell a_j q_{ij} \geq 0$ for these terms. We thus only need to deal with neurons in $T_{i,-}(\delta_{\text{sign}})$, we have the first line is bounded as

$$
\begin{aligned}
&\sum_{\ell \in \mathcal{T}_i(\delta_{\text{sign}})} a_\ell a_j q_{ij} \mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top \boldsymbol{x}||\boldsymbol{w}_i^{*\top} \boldsymbol{x}| \mathbb{1}_{\text{sign}(\boldsymbol{w}_\ell^\top \boldsymbol{x}) \neq \text{sign}(\boldsymbol{w}_i^{*\top} \boldsymbol{x})} \cdot \mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top} \boldsymbol{x}) \neq \text{sign}(\boldsymbol{w}_j^\top \boldsymbol{x})}] \\
\geq &\sum_{\ell \in \mathcal{T}_{i,-}(\delta_{\text{sign}})} a_\ell a_j q_{ij} \mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top \boldsymbol{x}||\boldsymbol{w}_i^{*\top} \boldsymbol{x}| \mathbb{1}_{\text{sign}(\boldsymbol{w}_\ell^\top \boldsymbol{x}) \neq \text{sign}(\boldsymbol{w}_i^{*\top} \boldsymbol{x})} \cdot \mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top} \boldsymbol{x}) \neq \text{sign}(\boldsymbol{w}_j^\top \boldsymbol{x})}] \\
\overset{(a)}{\geq} &- |a_j q_{ij}| \sum_{\ell \in \mathcal{T}_{i,-}(\delta_{\text{sign}})} |a_\ell| \left\|\boldsymbol{w}_\ell\right\|_2 \mathbb{E}_{\boldsymbol{x}}[|\overline{\boldsymbol{w}}_\ell^\top \widetilde{\boldsymbol{x}}||\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}| \mathbb{1}_{\text{sign}(\boldsymbol{w}_\ell^\top \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}})} \cdot \mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_j^\top \widetilde{\boldsymbol{x}})}] \\
\overset{(b)}{\geq} &- |a_j q_{ij}| \sum_{\ell \in \mathcal{T}_{i,-}(\delta_{\text{sign}})} |a_\ell| \left\|\boldsymbol{w}_\ell\right\|_2 \delta_\ell \delta_j \mathbb{E}_{\boldsymbol{x}}[\left\|\widetilde{\boldsymbol{x}}\right\|_2^2 \mathbb{1}_{\text{sign}(\boldsymbol{w}_\ell^\top \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}})} \cdot \mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_j^\top \widetilde{\boldsymbol{x}})}] \\
\overset{(c)}{\geq} &- |a_j q_{ij}| \sum_{\ell \in \mathcal{T}_{i,-}(\delta_{\text{sign}})} |a_\ell| \left\|\boldsymbol{w}_\ell\right\|_2 O(\delta_\ell \delta_j^2) \\
\overset{(d)}{\geq} &- |a_j q_{ij}| O(\tau \delta_{\text{sign}} \delta_{close}^2),
\end{aligned}
$$

where (a) $\widetilde{\boldsymbol{x}}$ is a 3-dimensional Gaussian since the expectation only depends on $\boldsymbol{w}_\ell, \boldsymbol{w}_j, \boldsymbol{w}_i^*$; (b) $|\overline{\boldsymbol{w}}_\ell^\top \widetilde{\boldsymbol{x}}| \leq \delta_\ell \left\|\widetilde{\boldsymbol{x}}\right\|_2$ when $\text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_\ell^\top \widetilde{\boldsymbol{x}})$ and $|\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}| \leq \delta_j \left\|\widetilde{\boldsymbol{x}}\right\|_2$ when $\text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_j^\top \widetilde{\boldsymbol{x}})$; (c) a direct calculation as in Lemma 34; (d) assumption that norm cancellation is small.

For the second term of (10), similar as above, we have

$$
\begin{aligned}
&2 \sum_{\ell \in \mathcal{T}_i \setminus \mathcal{T}_i(\delta_{\text{sign}})} a_\ell a_j q_{ij} \mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top \boldsymbol{x}||\boldsymbol{w}_i^{*\top} \boldsymbol{x}| \mathbb{1}_{\text{sign}(\boldsymbol{w}_\ell^\top \boldsymbol{x}) \neq \text{sign}(\boldsymbol{w}_i^{*\top} \boldsymbol{x})} \cdot \mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top} \boldsymbol{x}) \neq \text{sign}(\boldsymbol{w}_j^\top \boldsymbol{x})}] \\
\overset{(a)}{\geq} &- 2|a_j q_{ij}| \sum_{\ell \in \mathcal{T}_i \setminus \mathcal{T}_i(\delta_{\text{sign}})} |a_\ell| \left\|\boldsymbol{w}_\ell\right\|_2 \mathbb{E}_{\widetilde{\boldsymbol{x}}}[|\overline{\boldsymbol{w}}_\ell^\top \widetilde{\boldsymbol{x}}||\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}| \mathbb{1}_{\text{sign}(\boldsymbol{w}_\ell^\top \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}})} \cdot \mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_j^\top \widetilde{\boldsymbol{x}})}] \\
\overset{(b)}{\geq} &- 2|a_j q_{ij}| \sum_{\ell \in \mathcal{T}_i \setminus \mathcal{T}_i(\delta_{\text{sign}})} |a_\ell| \left\|\boldsymbol{w}_\ell\right\|_2 \delta_\ell \delta_j \mathbb{E}_{\widetilde{\boldsymbol{x}}}[\left\|\widetilde{\boldsymbol{x}}\right\|_2^2 \mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_j^\top \widetilde{\boldsymbol{x}})}] \\
\overset{(c)}{\geq} &- 2|a_j q_{ij}| O(\delta_j^2) \sum_{\ell \in \mathcal{T}_i \setminus \mathcal{T}_i(\delta_{\text{sign}})} |a_\ell| \left\|\boldsymbol{w}_\ell\right\|_2 \delta_\ell \\
\overset{(d)}{\geq} &- 2|a_j q_{ij}| O_*(\delta_{close}^2 \zeta \lambda^{-1} \delta_{\text{sign}}^{-1}),
\end{aligned}
$$

