On Learnability via Gradient Method for Two-Layer ReLU Neural Networks in Teacher-Student Setting

Shunta Akiyama [1] **Taiji Suzuki** [1 2]

Abstract

Deep learning empirically achieves high performance in many applications, but its training dynamics has not been fully understood theoretically. In this paper, we explore theoretical analysis on training two-layer ReLU neural networks in a teacher-student regression model, in which a student network learns an unknown teacher network through its outputs. We show that with a specific regularization and sufficient over-parameterization, the student network can identify the parameters of the teacher network with high probability via gradient descent with a norm dependent stepsize even though the objective function is highly non-convex. The key theoretical tool is the measure representation of the neural networks and a novel application of a dual certificate argument for sparse estimation on a measure space. We analyze the global minima and global convergence property in the measure space.

1. Introduction

Deep learning empirically achieves high performance in many applications, such as computer vision and speech recognition. To explain its success from the theoretical view point, we need to reveal its optimization dynamics and the generalization ability of the solution that is obtained by a particular optimization method such as gradient descent. However, its training dynamics has not been fully understood theoretically and thus the generalization ability of the solution is still an open question. One of the difficulties of this problem is non-convexity of the associated optimization problem (Li et al., 2018) for the optimization aspect, and the high dimensionality induced by over-parameterization for the generalization aspect. In this study, we tackle these two problems in a teacher-student problem with the ReLU activation under an over-parameterized setting. In this setting, we need to take care of the non-differentiability of the ReLU activation and the over-specification problem due to the over-parameterization which potentially causes difficulty to show favorable generalization ability such as exact recovery.

The teacher-student setting is one of the most common settings for theoretical studies, e.g., Tian (2017); Yehudai & Shamir (2020); Goldt et al. (2019); Safran & Shamir (2018); Safran et al. (2020); Tian (2020); Suzuki & Akiyama (2021); Zhang et al. (2019); Zhou et al. (2021) to name a few. Zhong et al. (2017) studied the case where the teacher and student have the same width, showed that the strong convexity holds around the parameters of the teacher network and proposed a special tensor method for initialization to achieve the global convergence to the global optimal. However, its global convergence is guaranteed only for a special initialization which excludes a pure gradient descent method. Moreover, the over-parameterized setting is not included in their analysis. Safran & Shamir (2018) empirically showed that gradient descent is likely to converge to non-global optimal local minima, even if we prepare a student that has the same size as the teacher. More recently, Yehudai & Shamir (2020) showed that even in the simplest case where the teacher and student have the width *one*, there exists distributions and activations in which gradient descent fails to learn. Safran et al. (2020) showed the strong convexity around the parameters of the teacher network in the case where the teacher and student have the same width for Gaussian inputs. They also studied the effect of over-parameterization and showed that over-parameterization will change the spurious local minima into the saddle points. However, it should be noted that this does not imply that a gradient descent can reach the global optima.

To alleviate the non-convexity of neural network optimization, over-parameterization is one of the promising approaches. Indeed, it is fully exploited by (i) Neural Tangent Kernel (NTK) (Allen-Zhu et al., 2019; Arora et al., 2019; Jacot et al., 2018; Du et al., 2019; Weinan et al., 2020) and (ii) mean field analysis (Nitanda & Suzuki, 2017; Chizat &

[1]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan [2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. Correspondence to: Shunta Akiyama <shunta_akiyama@mist.i-tokyo.ac.jp>, Taiji Suzuki <taiji@mist.i.u-tokyo.ac.jp>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Bach, 2018; Chizat, 2019; Suzuki & Akiyama, 2021; Mei et al., 2019; Tzen & Raginsky, 2020). (i) In the setting of NTK, the gradient descent of neural networks can be seen as the convex optimization in RKHS, and thus it is easier to analyze. On the other hand, in this regime, it is hard to explain the superiority of deep learning, because the estimation ability of the obtained estimator is reduced to that of the corresponding kernel. (ii) In the setting of the mean field analysis, a kind of continuous limit of neural network is considered and its convergence to some specific target functions has been analyzed. This regime is more suitable in terms of a "beyond kernel" perspective, but it essentially deals with a continuous limit and hence is difficult to show convergence to a teacher network with a *finite width*.

In this paper, we make full use of the "measure representation" of two-layer ReLU networks as in the mean field analysis, while our approach employs a *sparse* regularization on the measure of parameters to show the convergence of a gradient descent method to the global optimum where the teacher network has a finite width. The sparse regularization on a measure space is well studied in a so-called *BLASSO* problem (De Castro & Gamboa, 2012). Indeed, Chizat (2019) analyzed the gradient descent for two layer neural networks from the view point of BLASSO analyses, and showed the convergence to the global optimal. However they assumed several assumptions which are hard to clarify, and excluded a non-smooth activation such as the ReLU activation. On the other hand, we explicitly present a realistic condition under which a gradient descent converges to the global optimum. More specifically, our contributions can be summarized as follows:

- We show that with an appropriate sparse regularization, the optimal solution of a regularized empirical risk can be arbitrarily close to the true teacher-parameters for a sufficiently small regularization parameter. This implies effectiveness of a sparsity inducing regularization in deep learning.
- We prove that a gradient descent with a norm-dependent step size can converge to the global optimum of the regularized learning problem if the student network is appropriately over-parameterized.
- Combining the above results, we show that a gradient descent method with an over-parameterized initialization can find a network which is arbitrary close to the true teacher network. In particular, the size of the estimated network becomes "narrow" even though the initial solution is over-parameterized, which explains the feature learning ability of neural networks leading a better performance than shallow methods such as kernel methods.

1.1. Other Related Works

BLASSO problem The BLASSO problem (De Castro & Gamboa, 2012) is a regression problem with total variation regularization on a measure space, which is an extension of the LASSO problem to the measure space. One of the main theoretical interests of BLASSO studies (Bredies & Pikkarainen, 2013; Candès & Fernandez-Granda, 2013; Duval & Peyré, 2015; Poon et al., 2018; 2019) is to clarify whether the global minima of BLASSO can recover the "true" measure in the setting where the true measure is sparse, i.e., given by a sum of Dirac measures. Duval & Peyré (2015) showed that for a sufficiently small sample noise and an appropriate regularization, the global minimum will also be sparse and close to the true measure. A key theoretical tool is a dual certificate, which is motivated by the Fenchel duality. However, their analysis assumes smoothness on the objective function and thus is not directly applied to our setting because of the non-differentiability of the ReLU activation.

Sparse regularization It has been shown that explicit or implicit sparse regularization such as L_1-regularization is beneficial to obtain better performances of deep learning under certain situations (Chizat & Bach, 2020; Gunasekar et al., 2018; Woodworth et al., 2020; Klusowski & Barron, 2016). However, it is still an open question that a gradient descent can find the teacher model in a regression setting with the ReLU non-linear activation. Bach (2017) analyzed a neural network model with a sparse regularization (L_1-regularization) which can be regarded as an extension of Barron class (Barron, 1993), and derived its model capacity. It was shown that the Frank-Wolfe type method can estimate a target function in the neural network model, but unfortunately this does not imply that a gradient descent method can estimate the target function. Moreover, it is not clear that each update of the Frank-Wolfe method is computationally tractable.

Langevin dynamics approach The gradient Langevin dynamics (GLD) is a useful approach to obtain a global optimum of a non-convex objective function (Welling & Teh, 2011; Raginsky et al., 2017; Erdogdu et al., 2018; Suzuki & Akiyama, 2021). This approach can be also applied to neural network optimization but such analysis would not give any information about the landscape of the neural network training. Among them, Suzuki & Akiyama (2021) considered an infinite dimensional Langevin dynamics, but they excluded a non-differentiable activation such as ReLU and did not give any landscape analysis.

1.2. Notations

Here we give some notations used in the paper. Let $\mathcal{M}(\mathcal{C})$ be the set of the Radon measures on a topological space

$\mathcal{C}$ (we consider the Borel algebra of $\mathcal{C}$ as the σ-field on which the Radon measures are defined). Let $\delta_w(\cdot)$ be the Dirac measure on $w \in \mathbb{R}^d$, i.e., $\int f(x)\delta_w(\mathrm{d}x) = f(w)$. Let $[m] := \{1, \dots, m\}$ for a positive integer m. Let the inner product between $x, y \in \mathbb{R}^d$ be $\langle x, y\rangle := \sum_{j=1}^d x_i y_i$.

2. Problem Settings

In this section, we give the problem setting and the model that we consider in this paper. We focus on a regression problem where we observe n training examples $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ generated by the following model:

$$y_i = f^\circ(x_i), \tag{1}$$

where $f^\circ : \mathbb{R}^d \to \mathbb{R}$ is the unknown true function that we want to estimate, $(x_i)_{i=1}^n$ are independently identically distributed from $P_\mathcal{X}$. Later on we assume that $P_\mathcal{X}$ is the uniform distribution on the unit ball $\mathbb{S}^{d-1}$ (Assumption 3.1).

Based on the observed data D_n, we construct an estimator $\hat{f}$ which is supposed to be "close" to the true function f°. As its performance measure, we employ the mean squared error defined by $\|\widehat{f} - f^\circ\|_{L_2(P_\mathcal{X})}^2 := \mathbb{E}_{X\sim P_\mathcal{X}}[(\widehat{f}(X) - f^\circ(X))^2]$. Its empirical version is defined by $\|\widehat{f} - f^\circ\|_n^2 := \frac{1}{n}\sum_{i=1}^n (\widehat{f}(x_i) - f^\circ(x_i))^2$.

Teacher-Student Model In this section, we prepare the teacher-student model that we consider in this paper. The student model is the two-layer neural network with the ReLU-activation $\sigma(u) = \max\{x, 0\}$ (Glorot et al., 2011) and width M, which is defined as

$$f(x; \Theta) = \sum_{j=1}^M a_j \sigma(\langle w_j, x\rangle), \tag{2}$$

where $\Theta = ((a_1, w_1), \dots, (a_M, w_M)) \in (\mathbb{R} \times \mathbb{R}^d)^M$ is the trainable parameter. The teacher model is assumed to be included in the student model but the width could be smaller than M:

$$f^\circ(x) = \sum_{j=1}^m a_j^\circ \sigma(\langle w_j^\circ, x\rangle), \tag{3}$$

where m is the width of the teacher model and $(a_j^\circ, w_j^\circ) \in \mathbb{R} \times \mathbb{R}^d$ $(j \in [m])$. We consider an over-parameterized setting where $m \le M$ is assumed to be satisfied. Hence, the teacher model can be regarded as an element of the student model by setting $a_j = 0$ for $j = m+1, \dots, M$. For notational simplicity, we denote by $\Theta^\circ := (a_j^\circ, w_j^\circ)_{j=1}^m \in (\mathbb{R} \times \mathbb{R}^d)^m$.