where (a) $\widetilde{\boldsymbol{x}}$ is 3-dimensional Gaussian vector since the expectation only depends on $\boldsymbol{w}_\ell, \boldsymbol{w}_j, \boldsymbol{w}_i^*$; (b) $|\overline{\boldsymbol{w}}_\ell^\top \widetilde{\boldsymbol{x}}| \leq \delta_\ell \left\|\widetilde{\boldsymbol{x}}\right\|_2$ when $\text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_\ell^\top \widetilde{\boldsymbol{x}})$ and $|\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}| \leq \delta_j \left\|\widetilde{\boldsymbol{x}}\right\|_2$ when $\text{sign}(\boldsymbol{w}_i^{*\top} \widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_j^\top \widetilde{\boldsymbol{x}})$; (c) a direct calculation as in Lemma 34; (d) choice of q_{ij} and Lemma 19 and Lemma 25 that far-away neurons are small.

Thus, for (II.i) we have

$$
(II.i) \geq -2|a_j q_{ij}| O_*(\tau \delta_{\text{sign}} \delta_{close}^2 + \delta_{close}^2 \zeta \lambda^{-1} \delta_{\text{sign}}^{-1}).
$$

For (II.ii), we have

$$\begin{aligned}
|(II.ii)| \leq & 2\sum_{k\neq i}\sum_{\ell\in\mathcal{T}_k} |a_\ell||a_j q_{ij}|\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_\ell^\top \boldsymbol{x}||\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\text{sign}(\boldsymbol{w}_\ell^\top \boldsymbol{x})\neq\text{sign}(\boldsymbol{w}_k^{*\top}\boldsymbol{x})}\cdot\mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\neq\text{sign}(\boldsymbol{w}_j^\top \boldsymbol{x})}] \\
\overset{(a)}{\leq} & 2\sum_{k\neq i}\sum_{\ell\in\mathcal{T}_k} |a_\ell||a_j q_{ij}|\left\|\boldsymbol{w}_\ell\right\|_2 \delta_\ell\delta_j \mathbb{E}_{\widetilde{\boldsymbol{x}}}[\left\|\widetilde{\boldsymbol{x}}\right\|_2^2 \mathbb{1}_{\text{sign}(\boldsymbol{w}_\ell^\top \widetilde{\boldsymbol{x}})\neq\text{sign}(\boldsymbol{w}_k^{*\top}\widetilde{\boldsymbol{x}})}\cdot\mathbb{1}_{\text{sign}(\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}})\neq\text{sign}(\boldsymbol{w}_j^\top \widetilde{\boldsymbol{x}})}] \\
\overset{(b)}{\leq} & 2\sum_{k\neq i}\sum_{\ell\in\mathcal{T}_k} |a_\ell||a_j q_{ij}|\left\|\boldsymbol{w}_\ell\right\|_2 \delta_\ell\delta_j \mathbb{E}_{\widetilde{\boldsymbol{x}}}[\left\|\widetilde{\boldsymbol{x}}\right\|_2^2 \mathbb{1}_{|\boldsymbol{w}_k^{*\top}\widetilde{\boldsymbol{x}}|\leq\delta_\ell\left\|\widetilde{\boldsymbol{x}}\right\|_2}\cdot\mathbb{1}_{|\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}}|\leq\delta_j\left\|\widetilde{\boldsymbol{x}}\right\|_2}] \\
\overset{(c)}{\leq} & 2|a_j q_{ij}|\delta_j \sum_{k\neq i}\sum_{\ell\in\mathcal{T}_k} |a_\ell|\left\|\boldsymbol{w}_\ell\right\|_2 \delta_\ell\cdot O(\delta_\ell\delta_j/\Delta) \\
\overset{(d)}{=} & 2|a_j q_{ij}|O_*(\delta_{close}^2\zeta\lambda^{-1}\Delta^{-1}),
\end{aligned}$$

where (a)(b) $\widetilde{\boldsymbol{x}}$ is a 4-dimensional Gaussian vector, $|\overline{\boldsymbol{w}}_\ell^\top\widetilde{\boldsymbol{x}}| \leq \delta_\ell\left\|\widetilde{\boldsymbol{x}}\right\|_2$ when $\text{sign}(\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_\ell^\top\widetilde{\boldsymbol{x}})$ and $|\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}}| \leq \delta_j\left\|\widetilde{\boldsymbol{x}}\right\|_2$ when $\text{sign}(\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_j^\top\widetilde{\boldsymbol{x}})$; (c) by Lemma 33; (d) choice of q_{ij} and Lemma 19 and Lemma 25 that far-away neurons are small.

Combine (II.i) (II.ii), we have for (9)

$$\mathbb{E}_{\boldsymbol{x}}[R_2(\boldsymbol{x})a_j q_{ij}\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})-\sigma'(\boldsymbol{w}_j^\top\boldsymbol{x}))] \geq -2|a_j q_{ij}|O(\tau\delta_{\text{sign}}\delta_{close}^2+\delta_{close}^2\zeta\lambda^{-1}\delta_{\text{sign}}^{-1}).$$

This further gives the lower bound on (II):

$$\begin{aligned}
\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\mathbb{E}_{\boldsymbol{x}}[R_2(\boldsymbol{x})a_j q_{ij}\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})-\sigma'(\boldsymbol{w}_j^\top\boldsymbol{x}))] \geq & -2\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}|a_j q_{ij}|O(\tau\delta_{\text{sign}}\delta_{close}^2+\delta_{close}^2\zeta\lambda^{-1}\delta_{\text{sign}}^{-1}) \\
= & -O_*(\tau\delta_{\text{sign}}\delta_{close}^2+\delta_{close}^2\zeta\lambda^{-1}\delta_{\text{sign}}^{-1})
\end{aligned}$$

Bound (III) For (III), recall $R_3(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}}\left(\sum_{i\in[m_*]} a_i^*\left\|\boldsymbol{w}_i^*\right\|_2 - \sum_{i\in[m]} a_i\left\|\boldsymbol{w}_i\right\|_2\right) + \alpha - \hat{\alpha} + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{x}$. We have

$$\begin{aligned}
& \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\mathbb{E}_{\boldsymbol{x}}[R_3(\boldsymbol{x})a_j q_{ij}\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})-\sigma'(\boldsymbol{w}_j^\top\boldsymbol{x}))] \\
\overset{(a)}{\geq} & -O_*(\zeta/\lambda)\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}|a_j q_{ij}|\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\text{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})\neq\text{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})}] \\
& -\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}|a_j q_{ij}|\mathbb{E}_{\boldsymbol{x}}[|(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{x}||\boldsymbol{w}_i^{*\top}\boldsymbol{x}|\mathbb{1}_{\text{sign}(\boldsymbol{w}_j^\top\boldsymbol{x})\neq\text{sign}(\boldsymbol{w}_i^{*\top}\boldsymbol{x})}] \\
\overset{(b)}{\geq} & -O_*(\zeta/\lambda)\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}|a_j q_{ij}|O(\delta_j^2) \\
& -O(\zeta^{1/2})\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}|a_j q_{ij}|\delta_j\mathbb{E}_{\boldsymbol{x}}[\left\|\widetilde{\boldsymbol{x}}\right\|_2^2\mathbb{1}_{\text{sign}(\boldsymbol{w}_j^\top\widetilde{\boldsymbol{x}})\neq\text{sign}(\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}})}] \\
\overset{(c)}{\geq} & -O_*(\zeta/\lambda)\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}|a_j q_{ij}|O(\delta_j^2) \\
\overset{(d)}{\geq} & -O_*(\delta_{close}^2\zeta/\lambda),
\end{aligned}$$

where (a) plugging in the expression of R_3 and using Lemma 23 and Lemma 25; (b) using Lemma 35 and the fact that $\widetilde{\boldsymbol{x}}$ is a 3-dimensional Gaussian vector and $|\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}}| \leq \delta_j\left\|\widetilde{\boldsymbol{x}}\right\|_2$ when $\text{sign}(\boldsymbol{w}_i^{*\top}\widetilde{\boldsymbol{x}}) \neq \text{sign}(\boldsymbol{w}_j^\top\widetilde{\boldsymbol{x}})$; (c) Lemma 34; (d) choice of q_{ij}.

Combine all bounds Combine (I) (II) (III) we now get the last term of (8)

$$\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})a_j q_{ij}\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_j^\top\boldsymbol{x})-\sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}))] \geq -O_*((\zeta/\lambda)^{3/4}\delta_{close}^2+\tau\delta_{\text{sign}}\delta_{close}^2+\delta_{close}^2\zeta\lambda^{-1}\delta_{\text{sign}}^{-1})$$

From Lemma 20 we can choose $\delta_{close} = O_*(\zeta^{1/3})$ and from Lemma 28 we can choose $\delta_{\text{sign}} = \Theta_*(\lambda/\zeta^{1/2})$. Also with $\tau = O(\zeta^{5/6}/\lambda)$, we finally get

$$\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x})a_j q_{ij}\boldsymbol{w}_i^{*\top}\boldsymbol{x}(\sigma'(\boldsymbol{w}_j^\top\boldsymbol{x}) - \sigma'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}))] \geq \zeta/8,$$

as long as $\zeta = O(\lambda^{9/5}/\operatorname{poly}(r, m_*, \Delta, \|\boldsymbol{a}_*\|_1, a_{\min}))$ with small enough hidden constant.

Thus, we eventually get the lower bound of (8)

$$(\alpha + \alpha_*)\nabla_\alpha L_\lambda + \langle\nabla_{\boldsymbol{\beta}} L_\lambda, \boldsymbol{\beta} + \boldsymbol{\beta}_*\rangle + \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i}\langle\nabla_{\boldsymbol{w}_i} L_\lambda, \boldsymbol{w}_j - q_{ij}\boldsymbol{w}_i^*\rangle \geq \zeta/4 - \zeta/8 = \zeta/8.$$

□

G.4 Technical Lemma

In this section, we collect several technical lemmas that are useful in the proof.