For a neural network model, it is generally difficult to write the close form of the (regularized) empirical risk minimizer. Therefore, we typically optimize Θ via the gradient descent technique, but due to the non-convexity of the objective function, it is far from trivial that the global minima can be obtained by gradient descent.

Sparse Regularized Empirical Risk To estimate the true parameter Θ°, we define the following regularized empirical risk minimization problem on the parameter space $(\mathbb{R} \times \mathbb{R}^d)^M$:

$$\min_{\Theta \in (\mathbb{R}\times\mathbb{R}^d)^M} \frac{1}{2n}\sum_{i=1}^n (y_i - f(x_i; \Theta))^2 + \lambda \sum_{j=1}^M |a_j|\|w_j\|, \tag{4}$$

where $\lambda \ge 0$ is a regularization parameter. The regularization term $\lambda\sum_{j=1}^M |a_j|\|w_j\|$ can be seen as an L_1-regularization which induces sparsity. Indeed, by the scale homogeneity of ReLU ($a_j\sigma(\langle w_j, x\rangle) = a_j\|w_j\|\sigma(\langle w_j/\|w_j\|, x\rangle)$), we may reset the parameter as $a'_j = a_j\|w_j\|$ and $w'_j = w_j/\|w_j\|$ and then the regularization term can be rewritten as $\lambda\sum_{j=1}^M |a'_j|$. Apparently, this is the L_1-norm of $(a'_j)_{j=1}^M$.

In practice, we typically use the L_2-regularization $\frac{\lambda}{2}\sum_{j=1}^M (a_j^2 + \|w_j\|^2)$ instead of the L_1-regularization as induced above. However, the arithmetic-geometric mean relation yields that

$$|a_j|\|w_j\| = \min_{\substack{(a'_j, w'_j) \in \mathbb{R}\times\mathbb{R}^d: \\ |a_j|\|w_j\| = |a'_j|\|w'_j\|}} \tfrac{1}{2}(|a'_j|^2 + \|w'_j\|^2). \tag{5}$$

Therefore, our sparse regularization can be replaced by the L_2-regularization. In this paper, we directly consider the sparse regularization instead just for simplicity.

Remark 2.1. *We will see that the regularization term $\lambda\sum_{j=1}^M |a_j|\|w_j\|$ corresponds to the total-variation norm regularization for the measure representation of the network which we refer to in the next section. The same type of regularization has been considered in several studies, e.g., E et al. (2019); Neyshabur et al. (2015). In those studies, it plays an important role to show a better performance of deep learning compared with kernel methods. We further make full use of the sparsity to show the exact recovery of the true parameter Θ° even under the over-parameterized setting.*

3. Global Minima in the Teacher-Student Setting

In this section, we show that the minimizer of the regularized empirical risk (4) is arbitrarily close to the teacher network f° for a sufficiently large sample size n. Note that we are not arguing here that the optimal solution can be obtained by the gradient descent, but the computational issue will be addressed in the next section. We make the following assumptions for our analysis.

Assumption 3.1. *$(x_i)_{i=1}^n$ are i.i.d. observations from the uniform distribution on $\mathbb{S}^{d-1}$, that is, $P_{\mathcal{X}} = \mathrm{Unif}(\mathbb{S}^{d-1})$.*

Assumption 3.2. *The teacher network $f^\circ = \sum_{j=1}^m a_j^\circ \sigma(\langle w_j^\circ, \cdot \rangle)$ satisfies the following conditions:*

1. *$a_j^\circ > 0 \quad (\forall j \in [m])$.*
2. *$\langle w_{j_1}^\circ, w_{j_2}^\circ \rangle = 0 \quad (\forall j_1, j_2 \in [m],\ j_1 \neq j_2)$.*

The second assumption could be a bit strong, but the same assumption has been considered in several previous researches (Zhong et al., 2017; Tian, 2017; Safran & Shamir, 2018; Safran et al., 2020; Li et al., 2020). For example, Safran et al. (2020) analyzed the landscape of the objective under this assumption and showed a negative result that the loss landscape around the global minima is not even *locally convex*. On the other hand, they also showed that an over-parameterization turns a non-global optimal point into a saddle-point. However, they have not shown that a gradient descent can reach the optimal solution. Li et al. (2020) showed a global optimality of gradient descent in a specific teacher student setting under this condition. They consider a specific teacher model $f^\circ(x) = \sum_{j=1}^M a_j^\circ |\langle x, \theta_j^\circ \rangle|$ for $a_j^\circ > 0$ and a student model $f(x; W) = \frac{1}{M}\sum_{j=1}^M \|w_j\| \sigma(\langle w_j, x \rangle)$. This is relevant to ours, but specification of the teacher network is quite different from our setting.

The main ingredient of our analysis is the *measure representation* of the two layer ReLU-neural network. Using this representation, one can regard the neural network training as a sparse regularized learning on the measure space. This enables us to show (near) exact recovery. In particular, the *Beurling-LASSO (BLASSO)* analysis (De Castro & Gamboa, 2012) which could be seen as an infinite dimensional extension of sparse regularization theory is helpful.

3.1. Mesure Representation of Two-Layer Neural Networks and BLASSO Problem

We introduce the measure representation of the two-layer ReLU neural network. By using 1-homogeneity of the ReLU activation, it holds that

$$\begin{aligned}\sum_{j=1}^M a_j \sigma(\langle w_j, x \rangle) &= \sum_{j=1}^M a_j \|w_j\| \sigma\left(\left\langle \frac{w_j}{\|w_j\|}, x \right\rangle\right) \\ &= \int_{\mathbb{S}^{d-1}} \sigma(\langle \theta, x \rangle) \mathrm{d}\nu(\theta)\end{aligned} \tag{6}$$

with $\nu = \sum_{j=1}^m a_j \|w_j\| \delta_{w_j/\|w_j\|} \in \mathcal{M}(\mathbb{S}^{d-1})$. We call this ν a measure representation of the two-layer ReLU neural network. In the following, we write

$$f(x; \nu) = \int_{\mathbb{S}^{d-1}} \sigma(\langle \theta, x \rangle) \mathrm{d}\nu(\theta). \tag{7}$$

Under this representation, the teacher network is represented as $\nu^\circ = \sum_{j=1}^m r_j^\circ \delta_{\theta_j^\circ}$ with $r_j = a_j^\circ \|w_j^\circ\|$ and $\theta_j^\circ = w_j^\circ / \|w_j^\circ\|$.

Remark 3.3. *For a more general activation σ, we need to consider a measure on the product space $\mathbb{R} \times \mathbb{R}^d$. However, thanks to the 1-homogeneity of ReLU, we only need to consider a measure on $\mathbb{S}^{d-1}$ which is a compact metric space.*

With this measure representation, we may consider the following regression problem on the measure space instead:

$$\min_{\nu \in \mathcal{M}(\mathbb{S}^{d-1})} \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i; \nu))^2 + \lambda \|\nu\|_{\mathrm{TV}}, \tag{8}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the *total variation* norm of $\nu \in \mathcal{M}(\mathbb{S}^{d-1})$ that is defined by $\|\nu\|_{\mathrm{TV}} = \nu_+(\mathbb{S}^{d-1}) + \nu_-(\mathbb{S}^{d-1})$ for the Hahn–Jordan decomposition $\nu(\cdot) = \nu_+(\cdot) - \nu_-(\cdot)$. This can be seen as the continuous version of the original problem (4), which is called a BLASSO problem (De Castro & Gamboa, 2012). Since the measure representation covers any finite-width neural network, the following proposition holds.

Proposition 3.4. *Assume that a global minimum of (8) is obtained by a measure which is represented as a finite sum of Dirac measures:*

$$\nu^* = \textstyle\sum_{j=1}^{m^*} r_j^* \delta_{\theta_j^*},$$

then for the student network satisfying $M \geq m^*$, the global minima of (4) can be obtained by the form whose measure representation is written by ν^*.

There have been several studies that focused on the global minimum of the BLASSO problem (8). Duval & Peyré (2015) analyzed this problem in the context of sparse spike deconvolution, in which f is a Gaussian convolution filter and is an element of $L_2(\mathbb{T})$ (where $\mathbb{T}$ denotes the 1-dimensional torus), and showed that under the so-called *NDSC condition*, the global minima can be close to underlying measure. Poon et al. (2018; 2019) analyzed a more general setting and derived a sufficient condition for the NDSC condition. However, these analyses have required smoothness on the objective. Therefore, they can not be applied directly to our setting because of non-differentiability of the ReLU activation. We overcome this difficulty by directly deriving the *dual certificate* of the optimization problem.

3.2. Main Result 1: Global Minima of Regularized Empirical Risk

We prove that with a sufficiently small regularization parameter, the global minimizer of (8) is close to the teacher network with an arbitrarily small gap. We state this as the following theorem.

Theorem 3.5. *Assume that Assumptions 3.1 and 3.2 are satisfied. Suppose that $n > \mathrm{poly}(m, d, \log 1/\delta)$ for $\delta > 0$. Then, with probability at least $1 - \delta$, we have that, for any $\epsilon > 0$, with sufficiently small $\lambda > 0$, the optimal solution of (8) is uniquely determined and written by the form $\nu^* = \sum_{j=1}^m r_j^* \delta_{\theta_j^*}$ where $(r_j^*, \theta_j^*)_{j=1}^m \subset \mathbb{R} \times \mathbb{S}^{d-1}$ satisfy*

$$\begin{cases} \sum_{j=1}^m |r_j^\circ - r_j^*|^2 \leq \mathrm{O}(m\lambda^2) \\ \sum_{j=1}^m \mathrm{dist}^2(\theta_j^*, \theta_j^\circ) \leq \mathrm{O}(m\lambda^2) \end{cases} . \tag{9}$$

The proof can be found in Appendix A. From this theorem and Proposition 3.4, we immediately obtain the following corollary.