Lemma 33. *Consider $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^4$ with $\phi = \angle(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in [0, \pi]$ and $\|\boldsymbol{\alpha}\|_2 = \|\boldsymbol{\beta}\|_2 = 1$ and $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Then, for any $0 < \delta_1, \delta_2 \leq c\phi$ with a small enough constant c*

$$\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{x}\|_2^2\,\mathbb{1}_{|\boldsymbol{\alpha}^\top\boldsymbol{x}|\leq\delta_1\|\boldsymbol{x}\|_2, |\boldsymbol{\beta}^\top\boldsymbol{x}|\leq\delta_2\|\boldsymbol{x}\|_2}] = O(\delta_1\delta_2/\sin\phi).$$

Proof. WLOG, assume $\boldsymbol{\alpha} = (1, 0, 0, 0)^\top$, $\boldsymbol{\beta} = (\cos\phi, \sin\phi, 0, 0)$ and $\phi \in [0, \pi/2]$. Then we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{x}\|_2^2\,\mathbb{1}_{|\boldsymbol{\alpha}^\top\boldsymbol{x}|\leq\delta_1\|\boldsymbol{x}\|_2, |\boldsymbol{\beta}^\top\boldsymbol{x}|\leq\delta_2\|\boldsymbol{x}\|_2}]\\
=&\frac{1}{(2\pi)^2}\int_0^\infty r^5 e^{-r^2/2}\,\mathrm{d}r\int_{0\leq\theta_1\leq\pi, |\cos\theta_1|\leq\delta_1}\sin^2\theta_1\int_{0\leq\theta_2\leq\pi, |\cos\theta_1\cos\phi+\sin\theta_1\cos\theta_2\sin\phi|\leq\delta_2}\sin\theta_2\,\mathrm{d}\theta_2\,\mathrm{d}\theta_1\int_0^{2\pi}1\,\mathrm{d}\theta_3\\
=&O(1)\cdot\int_{0\leq\theta_1\leq\pi, |\cos\theta_1|\leq\delta_1}\sin^2\theta_1\int_{0\leq\theta_2\leq\pi, \frac{-\delta_2-\cos\theta_1\cos\phi}{\sin\theta_1\sin\phi}\leq\cos\theta_2\leq\frac{\delta_2-\cos\theta_1\cos\phi}{\sin\theta_1\sin\phi}}\sin\theta_2\,\mathrm{d}\theta_2\,\mathrm{d}\theta_1\\
=&\int_{0\leq\theta_1\leq\pi, |\cos\theta_1|\leq\delta_1}\sin^2\theta_1\cdot O\left(\frac{\delta_2}{\sin\theta_1\sin\phi}\right)\mathrm{d}\theta_1\\
=&O\left(\frac{\delta_1\delta_2}{\sin\phi}\right).
\end{aligned}$$

□

Lemma 34 (Lemma C.9 in [28]). *Consider $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^3$ with $\angle(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \phi$ and $\boldsymbol{\alpha}^\top\boldsymbol{\beta} \geq 0$. We have*

$$\mathbb{E}_{\boldsymbol{x}}[\|\boldsymbol{x}\|^2\,\mathbb{1}_{\text{sign}(\boldsymbol{\alpha}^\top\boldsymbol{x})\neq\text{sign}(\boldsymbol{\beta}^\top\boldsymbol{x})}] = O(\phi).$$

Lemma 35. *Consider $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d$ with $\angle(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \phi$, $\|\boldsymbol{\alpha}\|_2 = \|\boldsymbol{\beta}\|_2 = 1$ and $\boldsymbol{\alpha}^\top\boldsymbol{\beta} \geq 0$. We have*

$$\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{\alpha}^\top\boldsymbol{x}|\mathbb{1}_{\text{sign}(\boldsymbol{\alpha}^\top\boldsymbol{x})\neq\text{sign}(\boldsymbol{\beta}^\top\boldsymbol{x})}] = O(\phi^2).$$

Proof. It suffices to consider $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x} \in \mathbb{R}^2$. WLOG, assume $\boldsymbol{\alpha} = (1, 0)^\top$ and $\boldsymbol{\beta} = (\cos\phi, \sin\phi)^\top$
We have

$$\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{\alpha}^\top\boldsymbol{x}|\mathbb{1}_{\text{sign}(\boldsymbol{\alpha}^\top\boldsymbol{x})\neq\text{sign}(\boldsymbol{\beta}^\top\boldsymbol{x})}] = \frac{1}{2\pi}\int_0^\infty re^{-r^2/2}\,\mathrm{d}r\int_0^{2\pi}\cos\theta\mathbb{1}_{\text{sign}(\cos\theta)\neq\text{sign}(\cos(\theta-\phi))}\,\mathrm{d}\theta = O(\phi^2).$$

□

Lemma 36. *Under Lemma 6, let*

$$q_{ij} = \begin{cases}\frac{a_j a_i^*}{\sum_{j\in T_{i,+}(\delta_{close})} a_j^2} & \text{, if } j \in T_{i,+}(\delta_{close})\\ 0 & \text{, otherwise}\end{cases}$$

If $\sum_{i\in[m_*]}\left|a_i^2 - \|\boldsymbol{w}_i\|_2^2\right| \leq a_{\min}/2$, then $\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}^2 = O(\|\boldsymbol{a}_*\|_1)$.

Proof. We have

$$\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}^2 = \sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_{i,+}(\delta_{close})} \frac{a_j^2 a_i^{*2}}{(\sum_{j\in T_{i,+}(\delta_{close})} a_j^2)^2} = \sum_{i\in[m_*]} \frac{a_i^{*2}}{\sum_{j\in T_{i,+}(\delta_{close})} a_j^2}.$$

In the following, we aim to lower bound $\sum_{j\in T_{i,+}(\delta_{close})} a_j^2$. Given $\sum_{j\in T_{i,+}(\delta_{close})} |a_j^2 - \|\boldsymbol{w}_j\|_2^2| \le |a_i^*|/2$, we have

$$2\sum_{j\in T_{i,+}(\delta_{close})} a_j^2 \ge \sum_{j\in T_{i,+}(\delta_{close})} a_j^2 + \|\boldsymbol{w}_j\|_2^2 - |a_i^*|/2 \ge 2\sum_{j\in T_{i,+}(\delta_{close})} |a_j|\,\|\boldsymbol{w}_j\|_2 - |a_i^*|/2 \ge |a_i^*|/2,$$

where the last inequality is due to Lemma 20: $\sum_{j\in T_{i,+}(\delta_{close})} |a_j|\,\|\boldsymbol{w}_j\|_2 \ge |\sum_{j\in\mathcal{T}_i(\delta_{close})} a_j\,\|\boldsymbol{w}_j\|_2\,| \ge |a_i^*|/2$. Thus, we have $\sum_{i\in[m_*]}\sum_{j\in\mathcal{T}_i} q_{ij}^2 = O(\|\boldsymbol{a}_*\|_1)$. □

H Proofs in Section F (non-degenerate dual certificate)

In this section, we give the omitted proofs in Section F. The proofs are mostly direct computations with the properties of Hermite polynomials in Claim 2.

Lemma 30 (Non-degeneracy of kernel K)**.** *For $\ell \ge \Theta(\Delta^{-2}\log(m_*\ell/h\Delta))$, kernel $K_{\ge\ell}$ is non-degenerate in the sense that there exists $r = \Theta(\ell^{-1/2}), \rho_1 = \Theta(1), \rho_2 = \Theta(\ell)$ such that following hold:*

(i) $K(\boldsymbol{w},\boldsymbol{u}) \le 1-\rho_1$ *for all* $\delta(\boldsymbol{w},\boldsymbol{u}) := \angle(\boldsymbol{w},\boldsymbol{u}) \ge r$.

(ii) $K^{(20)}(\boldsymbol{w},\boldsymbol{u})[\boldsymbol{z},\boldsymbol{z}] \le -\rho_2\,\|\boldsymbol{z}\|^2$ *for tangent vector* $\boldsymbol{z}$ *that* $\boldsymbol{z}^\top\boldsymbol{w} = 0$ *and* $\delta(\boldsymbol{w},\boldsymbol{u}) \le r$.

(iii) $\left\|K^{(ij)}(\boldsymbol{w}_1^*,\boldsymbol{w}_k^*)\right\|_{\boldsymbol{w}_i^*,\boldsymbol{w}_k^*} \le h/m_*^2$ *for* $(i,j)\in\{0,1\}\times\{0,1,2\}$

Proof. With the property of Hermite polynomials in Claim 2, we have

$$
\begin{aligned}
K(\boldsymbol{w}, \boldsymbol{u}) =& \mathbb{E}_{\boldsymbol{x}}[\overline{\sigma_{\geq\ell}}(\overline{\boldsymbol{w}}^\top \boldsymbol{x})\overline{\sigma_{\geq\ell}}(\overline{\boldsymbol{u}}^\top \boldsymbol{x})] = \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 \cos^k\theta, \\
K^{(10)}(\boldsymbol{w}, \boldsymbol{u}) =& \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\frac{1}{\|\boldsymbol{w}\|_2}(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\overline{\boldsymbol{u}}, \\
K^{(11)}(\boldsymbol{w}, \boldsymbol{u}) =& \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\frac{1}{\|\boldsymbol{w}\|_2\|\boldsymbol{u}\|_2}(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\overline{\boldsymbol{u}}\overline{\boldsymbol{w}}^\top(\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top) \\
&+ \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\frac{1}{\|\boldsymbol{w}\|_2\|\boldsymbol{u}\|_2}(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)(\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top) \\
K^{(20)}(\boldsymbol{w}, \boldsymbol{u}) =& \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\frac{1}{\|\boldsymbol{w}\|_2^2}(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top) \\
&- \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\frac{1}{\|\boldsymbol{w}\|_2^2}\left((\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\overline{\boldsymbol{u}}\overline{\boldsymbol{w}}^\top + \overline{\boldsymbol{w}}\overline{\boldsymbol{u}}^\top(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top) + \overline{\boldsymbol{w}}^\top\overline{\boldsymbol{u}}(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\right), \\
K^{(21)}(\boldsymbol{w}, \boldsymbol{u})_i =& \partial_{u_i}K^{(20)}(\boldsymbol{w}, \boldsymbol{u}) \\
=& \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)(k-2)\cos^{k-3}\theta\frac{1}{\|\boldsymbol{w}\|_2^2\|\boldsymbol{u}\|_2}\boldsymbol{e}_i^\top(\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\overline{\boldsymbol{w}}\cdot(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top) \\
&+ \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\frac{1}{\|\boldsymbol{w}\|_2^2\|\boldsymbol{u}\|_2}(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\left((\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\boldsymbol{e}_i\overline{\boldsymbol{u}}^\top + \overline{\boldsymbol{u}}\boldsymbol{e}_i^\top(\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\right)(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top) \\
&- \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\frac{1}{\|\boldsymbol{w}\|_2^2\|\boldsymbol{u}\|_2}\boldsymbol{e}_i^\top(\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\overline{\boldsymbol{w}}\cdot\big((\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\overline{\boldsymbol{u}}\overline{\boldsymbol{w}}^\top \\
&+ \overline{\boldsymbol{w}}\overline{\boldsymbol{u}}^\top(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top) + \overline{\boldsymbol{w}}^\top\overline{\boldsymbol{u}}(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\big) \\
&- \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\frac{1}{\|\boldsymbol{w}\|_2^2}\big((\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)(\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\boldsymbol{e}_i\overline{\boldsymbol{w}}^\top + \overline{\boldsymbol{w}}\boldsymbol{e}_i^\top(\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top) \\
&+ \overline{\boldsymbol{w}}^\top(\boldsymbol{I} - \overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\boldsymbol{e}_i(\boldsymbol{I} - \overline{\boldsymbol{w}}\overline{\boldsymbol{w}}^\top)\big),
\end{aligned} \tag{11}
$$