Corollary 3.6. *Under the same assumption with Theorem 3.5, for the student network model with more than m nodes, the optimal solution of (4) achieves the same property with Theorem 3.5, i.e., the measure representation of the optimal network satisfies (9).*

Therefore, as long as the network size M is sufficiently large such that $M \geq m$, we can recover the true network with arbitrarily small error by tuning the regularization parameter. The event of this property is uniform over the choice of the accuracy ϵ and corresponding regularization parameter λ. Hence, by decreasing λ gradually, we can finally recover the teacher model exactly. This result only characterizes the globally optimal solution and it does not say anything about the algorithmic convergence of a gradient descent method. In the next section, we address this issue.

Proof Strategy: Dual Certificate Theorem 3.5 can be shown through a dual certificate characterization of the optimal solution. Let the optimization problem (8) be P_λ. By the Fenchel's duality theorem (Rockafeller, 1967; Borwein & Zhu, 2005; Duval & Peyré, 2015), its dual problem D_λ is given by

$$(D_\lambda) \qquad \max_{p \in \mathbb{R}^n : \|f^*(p)\|_\infty \leq 1} \frac{1}{n^2} \sum_{i=1}^n y_i p_i - \frac{\lambda}{2n^2} \|p\|^2,$$

where $f^*(p)(\cdot) \in \mathcal{C}(\mathbb{S}^{d-1})$[1] that is defined by $f^*(p)(\theta) := \frac{1}{n} \sum_{i=1}^n \sigma(\langle \theta, x_i \rangle)$, and the strong duality holds, that is, ν_λ^* is the optimal solution of P_λ if the following optimality condition is satisfied for the *unique* solution p_λ of D_λ (the uniqueness of the dual solution follows from the strong convexity of the dual problem):

$$\begin{cases} f^*(p_\lambda) \in \partial \|\nu_\lambda^*\|_{\mathrm{TV}}, \\ p_{\lambda,i} = -\frac{1}{\lambda}(f(x_i; \nu_\lambda^*) - y_i) \quad (\forall i \in [n]). \end{cases}$$

We call $f^*(p_\lambda)$ a *dual certificate* for ν_λ^*. Conversely, if this condition is satisfied by $(\nu_\lambda^*, p_\lambda) \in \mathcal{M}(\mathbb{S}^{d-1}) \times \mathbb{R}^n$, then the pair is the optimal solution of both P_λ and D_λ. Therefore, our strategy is to show that the dual certificate $f^*(p_\lambda)$ admits only a primal optimal solution ν_λ^* that satisfies the condition in the theorem, i.e., the support of ν_λ^* consists of only m distinct points each of which is close to the true parameters $(\theta_j^\circ)_{j=1}^m$. To prove this, we show that there exist $(\theta_j^*)_{j=1}^m$ such that $(\mathrm{dist}(\theta_j^*, \theta_j^\circ))_{j=1}^m$ are sufficiently small and satisfy

$$\begin{cases} f^*(p_\lambda)(\theta_j^*) = 1 \ \ (\forall j \in [m]), \\ |f^*(p_\lambda)(\theta)| < 1 \ \ (\forall \theta \in \mathbb{S}^{d-1} / \{\theta_1^*, \ldots, \theta_m^*\}) \end{cases} \tag{10}$$

for sufficiently small λ. From this inequality, we can show that $(|r_j^* - r_j^\circ|)_{j=1}^m$ will also be sufficiently small. Finally by using the form $\nu^* = \sum_{j=1}^m r_j^* \delta_{\theta_j^*}$ and strong convexity of the empirical risk term in P_λ w.r.t. r_j^* and θ_j^* around the teacher parameters $(r_j^\circ, \theta_j^\circ)_{j=1}^m$, we get the quantitative bound as Eq. (9).

For that purpose, we particularly consider a setting where $\lambda = 0$, and consider the minimal norm certificate:

$$p_0 := \min\{\|p\| \mid p \in \mathbb{R}^n \text{ is a feasible solution of } D_0\}.$$

The most difficult pint in our analysis is to show the property (10) for the minimal norm certificate p_0. This is accomplished by carefully evaluating the analytic form of $f^*(p_0)$. Indeed, by using the orthogonality of $(\theta_j^\circ)_{j=1}^m$ and the fact that the input distribution is the uniform distribution, we can write down the minimal norm certificate and analyze it.

4. Global Convergence of Gradient Method

In this section, we investigate a gradient descent method for the optimization problem (4). We show that under some assumptions, a gradient descent with a norm-dependent step size converges to the global optimum of the problem. We also show that these assumptions for the global convergence are satisfied under the conditions we made in the previous section, which implies the identifiability of the teacher parameters through the gradient descent method.

4.1. Norm-Dependent Gradient Descent

We consider a standard gradient descent for optimizing the objective (4). To incorporate the 1-homogeneity of the ReLU activation function, we employ a step size that can be dependent on the norm of each parameter. As we see in proof of the global convergence, this norm dependency is helpful to describe an update in the measure space. Let F be the regularized empirical risk given in (4), that is, $F(\Theta) := \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i; \Theta))^2 + \lambda \sum_{j=1}^M |a_j| \|w_j\|$. Then, the update rule of the norm-dependent gradient descent can be

[1] $\mathcal{C}(S)$ is the set of continuous functions on a topological space S.

written as

$$a_{j,k+1} = a_{j,k} - \eta_{j,k} g_j(\Theta_k) \text{ for } g_j(\Theta_k) \in \partial_{a_j} F(\Theta_k),$$
$$w_{j,k+1} = w_{j,k} - \eta_{j,k} h_j(\Theta_k) \text{ for } h_j(\Theta_k) \in \partial_{w_j} F(\Theta_k),$$

where $\Theta_k = ((a_{1,k}, w_{j,k}), \ldots, (a_{M,k}, w_{M,k}))$ is the parameter after k iterations, $\eta_{j,k} > 0$ is the norm-dependent step size which will be specified below. $\partial_a F(\Theta)$ denotes the sub-gradient of $F(\Theta)$ as a function of a. The sub-gradient is not always a singleton, but we employ the following one as g, h:

$$g_j(\Theta) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\Theta) - y_i)\sigma(\langle w_j, x_i\rangle) + \lambda \operatorname{sgn}(a_j)\|w_j\|,$$
$$h_j(\Theta) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\Theta) - y_i)a_j x_i \mathbb{1}\{\langle w_j, x_i\rangle \geq 0\} + \lambda \frac{|a_j| w_j}{\|w_j\|},$$

As for the norm-dependent step size $\eta_{j,k}$, we employ the following representation:

$$\eta_{j,k} = \alpha \frac{|a_{j,k}|\|w_{j,k}\|}{a_{j,k}^2 + \|w_{j,k}\|^2}, \tag{11}$$

where $\alpha > 0$ is a fixed constant. For the initialization, we consider the mean-field setting where each $a_{j,0} = \mathrm{O}(1/M)$:

$$a_{j,0} = \frac{2}{M} \quad (1 \leq j \leq M/2),$$
$$a_{j,0} = -\frac{2}{M} \quad (M/2 + 1 \leq j \leq M),$$
$$w_{j,0} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1}).$$

With the norm-dependent step size, the sign of $a_{j,k}$ will not be changed during the optimization, and thus we need the both positive and negative sign initializations for $(a_{j,0})_{j=1}^M$. As pointed out by several authors (Chizat & Bach, 2018; Chizat, 2019; Suzuki & Akiyama, 2021; Mei et al., 2019), it is essentially important to analyze the dynamics of "feature learning" in the mean field regime where each node is adaptively updated to represent the target function efficiently. This is in contrast to NTK analysis (a.k.a., lazy training regime) where the basis functions are almost fixed during the optimization. The algorithm is summarized in Algorithm 1.

The global optimality of the gradient descent can be shown through the measure representation of the neural network. Indeed, we have seen in the previous section that the optimization problem of a neural network model can be generalized to the BLASSO problem on the measure space as presented in Eq. (8). Let J be the BLASSO objective function on the measure space: $J(\nu) = \frac{1}{2n}\sum_{i=1}^n (y_i - f(x_i;\nu))^2 + \lambda\|\nu\|_{\mathrm{TV}}$. Note that in the over-parameterized setting, we cannot formally define the convergence of the parameter Θ_k to the true one Θ° because they have different dimensionality. Therefore, we consider convergence of the measure corresponding to the parameter Θ instead. We assume "sparsity" of the global minima of J on the measure space to ensure the convergence of the measure representation as follows.

Algorithm 1 Norm-Dependent Gradient Descent

Input: student width M (even), max iteration K, stepsize parameter $\alpha > 0$.
Initialization : $a_{j,0} = 2/M, 1 \leq j \leq M/2$, $a_{j,0} = -2/M, M/2+1 \leq j \leq M$, $w_{j,0} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$

```
for k = 1, 2, ..., K do
    for j = 1, ..., M do
        η_{j,k} = α |a_{j,k}| ||w_{j,k}|| / (a_{j,k}^2 + ||w_{j,k}||^2)
        choose g_j(Θ_k) ∈ ∂_{a_j} F(Θ_k)
        choose h_j(Θ_k) ∈ ∂_{w_j} F(Θ_k)
        a_{j,k+1} = a_{j,k} − η_{j,k} g_j(Θ_k)
        w_{j,k+1} = w_{j,k} − η_{j,k} h_j(Θ_k)
    end for
end for
```

Assumption 4.1. *ar The global minimum of J is uniquely attained by a sum of Dirac measures:*

$$\nu^* := \sum_{j=1}^{m^*} r_j^* \delta_{\theta_j^*}, \tag{12}$$

where m^* is a positive integer, $r_j^* \neq 0$, $\theta_j^* \in \mathbb{S}^{d-1}$ ($j \in [m^*]$) and $\theta_j^* \neq \theta_{j'}^*$ for any $j \neq j'$.

Remark 4.2. *Note that this condition can be satisfied under Assumptions 3.1 and 3.2 by Theorem 3.5.*

By the same argument as Proposition 3.4, if we set $M \geq m^*$, the sparsity and uniqueness of the global minimum of J leads to the existence of the global minimum of F, which is essentially represented by m^* nodes. Even in this case, by the non-convexity of F, it is far from trivial to show the convergence of the gradient method to the global optimal solution. As we have stated, we show this through the measure representation of the network.

To show the result, we prepare some additional notations. For the intermediate solution $\Theta_k = \{(a_{j,k}, w_{j,k})\}_{k=1}^M$, we define $r_{j,k} = a_{j,k}\|w_{j,k}\|, \theta_{j,k} = \frac{w_{j,k}}{\|w_{j,k}\|}$ (if $\|w_{j,k}\| = 0$, we set $\theta_{j,k}$ be arbitrary fixed point in $\mathbb{S}^{d-1}$). Accordingly, the measure representation corresponds to Θ_k be

$$\nu_k := \textstyle\sum_{j=1}^M r_{j,k}\delta_{\theta_{j,k}}.$$

For two Radon measures $\mu_1, \mu_2 \in \mathcal{M}(\mathbb{S}^{d-1})$, $W_\infty(\mu_1, \mu_2)$ denotes the Wasserstein distance between them: $W_\infty(\mu_1, \mu_2) := \inf_{\gamma \in \Pi(\mu_1,\mu_2)} \sup_{(\theta_1,\theta_2)\in \mathrm{supp}(\gamma)} \mathrm{dist}(\theta_1, \theta_2)$, where $\Pi(\mu_1, \mu_2)$ is a set of product measures with marginals μ_1 and μ_2, $\mathrm{supp}(\gamma)$ is the support of γ, and $\mathrm{dist}(\theta_1, \theta_2) := \arccos(\langle \theta_1, \theta_2 \rangle)$ for $\theta_1, \theta_2 \in \mathbb{S}^{d-1}$.

Since $f(x;\nu)$ is a linear model with respect to ν and the squared loss is differentiable, the Fréchet subdifferential of $J(\nu)$ on $\mathcal{M}(\mathbb{S}^{d-1})$ can be defined and be represented as a set of functions $G(\cdot) : \mathbb{S}^{d-1} \to \mathbb{R}$ defined by

$$G(\theta) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\nu) - y_i)\sigma(\langle \theta, x_i \rangle) + \lambda\eta(\theta),$$

where $\eta \in \mathcal{C}(\mathbb{S}^{d-1})$ satisfies $\|\eta\|_\infty \leq 1$ and $\int \eta \mathrm{d}\nu = \|\nu\|_{\mathrm{TV}}$. Note that we have that $\partial J(\nu_k) := \{G \in \mathcal{C}(\mathbb{S}^{d-1}) \mid J(\mu) - J(\nu_k) \geq \int G(\theta)\mathrm{d}(\mu - \nu_k)$ for any $\mu \in \mathcal{M}(\mathbb{S}^{d-1})\}$ which is well defined because $J(\cdot)$ is a convex function on the measure space $\mathcal{M}(\mathbb{S}^{d-1})$.

4.2. Main Result 2: Global Optimality of Gradient Method

Here, we give the global convergence property of the norm-dependent gradient descent under a bit milder conditions than those assumed in the previous section. The analysis basically follows that of Chizat (2019), but they assumed smoothness on the activation and excluded the ReLU activation. To overcome this difficulty, our norm-dependent step size (Eq. (11)) plays the important role. Moreover, we carefully divide the parameter space into "smooth region" and "non-smooth irrelevant-region" to show a descent property of the objective. The assumptions below are made under a condition of a training data observation $D_n = (x_i, y_i)_{i=1}^n$.

Assumption 4.3 (Non-orthogonality between x and θ)**.** *For any $i \in [n], j \in [m]^*$, we have $\langle x_i, \theta_j^* \rangle \neq 0$.*

Assumption 4.4 (Strong convexity w.r.t. r)**.** *There exists a constant $\kappa > 0$ such that for any $r_1, \dots, r_m \in \mathbb{R}$, $\|\sum_{j=1}^m r_j\sigma(\langle \theta_j^*, \cdot \rangle)\|_n^2 \geq \kappa(r_1^2 + \cdots + r_m^2)$.*

Assumption 4.5 (Non-degeneracy)**.** *There exists no $\theta \notin \mathrm{supp}(\nu^*)$ such that $J'(\nu^*)(\theta) = 0$.*

Assumption 4.6 (Boundedness)**.** *There exists a constant $C_F > 0$ such that, for any k, it holds that $F(\Theta_k) \leq C_F$.*

Assumption 4.7 (Boundedness of input)**.** $\|x_i\| \leq 1$ *for all* $i \in [n]$.

Assumption 4.3 is satisfied almost surely if $x_i \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. This is required to ensure the smoothness of the objective around the optimal parameter $(r_j^*, \theta_j^*)_{j=1}^{m^*}$. Otherwise the objective function F is non-differentiable at the global optimal with respect to θ_j, which causes difficulty to show the local convergence around the global optimal. Assumption 4.4 is also almost surely satisfied if the nodes $x \mapsto \sigma(\langle x, \theta_j^* \rangle)$ $(j \in [m^*])$ are linearly independent in $L_2(P_\mathcal{X})$. Assumption 4.5 is a bit tricky but is assumed in several existing work (Duval & Peyré, 2015; Chizat, 2019; Flinth et al., 2020) ensures that the true parameters $(\theta_j^*)_{j=1}^{m^*}$ are uniquely determined. Assumption 4.5 is also needed to ensure that in a local convergence phase, which we describe in Theorem 4.8, ν_k vanishes rapidly far away from $(\theta_j)_{j=1}^{m^*}$. This assumption can be verified under the same setting as Theorem 3.5 by utilizing a dual certificate argument. Assumption 4.7 is just fixing the scaling factor and is satisfied under the setting $x_i \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ (Assumption 3.1).

Theorem 4.8. *Assume that Assumptions 4.1, 4.3–4.7 hold. Let $\tau = \mathrm{Unif}(\mathbb{S}^{d-1})$, $\nu_0^+ = 2/M\sum_{j=1}^{M/2}\delta_{w_{j,0}}, \nu_0^- = 2/M\sum_{j=M/2+1}^{M}\delta_{w_{j,0}}$ and $J^* = J(\nu^*)$. Then, for any $0 < \epsilon < 1/2$, there exist constants $\rho, C, C', C_M > 0$, $J_0 > J^*$, $\kappa_0 > 0$ such that if $\alpha > 0$ satisfies*

$$\alpha < \min\{(J_0 - J^*)^{1+\epsilon/2}/C, 1/8C_1, 1/10C_2, \rho/C_2, \lambda^2/8C_F^2\}$$

with $C_1 = 2\sqrt{n}C_F + \lambda$ and $C_2 = 2\sqrt{n}C_F$, the width M is sufficiently over-parameterized as $M \geq C_M\exp(\alpha^{-2})/\alpha$, and the initial solution satisfies

$$\max\{W_\infty(\tau, \nu_0^+),\ W_\infty(\tau, \nu_0^-)\} \leq (J_0 - J^*)/C,$$

then we have the following convergence properties:
(1) Global exploration: There exists $k_0 \geq C'(J_0 - J^*)^{-(2+\epsilon)}$ such that for any $k \geq k_0$, it holds that

$$J(\nu_k) - J^* \leq J_0 - J^*.$$

(2) Local convergence: For any $k \geq k_0$, it holds that

$$J(\nu_k) - J^* \leq (J(\nu_0) - J^*)(1 - \kappa_0)^{k-k_0}.$$

Therefore, combining these results, we see that $J(\nu_k)$ converges to $J(\nu^*)$.

The proof can be found in Appendix B. This theorem implies that the norm-dependent gradient descent can converge to the global optimal solution in terms of both the measure on parameters and the function value. Its dynamics consists of two phases: (1) the global exploration regime, and (2) the local linear convergence regime. In the first phase, the gradient descent explores the parameter space to roughly capture the location of the optimal parameters. In the second phase, the dynamics enters a local region around the optimal parameters where the objective is locally strongly convex. After entering this phase, the parameters converge to the optimal solution linearly. In that sense, J_0 represents a threshold that separates the global region and local near strongly convex region. During the optimization, the sparse regularization works for eliminating the amplitudes of nodes that are far away from the optimal parameters. This kind of "two phase" dynamics has been pointed out by several authors (e.g., Li & Yuan (2017); Chizat (2019)), but it has not been shown for the ReLU fully connected neural networks.

The condition $\max\{W_\infty(\tau, \nu_0^+), \quad W_\infty(\tau, \nu_0^-)\} \le (J_0 - J^*)/C$ requires that M is sufficiently over-parameterized. It is known that $W_\infty(\tau, \nu_0^\pm) = \mathrm{O}_p((\log M)^{1/(d-1)} M^{-1/(d-1)})$ for $d > 3$ (Trillos & Slepčev, 2015). Therefore, it is implicitly assumed that $M \ge \Omega((J_0 - J^*)^{-(d-1)} \log_+(1/(J_0 - J^*))^{(d-1)})$[2]. The condition $M \ge C_M \exp(\alpha^{-2})/\alpha$ also requires the over-parameterization and the right side may be quite large. This condition is only required for the global exploration ((1) in Theorem 4.8). The over-parameterization and the norm-dependency of stepsize ensure that $(\theta_{j,k})_{j=1}^M$ do not move far away from initialization until the function value decrease enough. By this property, the gradient descent can "identify" an informative subset of parameters $(\theta_{j,k})_{j=1}^M$, which are close to the optimal parameters $(\theta_j^*)_{j=1}^{m*}$. It may be possible to ensure that under the less number of parameters M, the gradient descent "automatically" reaches around each of the optimal parameters and can accomplish the global exploration. We leave this issue for future work. Finally, we mention a remark on a condition on the constant ρ and the regularization parameter λ for Theorem 4.8. Roughly speaking, ρ represents a diameter of a local smooth region around each optimal parameter θ_j^*. Under Assumptions 3.1 and 3.2, it suffices to take $\rho = \mathrm{O}_p(1/nm)$ if θ_j^* and θ_j° are sufficiently close for any $j \in [m]$ (see Lemma B.18). It can be shown that this closeness condition between θ_j^* and θ_j° holds with high probability by setting $\lambda = \mathrm{O}(1/nm^{3/2})$ by Theorem 3.5. These estimates are derived from conservative evaluations and could be larger for each concrete realization of $(x_i)_{i=1}^n$.

In addition to this convergence property in terms of the objective function, we can show convergence in terms of the L_∞-norm.

Theorem 4.9. *Under Assumptions 4.1, 4.3–4.7, there exists $C'' > 0$ such that for all $k \ge k_0$, it holds that*

$$\|f(x; \nu_k) - f(x; \nu^*)\|_\infty \le C''(J(\nu_0) - J^*)(1 - \kappa_0)^{k-k_0},$$

where k_0 and κ_0 are those introduced in Theorem 4.8.

To show this, we prove that the measure representation ν_k converges to the optimal representation ν^* in terms of a modified 2-Wasserstein distance. The details can be found in Section B.6.