where $\theta = \arccos(\overline{\boldsymbol{w}}^\top\overline{\boldsymbol{u}})$.

Part (i) Given that $r = \Theta(1/\sqrt{\ell})$ with a small enough hidden constant, we know for $\delta(\boldsymbol{w}, \boldsymbol{u}) \geq r$

$$
K(\boldsymbol{w}, \boldsymbol{u}) = \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2\cos^k\theta \leq \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2\cdot(1 - r^2/5)^\ell = c < 1,
$$

where c is a constant less than 1. Thus, $\rho_1 = \Theta(1)$.

Part (ii) For tangent vector $\boldsymbol{z}$ that $\boldsymbol{z}^\top\boldsymbol{w} = 0$, we have ($\|\boldsymbol{w}\|_2 = \|\boldsymbol{u}\|_2 = 1$, $\delta(\boldsymbol{w}, \boldsymbol{u}) \leq r$)

$$
\begin{aligned}
K^{(20)}(\boldsymbol{w}, \boldsymbol{u})[\boldsymbol{z}, \boldsymbol{z}] =& \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\cdot(\overline{\boldsymbol{u}}^\top\boldsymbol{z})^2 - \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\cdot\overline{\boldsymbol{w}}^\top\overline{\boldsymbol{u}}\,\|\boldsymbol{z}\|_2^2 \\
=& \frac{\|\boldsymbol{z}\|_2^2}{Z_\sigma^2}\left(\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\cdot(\overline{\boldsymbol{u}}^\top\overline{\boldsymbol{z}})^2 - \sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\cdot\overline{\boldsymbol{w}}^\top\overline{\boldsymbol{u}}\right) \\
\leq& \frac{\|\boldsymbol{z}\|_2^2}{Z_\sigma^2}\left(\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\sin^2\theta - \sum_{\ell\leq k\leq 2\ell}\hat{\sigma}_k^2 k\cos^k\theta\right).
\end{aligned}
$$

For the first term, we have

$$\begin{aligned}
&\sum_{k\geq \ell} \hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta \sin^2\theta \\
&\leq \sum_{k\geq 1/r^2} \hat{\sigma}_k^2 k(k-1)\cdot \Theta(1/k) + \sum_{\ell\leq k\leq 1/r^2} \hat{\sigma}_k^2 k(k-1) r^2 \\
&= \sum_{k\geq 1/r^2} \Theta(k^{-3/2}) + \sum_{\ell\leq k\leq 1/r^2} \Theta(k^{-1/2}) r^2 = \Theta(r),
\end{aligned}$$

where we use $\hat{\sigma}_k^2 = \Theta(k^{-5/2})$ in Lemma 37.

For the second term, we have

$$\sum_{\ell\leq k\leq 2\ell} \hat{\sigma}_k^2 k \cos^k\theta \geq \Theta(\ell^{-1/2})(1-r^2)^{2\ell}.$$

Given that $r = \Theta(1/\sqrt{\ell})$ with a small enough hidden constant, we know

$$K^{(20)}(\boldsymbol{w},\boldsymbol{u})[\boldsymbol{z},\boldsymbol{z}] \leq -\frac{\|\boldsymbol{z}\|_2^2}{Z_\sigma^2}\Theta(\ell^{-1/2}) = -\Theta(\ell)\,\|\boldsymbol{z}\|_2^2\,,$$

since $Z_\sigma^2 = \Theta(\ell^{-3/2})$.