Near Exact Recovery by Gradient Descent Finally, combining Theorem 3.5 and Theorem 4.9, we obtain the following corollary that asserts that the student network converges near the teacher network by the gradient descent method. To show this, we need to prove that Assumptions 3.1 and 3.2 implies Assumptions 4.1, 4.3–4.7. The proof can be found in Section B.6.

[2] $\log_+(x)$ denotes $\max\{\log(x), 1\}$.

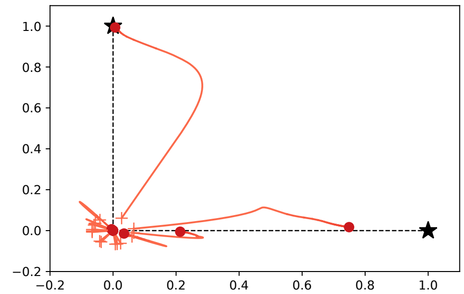

Figure 1. Illustration of the optimization dynamics with $d = 2$ and $m = 2$. The true parameters are indicated by $\star$, the initial solution of each node is indicated by an orange $+$, and its final state is indicated by the red $\circ$.

Corollary 4.10. *Under Assumptions 3.1 and 3.2, suppose that $n > \mathrm{poly}(m, d, \log 1/\delta)$ for $\delta > 0$, then, with probability at least $1 - \delta$, it holds that the $L_2(P_\mathcal{X})$-norm between $f(\cdot; \nu_k)$ and f° can be bounded as*

$$\begin{aligned}&\|f(\cdot; \nu_k) - f^\circ\|_{L_2(P_\mathcal{X})}^2 \\ &\le 2C''^2(J(\nu_0) - J^*)^2(1 - \kappa_0)^{2(k-k_0)} + O(m\lambda^2),\end{aligned}$$

dependent on the observation D_n. for all $k \ge k_0$ where k_0 and κ_0 are constants introduced in Theorem 4.8 that could depend on the observation D_n.

5. Numerical Experiments

In this section, we conduct numerical experiments to justify our theoretical results.

Illustration in two dimensional space. First, we give an illustrative example in which the dynamics of the student network is depicted in a two dimensional setting $d = 2$. In this experiment, we employ $m = 2$ with $r_1^\circ = r_2^\circ = 1$ and $\theta_1^\circ = (1, 0)^\top$, $\theta_2^\circ = (0, 1)^\top$, $M = 15$, and $n = 100$. Figure 1 shows the optimization trajectory of $(a_{j,k}, w_{j,k})_{j=1}^M$. We can see that the nodes with initialization near to a teacher parameter approaches one of the nodes in the teacher network and, on the other hand, the nodes with initialization far away from any teacher node finally vanish. This behavior is induced by the sparse regularization, that is, the sparse regularization "selects" informative nodes and discard non-informative nodes. We also see that the selected nodes explore a wide area in the early stage and after that they finally head to the direction of one of the teacher nodes. This well justifies our theoretical analysis.

Effect of over-parameterization for convergence. Next, we investigate how the over-parameterization affects the dynamics. In this experiment, we employ $m = 5$ for the

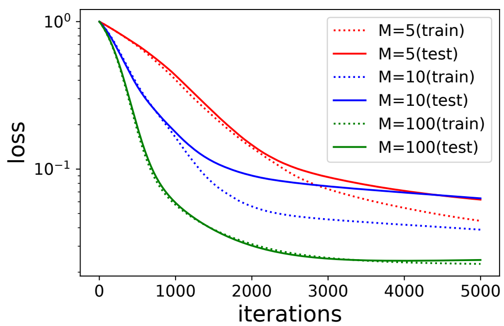

Figure 2. Convergence of the training/test loss for different student width $M = 5, 10, 100$.

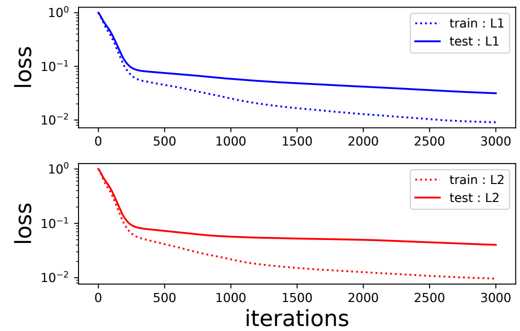

Figure 3. Comparison of L_1 and L_2- regularizations.

teacher width, $d = 5$ for the dimensionality and $n = 100$ for the sample size. As for the student network, we compare the dynamics between $M = 5, 10, 100$. Figure 2 depicts the training loss and test loss against the number of iterations. Each line corresponds to different setting of M. We can see that a sufficiently over-parameterized network ($M = 100$) appropriately estimates the true function while a narrow network ($M = 5$) does not reach the global optimal solution. We also note that the test loss is almost same as the training loss in the over-parameterized setting while we observe over-fitting for $M = 5$ and $M = 10$. This means that the solution in the over-parameterized setting ($M = 100$) finally converges to the optimal "sparse" solution that avoids the over-fitting. This is consistent to the findings by the existing studies (Safran & Shamir, 2018; Safran et al., 2020).

Comparison of L_1 and L_2 Regularization Inspired by Eq. (5), we also conduct norm-dependent gradient descent for the L_2-regularized problem:

$$\min_{\Theta \in (\mathbb{R}\times\mathbb{R}^d)^M} \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i;\Theta))^2 + \frac{\lambda}{2}\sum_{j=1}^{M}(a_j^2 + \|w_j\|^2). \tag{13}$$

We give a comparison of the loss evolution between the L_1-regularization and L_2-regularization in Figure 3. In this experiment, we employ $m = 5$ for the teacher width, $d = 5$ for the dimensionality, $n = 100$ for the sample size and $M = 10$ for the student width. We can see that both regularizations show the almost same trajectory of the loss functions. This indicates the usefulness of the practical use of the L_2-regularization.

6. Conclusion

In this paper, we have investigated identifiability of the true target function via the gradient descent method for two-layer ReLU neural networks in teacher-student settings. We have shown that with the sparse regularization, the global minima can be arbitrarily close to the teacher network. Furthermore, we have proposed a gradient method with norm-dependent step size which is guaranteed to converge to the global minima, and shown that this framework can be applied to the teacher-student setting. The key ingredient in this analysis is the measure representation of the ReLU network. With this perspective, the gradient method can be associated with gradient descent in the measure space. We believe that this analysis gives a new insight into learnability in the teacher-student setting.

Acknowledgement

TS was partially supported by JSPS KAKENHI (18H03201, and 20H00576), Japan Digital Design and JST CREST.

References

Absil, P. A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *Proceedings of International Conference on Machine Learning*, pp. 242–252, 2019.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Borwein, J. M. and Zhu, Q. J. *Techique of Variational Analysis*. Springer, 2005.

Bredies, K. and Pikkarainen, H. K. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

Cai, T. T., Fan, J., and Jiang, T. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 14:1837–1864, 2013.

Candès, E. J. and Fernandez-Granda, C. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

Chizat, L. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. volume 31, pp. 3036–3046, 2018.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

De Castro, Y. and Gamboa, F. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

de Dios, J. and Bruna, J. On sparsity in overparametrised shallow ReLU networks. *arXiv preprint arXiv:2006.10225*, 2020.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations 7*, 2019.

Duval, V. and Peyré, G. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

E, W., Ma, C., and Wu, L. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

Erdogdu, M. A., Mackey, L., and Shamir, O. Global nonconvex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems 31*, pp. 9671–9680. 2018.

Flinth, A., de Gournay, F., and Weiss, P. On the linear convergence rates of exchange and continuous methods for total variation minimization. *Mathematical Programming*, pp. 1–37, 2020.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 315–323, 2011.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., and Zdeborová, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, pp. 6981–6991, 2019.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9482–9491, 2018.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 8580–8589, 2018.

Klusowski, J. M. and Barron, A. R. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pp. 6389–6399, 2018.

Li, Y. and Yuan, Y. Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 597–607. Curran Associates, Inc., 2017.

Li, Y., Ma, T., and Zhang, H. R. Learning over-parametrized two-layer neural networks beyond NTK. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2613–2682. PMLR, 2020.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pp. 2388–2464, 2019.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401, 2015.

Nitanda, A. and Suzuki, T. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Poon, C., Keriven, N., and Peyré, G. The geometry of off-the-grid compressed sensing. *arXiv preprint arXiv:1802.08464*, 2018.

Poon, C., Keriven, N., and Peyré, G. Support localization and the fisher metric for off-the-grid sparse regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1341–1350. PMLR, 2019.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *arXiv e-prints*, pp. arXiv:1702.03849, 2017.

Rockafeller, R. T. Duality and stability in extremum problems involving convex functions. *Pacific Journal of Mathematics*, (1):167–188, 1967.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning*, pp. 4433–4441. PMLR, 2018.

Safran, I., Yehudai, G., and Shamir, O. The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. *arXiv preprint arXiv:2006.01005*, 2020.

Suzuki, T. and Akiyama, S. Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. In *International Conference on Learning Representations*, 2021.

Tian, Y. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560*, 2017.

Tian, Y. Student specialization in deep rectified networks with finite width and input dimension. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9470–9480. PMLR, 2020.

Trillos, N. G. and Slepčev, D. On the rate of convergence of empirical measures in-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

Tropp, J. A. *An Introduction to Matrix Concentration Inequalities*, volume 8 of *Foundations and Trends in Machine Learning*. Now Publishers Inc. Hanover, MA, USA, 2015.

Tzen, B. and Raginsky, M. A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled McKean-Vlasov dynamics. *arXiv preprint arXiv:2002.01987*, 2020.

Weinan, E., Ma, C., and Wu, L. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics*, pp. 1–24, 2020.

Welling, M. and Teh, Y.-W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, 2011.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 09–12 Jul 2020.

Yehudai, G. and Shamir, O. Learning a single neuron with gradient methods. *arXiv preprint arXiv:2001.05205*, 2020.

Zhang, X., Yu, Y., Wang, L., and Gu, Q. Learning one-hidden-layer relu networks via gradient descent. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1524–1534. PMLR, 2019.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

Zhou, M., Ge, R., and Jin, C. A local convergence theory for mildly over-parameterized two-layer neural network. *arXiv preprint arXiv:2102.02410*, 2021.

A. Proof of Theorem 3.5 and related topics

In this section, we give the proof of Theorem 3.5 and auxiliary lemmas to prove it.

A.1. Preliminaries

In this section, we give the proof of the main result I (Theorem 3.5). The key tool is the *dual certificate* and the *NDSC condition* (Definition A.5) which were introduced by Duval & Peyré (2015). We firstly introduce these concepts, and then prove the assertion by using them.

A.1.1. Dual Problem and Optimality Condition

As described in Eq. (8), we consider the following optimization problem on the measure space:

$$\min_{\nu\in\mathcal{M}(\mathbb{S}^{d-1})} \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i;\nu))^2 + \lambda\|\nu\|_{\mathrm{TV}}. \tag{P_λ}$$

By regarding f as a linear operator $f(\cdot):\mathcal{M}(\mathbb{S}^{d-1})\to\mathbb{R}^n$, $\nu\mapsto(f(x_1;\nu),\ldots,f(x_n;\nu))^\top$, we can define its adjoint operator $f^*:\mathbb{R}^n\to\mathcal{C}(\mathbb{S}^{d-1})$ as

$$f^*(p)(\theta) = \frac{1}{n}\sum_{i=1}^{n} p_i\sigma(\langle\theta, x_i\rangle).$$

Then, we can obtain the dual problem of (P_λ) through the Fenchel duality theorem (Rockafeller, 1967; Borwein & Zhu, 2005; Duval & Peyré, 2015):

$$\max_{p\in\mathbb{R}^n:\|f^*(p)\|_\infty\le 1} \quad \frac{1}{n^2}\sum_{i=1}^{n} y_i p_i - \frac{\lambda}{2n^2}\|p\|^2. \tag{D_λ}$$

This dual problem (D_λ) can be reformulated as

$$\min_{p\in\mathbb{R}^n:\|f^*(p)\|_\infty\le 1} \quad \frac{1}{n^2}\left\| p - \frac{1}{\lambda}\begin{pmatrix} y_1\\ \vdots \\ y_n\end{pmatrix}\right\|^2. \tag{$\widetilde{D}_\lambda$}$$

Note that solutions of this problem are expressed by a projection of $(y_1,\ldots,y_n)^\top\in\mathbb{R}^n$ onto a closed convex subset $\{p\in\mathbb{R}^n \mid \|f^*(p)\|_\infty\le 1\}$ which is uniquely determined by the Hilbert projection theorem.

By taking the limit of $\lambda\to+0$ in Eq. (8), we obtain the following problem:

$$\min_{\mu\in\mathcal{M}(\mathbb{S}^{d-1})}\|\mu\|_{\mathrm{TV}} \quad \text{s.t.} \quad f(x_i;\mu)=y_i \quad (\forall i\in[n]). \tag{P_0}$$

The dual problem of this is given by

$$\max_{\|f^*(p)\|_\infty\le 1} \quad \frac{1}{n^2}\sum_{i=1}^{n} y_i p_i. \tag{D_0}$$

The strong duality between these problems can be characterized by the *subdifferential* of the object function. In particular, we require the subdifferential $\partial\|\nu\|_{\mathrm{TV}}$ of the total variation norm which is expressed by

$$\partial\|\nu\|_{\mathrm{TV}} = \left\{\eta\in\mathcal{C}(\mathbb{S}^{d-1}) \mid \|\eta\|_\infty\le 1, \int\eta\mathrm{d}\mu = \|\nu\|_{\mathrm{TV}}\right\}.$$

For $\lambda>0$, we can show that the strong duality holds between (P_λ) and (D_λ), which means that both problems have the same optimal value and any solution ν of (P_λ) is linked with the unique solution p of (D_λ) by

$$\begin{cases} f^*(p)\in\partial\|\nu\|_{\mathrm{TV}}, \\ p_i = -\frac{1}{\lambda}(f(x_i;\nu)-y_i) \quad (\forall i\in[n]). \end{cases} \tag{14}$$

Conversely, if there exists a pair $(\nu, p) \in \mathcal{M}(\mathbb{S}^{d-1}) \times \mathbb{R}^n$ satisfying Eq. (14), then ν is an optimal solution of (P_λ) and p is the unique solution of (D_λ).

Strong duality also holds between (P_0) and (D_0). If an optimal solution p^* of (D_0) exists, then it is linked to any solution ν of (P_0) by

$$\begin{cases} f^*(p) \in \partial \|\nu\|_{\mathrm{TV}}, \\ f(x_i; \mu) = y_i \quad (\forall i \in [n]), \end{cases} \tag{15}$$

and similarly, if there exists a pair $(\nu, p) \in \mathcal{M}(\mathbb{S}^{d-1}) \times \mathbb{R}^n$ satisfying Eq. (15), then ν is an optimal solution of (P_0) and p is a solution of (D_0).

In particular when ν is written by a sum of Dirac measures as $\nu = \sum_{j=1}^m r_j \delta_{\theta_j}$, $f^*(p) \in \partial \|\nu\|_{\mathrm{TV}}$ is equivalent to

$$\begin{cases} f^*(p)(\theta_j) = \mathrm{sgn}(r_j) & (\forall j \in [m]), \\ |f^*(p)(\theta)| \leq 1 & (\forall \theta \in \mathbb{S}^{d-1}). \end{cases} \tag{16}$$

We use the next proposition to prove the main theorem. The proof is remained to the latter of this section.

Proposition A.1. *Let $n > \mathrm{poly}(m, d, \log 1/\delta)$. Then, with probability at least $1-\delta$, for any $\epsilon > 0$, there exists $\lambda = \lambda(\epsilon)$ such that the optimal solution p_λ of (D_λ) satisfies*

$$\begin{cases} f^*(p)(\theta_j^*) = 1 & (\forall j \in [m]), \\ |f^*(p)(\theta)| < 1 & (\forall \theta \in \mathbb{S}^{d-1},\ \theta \neq \theta_j^*), \end{cases}$$

where $(\theta_j^*)_{j\in[m]} \subset \mathbb{S}^{d-1}$ satisfying $\|\theta_j^\circ - \theta_j^*\| < \epsilon$ $(\forall j \in [m])$. Moreover, the global minima of (P_λ) is written by $\nu^* = \sum_{j=1}^m r_j^* \delta_{\theta_j^*}$, where $(r_j^*, \theta_j^*)_{\in[m]}$ satisfies Eq. (9).

Remark A.2. *Since $f^*(p)$ is piecewise-linear function for any p (following from the same property of ReLU), we know that the global minima of (P_λ) is expressed by a sum of at most $O(n^{d+1})$ Dirac measures independently of the sample $(x_i)_{i=1}^n$. This result can be extended to any other 1-homogeneous activation function. The same result is derived in de Dios & Bruna (2020) by another approach, our argument above gives another perspective to the characterization of the optimal solution.*

A.1.2. NON DEGENERATE SOURCE CONDITION

Unlike (D_λ), (D_0) does not always have a unique solution. Then we consider the following concept, which is crucial in this proof.

Definition A.3 (minimal norm certificate (Duval & Peyré, 2015))**.** *The minimal norm certificate associated with (P_λ) is defined as $f^*(p_0)$, where p_0 is the minimum norm solution of (D_0) if it exists, i.e.,*

$$p_0 = \arg\min \ \{\|p\| \mid p \textit{ is a solution of } \ (D_0)\}.$$

Minimum norm certificate is linked with the unique solution of (D_λ) in the following sense:

Lemma A.4 (Duval & Peyré (2015))**.** *Let p_λ be the unique solution of (D_λ). Then p_λ converge to p_0 as $\lambda \to +0$, where p_0 is the minimal norm solution of D_0.*

Using this Lemma, we can show that under $\lambda \to +0$, the global minima of (P_λ) has its support which is arbitrary close to that of (P_0). Therefore we focus on (P_0) and introduce the following concept. Let $\nabla_{\mathbb{S}^{d-1}} := (\mathrm{I}_d - \theta\theta^\top)\nabla$ which represents the derivative on $\mathbb{S}^{d-1}$. We note that $\nabla_{\mathbb{S}^{d-1}} f(\theta) = 0$ means $\nabla f(\theta) = a\theta$ for some $a \in \mathbb{R}$.

Definition A.5 (NDSC (Non-Degenerate Source Condition) (Duval & Peyré, 2015))**.** *We say that $\nu = \sum_{j=1}^m r_j \delta_{\theta_j}$ satisfies NDSC if the minimal norm certificate $f^*(p_0)$ satisfies the following condition:*

- $f^*(p_0)(\theta_j) = \mathrm{sgn}(r_j)$ $(\forall j \in [m])$*,*
- $|f^*(p_0)(\theta)| < 1$ *for any* $\theta \in \mathbb{S}^{d-1}$ *such that* $\theta \neq \theta_j$ $(\forall j \in [m])$*,*
- $\nabla^2_{\mathbb{S}^{d-1}} f^*(p_0)(\theta_j)$ *is invertible for any* $j \in [m]$*.*

Through the second and the third conditions, we can verify that for the unique solution p_λ of (D_λ) $f^*(p_\lambda)(\theta) = 1$ holds only in the neighborhood of $\theta = \theta_j$. Hence, the optimal solution of (P_λ) has its support only around θ_j. This yields that θ_j^* is close to the teacher parameter θ_j° for sufficiently small λ. Therefore, we just need to show NDSC for p_0, but it is hard to obtain the closed form of p_0. To overcome this difficulty, we consider a "loose" version of p_0, which is called *pre-certificate*.

Definition A.6 (pre-certificate (Duval & Peyré, 2015))**.** *The pre-certificate associated with* (P_λ) *is defined as* $f^*(p^\dagger)$*, where*

$$p^\dagger = \arg\min\ \{\|p\| \mid 1 \leq \forall j \leq m, f^*(p)(\theta_j) = 1, \nabla_{\mathbb{S}^{d-1}} f^*(p)(\theta_j) = 0\}.$$

Pre-certificate can be expressed by the minimal norm solution of a linear equation as we see below. If the pre-certificate $f^*(p^\dagger)$ satisfies the conditions in NDSC by replacing p_0 with $p^\dagger$, then $p^\dagger$ is an optimal solution of D_0 by the optimality condition (16). Moreover, by noticing that $\|p^\dagger\| \leq \|p_0\|$, if $f^*(p^\dagger)$ achieves the conditions in NDSC, it holds that $p^\dagger = p_0$ and thus the NDSC condition holds for ν, which yields the optimality of ν. Therefore, we show that the pre-certificate $f^*(p^\dagger)$ satisfies the conditions in NDSC instead of directly showing it for the minimal norm certificate $f^*(p_0)$.