Part (iii) Recall that $\delta(\boldsymbol{w}_i^*, \boldsymbol{w}_j^*) \geq \Delta$ for $i \neq j$. It suffices to bound $\left\|K^{(ij)}(\boldsymbol{w},\boldsymbol{u})\right\|_2 \leq h/m_*^2$ for $\theta = \delta(\boldsymbol{w},\boldsymbol{u}) \geq \Delta$. Given that $\ell \geq \Theta(\Delta^{-2}\log(m_*\ell/h\Delta))$ with large enough hidden constant, from

(11) we have for $\|\boldsymbol{w}\| = \|\boldsymbol{u}\| = 1$

$$
\begin{aligned}
K(\boldsymbol{w},\boldsymbol{u}) \leq & \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2(1-\Delta^2/5)^\ell \leq h/m_*^2,\\
\left\|K^{(10)}(\boldsymbol{w},\boldsymbol{u})\right\|_{\boldsymbol{w}} \leq & \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\sin\theta \leq \Theta(\ell)(1-\Delta^2/5)^{\ell-1} \leq h/m_*^2,\\
\left\|K^{(11)}(\boldsymbol{w},\boldsymbol{u})\right\|_{\boldsymbol{w},\boldsymbol{u}} = & \frac{1}{Z_\sigma^2}\sup_{\substack{\boldsymbol{z}_1^\top\boldsymbol{w}=\boldsymbol{z}_2^\top\boldsymbol{u}=0,\\ \|\boldsymbol{z}_1\|_2=\|\boldsymbol{z}_2\|_2=1}}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\overline{\boldsymbol{u}}^\top\boldsymbol{z}_1\cdot\overline{\boldsymbol{w}}^\top\boldsymbol{z}_2 + \sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\boldsymbol{z}_1^\top\boldsymbol{z}_2\\
\leq & \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\sin^2\theta + \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\\
\leq & \Theta(\ell^{3/2})\sum_{k\geq\ell}\Theta(k^{-1/2})(1-\Delta^2/5)^{k-2} + \Theta(\ell)(1-\Delta^2/5)^{\ell-1} \leq h/m_*^2,\\
\left\|K^{(20)}(\boldsymbol{w},\boldsymbol{u})\right\|_{\boldsymbol{w}} = & \frac{1}{Z_\sigma^2}\sup_{\substack{\boldsymbol{z}_1^\top\boldsymbol{w}=\boldsymbol{z}_2^\top\boldsymbol{w}=0,\\ \|\boldsymbol{z}_1\|_2=\|\boldsymbol{z}_2\|_2=1}}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\cdot\overline{\boldsymbol{u}}^\top\boldsymbol{z}_1\cdot\overline{\boldsymbol{u}}^\top\boldsymbol{z}_2 - \sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\cdot\overline{\boldsymbol{w}}^\top\overline{\boldsymbol{u}}\cdot\boldsymbol{z}_1^\top\boldsymbol{z}_2\\
\leq & \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\sin^2\theta + \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\\
\leq & \Theta(\ell^{3/2})\sum_{k\geq\ell}\Theta(k^{-1/2})(1-\Delta^2/5)^{k-2} + \Theta(\ell)(1-\Delta^2/5)^{\ell-1} \leq h/m_*^2,\\
\left\|K^{(21)}(\boldsymbol{w},\boldsymbol{u})\right\|_{\boldsymbol{w},\boldsymbol{u}} = & \sup_{\substack{\boldsymbol{z}_1^\top\boldsymbol{w}=\boldsymbol{z}_2^\top\boldsymbol{w}=\boldsymbol{q}^\top\boldsymbol{u}=0,\\ \|\boldsymbol{z}_1\|_2=\|\boldsymbol{z}_2\|_2=\|\boldsymbol{q}\|_2=1}}\frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)(k-2)\cos^{k-3}\theta\sum_i q_i\boldsymbol{e}_i^\top(\boldsymbol{I}-\overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\overline{\boldsymbol{w}}\cdot\overline{\boldsymbol{u}}^\top\boldsymbol{z}_1\cdot\overline{\boldsymbol{u}}^\top\boldsymbol{z}_2\\
& + \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\left(\sum_i q_i\boldsymbol{z}_1^\top(\boldsymbol{I}-\overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\boldsymbol{e}_i\cdot\overline{\boldsymbol{u}}^\top\boldsymbol{z}_2 + \sum_i q_i\boldsymbol{z}_2^\top(\boldsymbol{I}-\overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\boldsymbol{e}_i\cdot\overline{\boldsymbol{u}}^\top\boldsymbol{z}_1\right)\\
& - \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\sum_i q_i\boldsymbol{e}_i^\top(\boldsymbol{I}-\overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\overline{\boldsymbol{w}}\cdot\overline{\boldsymbol{w}}^\top\overline{\boldsymbol{u}}\cdot\boldsymbol{z}_1^\top\boldsymbol{z}_2\\
& - \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\sum_i q_i\overline{\boldsymbol{w}}^\top(\boldsymbol{I}-\overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^\top)\boldsymbol{e}_i\cdot\boldsymbol{z}_1^\top\boldsymbol{z}_2\\
\leq & \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)(k-2)\cos^{k-3}\theta\sin^3\theta + \frac{2}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)\cos^{k-2}\theta\sin\theta + \frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k\cos^{k-1}\theta\\
\overset{(a)}{\leq} & h/m_*^2,
\end{aligned}
$$

where we use $\hat{\sigma}_k^2 = \Theta(k^{-5/2})$ in Lemma 37 and (a) the last two terms bound similarly as in $K^{(20)}$ and first term $\frac{1}{Z_\sigma^2}\sum_{k\geq\ell}\hat{\sigma}_k^2 k(k-1)(k-2)\cos^{k-3}\theta\sin^3\theta \leq \Theta(\ell^{3/2})\sum_{k\geq\ell}\Theta(k^{1/2})(1-\Delta^2/5)^k \leq h/3m_*^2$. □