A.2. NDSC in the teacher-student settings

As we discussed in the previous section, we show the following property:

Proposition A.7 (NDSC in the teacher-student setting)**.** *Under Assumptions 3.1 and 3.2, for* $n > \mathrm{poly}(m, d, \log(1/\delta))$ *with* $\delta > 0$*, the pre-certificate associated with the teacher-student settings satisfies the following conditions with probability at least* $1 - \delta$*:*

- $f^*(p^\dagger)(\theta_j^\circ) = 1\ \ (\forall j \in [m])$.
- $|f^*(p^\dagger)(\theta)| < 1$ *for any* $\theta \neq \theta_j^\circ$ $(\forall j \in [m])$.
- $f^*(p^\dagger)$ *is strictly concave at* $\theta = \theta_j^\circ$ $(\forall j \in [m])$.

From now on, we show this proposition. At first, we consider how the pre-certificate can be characterized in this setting. When $f^*(p)(\cdot)$ is differentiable at θ_j° as a function of θ ($\Leftrightarrow$ there is no x_i that is orthogonal to θ_j°, which holds a.s. for all $j \in [m]$), the extremality condition is given as follows:

$$\begin{cases} f^*(p)(\theta_j^\circ) = 1, \\ \nabla f^*(p)(\theta_j^\circ) = \alpha \theta_j^\circ. \end{cases}$$

For the ReLU activation, these are equivalent to

$$\nabla f^*(p_0)(\theta_j^\circ) = \theta_j^\circ, \tag{17}$$

since it holds that $\langle \theta_j^\circ, \nabla f^*(p_0)(\theta_j^\circ)\rangle = f^*(p_0)(\theta_j^\circ)$. By writing down this equation, we get

$$\frac{1}{n}\sum_{i=1}^{n} p_i x_i \mathbb{1}\{\langle \theta_j^\circ, x_i\rangle \geq 0\} = \theta_j^\circ.$$

By considering the same equation for all $j \in [m]$ and combining them, we get a linear equation about p as

$$\frac{1}{n}\begin{pmatrix} x_1 \mathbb{1}\{\langle \theta_1^\circ, x_1\rangle \geq 0\} & x_2 \mathbb{1}\{\langle \theta_1^\circ, x_2\rangle \geq 0\} & \dots & x_n \mathbb{1}\{\langle \theta_1^\circ, x_n\rangle \geq 0\} \\ x_1 \mathbb{1}\{\langle \theta_2^\circ, x_1\rangle \geq 0\} & x_2 \mathbb{1}\{\langle \theta_2^\circ, x_2\rangle \geq 0\} & \dots & x_n \mathbb{1}\{\langle \theta_2^\circ, x_n\rangle \geq 0\} \\ \vdots & \vdots & \ddots & \vdots \\ x_1 \mathbb{1}\{\langle \theta_m^\circ, x_1\rangle \geq 0\} & x_2 \mathbb{1}\{\langle \theta_m^\circ, x_2\rangle \geq 0\} & \dots & x_n \mathbb{1}\{\langle \theta_m^\circ, x_n\rangle \geq 0\} \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} = \begin{pmatrix} \theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ \end{pmatrix}.$$

By definition, $p^\dagger$ is the minimum norm solution of this equation and represented by

$$p^\dagger = n X_0^\dagger \begin{pmatrix} \theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ \end{pmatrix},$$

where

$$X_0 = \begin{pmatrix} x_1 \mathbb{1}\{\langle \theta_1^\circ, x_1 \rangle \geq 0\} & x_2 \mathbb{1}\{\langle \theta_1^\circ, x_2 \rangle \geq 0\} & \dots & x_n \mathbb{1}\{\langle \theta_1^\circ, x_n \rangle \geq 0\} \\ x_1 \mathbb{1}\{\langle \theta_2^\circ, x_1 \rangle \geq 0\} & x_2 \mathbb{1}\{\langle \theta_2^\circ, x_2 \rangle \geq 0\} & \dots & x_n \mathbb{1}\{\langle \theta_2^\circ, x_n \rangle \geq 0\} \\ \vdots & \vdots & \ddots & \vdots \\ x_1 \mathbb{1}\{\langle \theta_m^\circ, x_1 \rangle \geq 0\} & x_2 \mathbb{1}\{\langle \theta_m^\circ, x_2 \rangle \geq 0\} & \dots & x_n \mathbb{1}\{\langle \theta_m^\circ, x_n \rangle \geq 0\} \end{pmatrix} \in \mathbb{R}^{md \times n}$$

and $X_0^\dagger$ denotes the Moore-Penrose inverse. Especially when X_0 has full row rank (which we verify in the latter w.h.p.), it holds that

$$p^\dagger = n X_0^\mathsf{T} (X_0 X_0^\mathsf{T})^{-1} \begin{pmatrix} \theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ \end{pmatrix}. \tag{18}$$

Therefore, we get the closed form of $f^*(p^\dagger)$ as follows.

Lemma A.8. *Suppose that X_0 has full row rank. Let $X(\theta)$ be*

$$X(\theta) = \left(x_1 \mathbb{1}\{\langle \theta, x_1 \rangle \geq 0\}, \quad x_2 \mathbb{1}\{\langle \theta, x_2 \rangle \geq 0\}, \quad \dots \quad , \quad x_n \mathbb{1}\{\langle \theta, x_n \rangle \geq 0\} \right).$$

Then the following equality holds:

$$f^*(p^\dagger)(\theta) = \frac{1}{n} \theta^\mathsf{T} \left(X(\theta) X_0^\mathsf{T} \right) \left(\frac{1}{n} X_0 X_0^\mathsf{T} \right)^{-1} \begin{pmatrix} \theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ \end{pmatrix}.$$

Each matrix in the expression of $f^*(p^\dagger)$ of the above lemma can be written as follows:

$$\frac{1}{n} X_0 X_0^\mathsf{T} = \begin{pmatrix} K_{1,1} & K_{1,2} & \dots & K_{1,m} \\ K_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & K_{m-1,m} \\ K_{m,1} & \dots & K_{m,m-1} & K_{m,m} \end{pmatrix} \in \mathbb{R}^{dm \times dm},$$

where

$$K_{j_1, j_2} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\mathsf{T} \mathbb{1}\{\langle \theta_{j_1}^\circ, x_i \rangle \geq 0, \langle \theta_{j_2}^\circ, x_i \rangle \geq 0\} \in \mathbb{R}^{d \times d},$$

and

$$\frac{1}{n} X(\theta) X_0^\mathsf{T} = (K_1(\theta), K_2(\theta), \dots, K_m(\theta)) \in \mathbb{R}^{d \times dm},$$

where

$$K_j(\theta) = \frac{1}{n} \sum_{i=1}^n x_i x_i^\mathsf{T} \mathbb{1}\{\langle \theta_j^\circ, x_i \rangle \geq 0, \langle \theta, x_i \rangle \geq 0\} \in \mathbb{R}^{d \times d}.$$

Since these two matrices $\frac{1}{n} X_0 X_0^\mathsf{T}$ and $\frac{1}{n} X(\theta) X_0^\mathsf{T}$ depend on the sample observation $(x_i)_{i=1}^n$, it is hard to obtain its close form expression. On the other hand, these are empirical versions of $\mathbb{E}_{D_n}\left[\frac{1}{n} X_0 X_0^\mathsf{T}\right]$ and $\mathbb{E}_{D_n}\left[\frac{1}{n} X(\theta) X_0^\mathsf{T}\right]$, respectively. Fortunately, we can write them down by closed forms, and thus we consider the population version $\bar{f}(\theta)$ of $f^*(p^\dagger)(\theta)$ instead, i.e.,

$$\bar{f}(\theta) = \theta^\mathsf{T} \mathbb{E}_{D_n} \left[\frac{1}{n} X(\theta) X_0^\mathsf{T} \right] \mathbb{E}_{D_n} \left[\frac{1}{n} X_0 X_0^\mathsf{T} \right]^{-1} \begin{pmatrix} \theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ \end{pmatrix}. \tag{19}$$

Lemma A.9. *Under the Assumption 3.1, the matrices in (19) are written by follows:*

$$\mathbb{E}_{D_n}\left[\frac{1}{n}X_0X_0^\top\right] = \frac{1}{d}\begin{pmatrix} \frac{1}{2}\mathrm{I}_d & \frac{1}{4}\mathrm{I}_d & \cdots & \frac{1}{4}\mathrm{I}_d \\ \frac{1}{4}\mathrm{I}_d & \frac{1}{2}\mathrm{I}_d & \cdots & \frac{1}{4}\mathrm{I}_d \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{4}\mathrm{I}_d & \frac{1}{4}\mathrm{I}_d & \cdots & \frac{1}{2}\mathrm{I}_d \end{pmatrix} + \frac{1}{2\pi d}\begin{pmatrix} 0_d & E_{1,2} & \cdots & E_{1,m} \\ E_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & E_{m-1,m} \\ E_{m,1} & \cdots & E_{m,m-1} & 0_d \end{pmatrix},$$

where E_{j_1,j_2} is the symmetric matrix $\theta^\circ_{j_1}\theta^{\circ\top}_{j_2} + \theta^\circ_{j_2}\theta^{\circ\top}_{j_1}$.

$$\theta^\top\mathbb{E}_{D_n}\left[\frac{1}{n}X(\theta)X_0^\top\right] = \frac{1}{2d}\left(\tfrac{\pi-\phi_1}{\pi}\theta^\top + \tfrac{\sin\phi_1}{\pi}\theta_1^{\circ\top}, \quad \tfrac{\pi-\phi_2}{\pi}\theta^\top + \tfrac{\sin\phi_2}{\pi}\theta_2^{\circ\top}, \quad \ldots\ , \quad \tfrac{\pi-\phi_m}{\pi}\theta^\top + \tfrac{\sin\phi_m}{\pi}\theta_m^{\circ\top}\right)$$

where $\phi_j = \arccos(\langle\theta, \theta_j^\circ\rangle)$.