Lemma 31 (Regularity conditions on kernel K)**.** *Let $B_{ij} := \sup_{\boldsymbol{w},\boldsymbol{u}}\left\|K^{(ij)}(\boldsymbol{w},\boldsymbol{u})\right\|_{\boldsymbol{w},\boldsymbol{u}}$ and $B_0 = B_{00}+B_{10}+1$, $B_2 = B_{20}+B_{21}+1$. We have $B_{00} = O(1)$, $B_{10} = O(1)$, $B_{11} = O(\ell)$, $B_{20} = O(\ell)$, $B_{21} = O(\ell^{3/2})$, and therefore $B_0 = O(1)$, $B_2 = O(\ell^{3/2})$.*

Proof. We compute B_{ij} one by one from (11) (see part (iii) proof in Lemma 30).

$$
\begin{aligned}
B_{00} =& \sup_{\boldsymbol{w},\boldsymbol{u}} \left| \frac{1}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 \cos^k\theta \right| \le 1, \\
B_{10} \le& \frac{1}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 k \cos^{k-1}\theta \sin\theta = O(1), \\
B_{11} \le& \frac{1}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 k(k-1) \cos^{k-2}\theta \sin^2\theta + \frac{1}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 k \cos^{k-1}\theta = O(\ell), \\
B_{20} \le& \frac{1}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 k(k-1) \cos^{k-2}\theta \sin^2\theta + \frac{1}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 k \cos^{k}\theta = O(\ell), \\
B_{21} \le& \frac{1}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 k(k-1)(k-2) \cos^{k-3}\theta \sin^3\theta + \frac{2}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 k(k-1) \cos^{k-1}\theta \sin\theta + \frac{1}{Z_\sigma^2} \sum_{k\ge\ell} \hat{\sigma}_k^2 k \cos^{k-1}\theta \sin\theta \\
\overset{(a)}{\le}& O(\ell^{3/2}),
\end{aligned}
$$

where (a) the last two terms follow the same as in B_{11} and the first term $\frac{1}{Z_\sigma^2}\sum_{k\ge\ell}\hat{\sigma}_k^2 k(k-1)(k-2)\cos^{k-3}\theta\sin^3\theta \le \Theta(\ell^{3/2})\sum_{k\ge\ell}\Theta(k^{1/2})(1-\theta^2/5)^{k-3}\theta^3 \le \Theta(\ell^{3/2})\sum_{k\ge\ell}(1-\theta^2/5)^{(k-3)/2}\theta^3/\sqrt{\log(1/(1-\theta^2))} = \Theta(\ell^{3/2})$. □

I Some Properties of Hermite Polynomials

In this section, we give several properties of Hermite Polynomials that are useful in our analysis. See [25] for a more complete discussion on Hermite polynomials. Let H_k be the probabilists' Hermite polynomial where

$$H_k(x) = (-1)^k e^{x^2/2} \frac{\mathrm{d}^k}{\mathrm{d}x^k}(e^{-x^2/2})$$

and $h_k = \frac{1}{\sqrt{k!}}H_k$ be the normalized Hermite polynomials. Given a function σ, we call $\sigma(x) = \sum_{k=0}^{\infty}\hat{\sigma}_k h_k(x)$ as the Hermit expansion of σ and $\hat{\sigma}_k = \mathbb{E}_{x\sim N(0,1)}[\sigma(x)h_k(x)]$ as the k-th Hermite coefficient of σ.

The following is a useful property of Hermite polynomial.

Claim 2 ([25], Section 11.2)**.** *Let (x,y) be ρ-correlated standard normal variables (that is, both x,y have marginal distribution $N(0,1)$ and $\mathbb{E}[xy]=\rho$). Then, $\mathbb{E}[h_m(x)h_n(y)] = \rho^n\delta_{mn}$.*

The following lemma gives the Hermite coefficients for absolute value function and ReLU.

Lemma 37. *Let $\hat{\sigma}_k = \mathbb{E}_{x\sim N(0,1)}[\sigma(x)h_k(x)]$ be the Hermite coefficient of σ. For σ is ReLU or absolute function, we have $|\hat{\sigma}_k| = \Theta(k^{-5/4})$.*

Proof. From Goel et al. [18], Zhou et al. [28] we have

$$
\hat{\sigma}_{abs,k} = \begin{cases} 0 & ,k \text{ is odd} \\ \sqrt{2/\pi} & ,k=0 \\ (-1)^{\frac{k}{2}-1}\sqrt{\frac{2}{\pi}}\frac{(k-2)!}{\sqrt{k!}2^{k/2-1}(k/2-1)!} & ,k \text{ is even and } k\ge 2 \end{cases}
$$

$$
\hat{\sigma}_{relu,k} = \begin{cases} 0 & ,k \text{ is odd and } k\ge 3 \\ \sqrt{1/2\pi} & ,k=0 \\ 1/2 & ,k=1 \\ (-1)^{\frac{k}{2}-1}\sqrt{\frac{1}{2\pi}}\frac{(k-2)!}{\sqrt{k!}2^{k/2-1}(k/2-1)!} & ,k \text{ is even and } k\ge 2 \end{cases}
$$

Using Stirling's formula, we get $|\hat{\sigma}_{abs,k}|, |\hat{\sigma}_{relu,k}| = \Theta(k^{-5/4})$. □