We know the matrix $\mathbb{E}_{D_n}\left[\frac{1}{n}X_0X_0^\top\right]$ is a positive definite. Indeed, Safran et al. (2020, Theorem 3.2) shows that

$$\begin{pmatrix} \frac{1}{2}\mathrm{I}_d & \frac{1}{4}\mathrm{I}_d & \cdots & \frac{1}{4}\mathrm{I}_d \\ \frac{1}{4}\mathrm{I}_d & \frac{1}{2}\mathrm{I}_d & \cdots & \frac{1}{4}\mathrm{I}_d \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{4}\mathrm{I}_d & \frac{1}{4}\mathrm{I}_d & \cdots & \frac{1}{2}\mathrm{I}_d \end{pmatrix} + \frac{1}{2\pi}\begin{pmatrix} 0_d & E_{1,2} & \cdots & E_{1,m} \\ E_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & E_{m-1,m} \\ E_{m,1} & \cdots & E_{m,m-1} & 0_d \end{pmatrix} \succeq \left(\frac{1}{4} - \frac{1}{2\pi}\right) I_{md},$$

which leads to

$$\mathbb{E}_{D_n}\left[\frac{1}{n}X_0X_0^\top\right] \succeq \frac{1}{d}\left(\frac{1}{4} - \frac{1}{2\pi}\right) I_{md}. \tag{20}$$

By the straight forward calculation, we can check that

$$\left(\mathbb{E}_{D_n}\left[\frac{1}{n}X_0X_0^\top\right]\right)^{-1}\begin{pmatrix}\theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ\end{pmatrix} = ad\begin{pmatrix}\theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ\end{pmatrix} + bd\begin{pmatrix}\sum_{j=1}^m \theta_j^\circ \\ \sum_{j=1}^m \theta_j^\circ \\ \vdots \\ \sum_{j=1}^m \theta_j^\circ\end{pmatrix}, \tag{21}$$

where a and b satisfy

$$\begin{cases} \left(1 + \frac{m-1}{\pi}\right)a + \left(\frac{m+1}{2} + \frac{m-1}{\pi}\right)b = 1, \\ a + \left(\frac{2}{\pi} + m + 1\right)b = 0. \end{cases} \tag{22}$$

By solving the equation(22), we get the closed form of a, b as

$$a = \frac{2\pi(\pi m + \pi + 2)}{2\pi m^2 + (\pi^2 - 2\pi + 4)m + \pi^2 + 4\pi - 4}, b = -\frac{2\pi^2}{2\pi m^2 + (\pi^2 - 2\pi + 4)m + \pi^2 + 4\pi - 4}.$$

Note that for any integer m, it holds that $a > 0, b < 0$ and $a = -\left(\frac{2}{\pi} + m + 1\right)b$.

By combining Lemma A.9 and Eq. (21), we can write $\bar{f}$ by an explicit form given as

$$\bar{f}(\theta) = (a+b)\sum_{j=1}^m\left(\frac{\pi-\phi_j}{\pi}\cos\phi_j + \frac{\sin\phi_j}{\pi}\right) + b\sum_{j=1}^m\sum_{j'\neq j}\frac{\pi-\phi_j}{\pi}\cos\phi_{j'}.$$

By the construction, it is expected that the function $f^*(p^\dagger)$ converges to $\bar{f}$ with $n \to \infty$. Indeed, we can show that

1. $\bar{f}$ satisfies the conditions of NDSC.
2. $f^*(p^\dagger)$ converges to $\bar{f}$ while satisfying the conditions in NDSC.

At first, we give the first assertion.

Lemma A.10. $\bar{f}(\cdot)$ *satisfies*

$$\begin{cases} \bar{f}(\theta_j^\circ) = 1 & (\forall j \in [m]), \\ 0 < \bar{f}(\theta) < 1 & (\theta \neq \theta_j^\circ \ (\forall j \in [m]),\ \theta \in \mathbb{S}^{d-1}). \end{cases}$$

Proof. Let us consider the induction on m. If $m = 1$, Lemma holds clearly with

$$\bar{f}(\theta) = \frac{\pi - \phi_1}{\pi} \cos \phi_1 + \frac{\sin \phi_1}{\pi}.$$

Below we consider the case $m \geq 2$ and assume that the conclusion holds for $m - 1$. At first, if $\theta = \theta_j^\circ$ for a $j \in [m]$, it holds that

$$\begin{aligned} \bar{f}(\theta) = (a+b)\left(\frac{\pi - 0}{\pi} \cos 0 + \frac{\sin 0}{\pi}\right) &+ (a+b) \sum_{j \neq j} \left(\frac{\pi - \pi/2}{\pi} \cos \pi/2 + \frac{\sin \pi/2}{\pi}\right) \\ &+ b \sum_{j' \neq j} \frac{\pi - \pi/2}{\pi} \cos 0 \\ &= \left(1 + \frac{m-1}{\pi}\right) a + \left(\frac{m+1}{2} + \frac{m-1}{\pi}\right) b = 1, \end{aligned}$$

which gives the first equality. To prove the other case, we consider the expansion

$$\theta = \sum_{j=1}^m k_j \theta_j^\circ + \left(\text{orthogonal term to } \operatorname{span}\{\theta_1^\circ, \theta_2^\circ, \ldots, \theta_m^\circ\}\right).$$

Because of the orthogonality of $(\theta_j^\circ)_{j=1}^m$, for each $\theta \in \mathbb{S}^{d-1}$, $(k_j)_{j=1}^m$ are uniquely determined and satisfy the inequality $\sum_{j=1}^m k_j^2 \leq 1$.

Then, because the orthogonal term does not affect the value of $\bar{f}$, we can write

$$\bar{f}(k_1, \ldots, k_m) = (a+b) \sum_{j=1}^m \left(\frac{\pi - \arccos(k_j)}{\pi} k_j + \frac{\sqrt{1 - k_j^2}}{\pi}\right) + b \sum_{j=1}^m \sum_{j' \neq j} \frac{\pi - \arccos(k_j)}{\pi} k_{j'}.$$

Firstly we show $\bar{f} < 1$. Suppose that there exists $j \in [m]$ such that $k_j = 0$. Without loss of generality, we consider the case $k_m = 0$. Then we have

$$\begin{aligned} &\bar{f}(k_1, \ldots, k_{m-1}, 0) \\ &= (a+b) \sum_{j=1}^{m-1} \left(\frac{\pi - \arccos(k_j)}{\pi} k_j + \frac{\sqrt{1 - k_j^2}}{\pi}\right) \\ &+ (a+b)\frac{1}{\pi} + b \sum_{j=1}^{m-1} \sum_{\substack{j' \neq j \\ 1 \leq j' \leq m-1}} \frac{\pi - \arccos(k_j)}{\pi} k_{j'} + \frac{1}{2} b \sum_{j=1}^{m-1} k_j \\ &\leq -b \left\{ \left(m - 1 + \frac{2}{\pi}\right) \sum_{j=1}^{m-1} \left(\frac{\pi - \arccos(k_j)}{\pi} k_j + \frac{\sqrt{1 - k_j^2}}{\pi}\right) - \sum_{j=1}^{m-1} \sum_{\substack{j' \neq j \\ 1 \leq j' \leq m-1}} \frac{\pi - \arccos(k_j)}{\pi} k_{j'} \right\} \\ &- b \sum_{j=1}^{m-1} \left(\frac{\pi - \arccos(k_j)}{\pi} k_j + \frac{\sqrt{1 - k_j^2}}{\pi}\right) + (a+b)\frac{1}{\pi} + \frac{1}{2} b \sum_{j=1}^{m-1} k_j. \end{aligned}$$

By the induction assumption, the first term takes maximum value only when $k_j = 1$ for some $1 \le j \le m-1$. For the rest term, we have

$$\begin{aligned}
&-b\sum_{j=1}^{m-1}\left(\frac{\pi-\arccos(k_j)}{\pi}k_j+\frac{\sqrt{1-k_j^2}}{\pi}\right)+(a+b)\frac{1}{\pi}+\frac{1}{2}b\sum_{j=1}^{m-1}k_j\\
&\quad=-b\left\{\sum_{j=1}^{m-1}\left(\frac{\pi-\arccos(k_j)}{\pi}k_j+\frac{\sqrt{1-k_j^2}}{\pi}\right)-\left(m+\frac{2}{\pi}\right)\frac{1}{\pi}-\frac{1}{2}\sum_{j=1}^{m-1}k_j\right\}\\
&\quad\le-b\left\{\sum_{j=1}^{m-1}\left(\frac{1}{\pi}+\frac{1}{2}k_j+\left(\frac{1}{2}-\frac{1}{\pi}\right)k_j^2\right)-\left(m+\frac{2}{\pi}\right)\frac{1}{\pi}-\frac{1}{2}\sum_{j=1}^{m-1}k_j\right\}\\
&\quad=-b\left\{-\left(1+\frac{2}{\pi}\right)\frac{1}{\pi}+\left(\frac{1}{2}-\frac{1}{\pi}\right)\right\},
\end{aligned}$$

where we use the inequality

$$\frac{\pi-\arccos(k_j)}{\pi}k_j+\frac{\sqrt{1-k_j^2}}{\pi}\le\frac{1}{\pi}+\frac{1}{2}k+\left(\frac{1}{2}-\frac{1}{\pi}\right)k^2,$$

which holds with equality if $k = 0$ or 1. Therefore $\bar{f}$ takes maximum value at $\theta = \theta_j^\circ$ for some $1 \le j \le m$, which gives the conclusion. Then we consider the case where $k_j \neq 0$ for all $j \in [m]$. We only need to consider a case $\sum_{j=1}^m k_j \ge 0$. Indeed, it holds that

$$\begin{aligned}
&\bar{f}(k_1,\dots,k_m)-\bar{f}(-k_1,\dots,-k_m)\\
&=(a+b)\sum_{j=1}^{m}\left(\frac{\pi-\arccos(k_j)}{\pi}k_j+\frac{\sqrt{1-k_j^2}}{\pi}-\frac{\arccos(k_j)}{\pi}(-k_j)-\frac{\sqrt{1-k_j^2}}{\pi}\right)\\
&+b\sum_{j=1}^{m}\sum_{j'\neq j}\left(\frac{\pi-\arccos(k_j)}{\pi}k_{j'}-\frac{\arccos(k_j)}{\pi}(-k_{j'})\right)\\
&=(a+b)\sum_{j=1}^{m}k_j+b\sum_{j=1}^{m}\sum_{j'\neq j}k_j=-b\left(1+\frac{2}{\pi}\right)\sum_{j=1}^{m}k_j>0.
\end{aligned}$$

Now we consider the conversion $(k_1,\dots,k_{j_1},\dots,k_{j_2},\dots,k_m)\mapsto(k_1,\dots,\sqrt{k_{j_1}^2+k_{j_1}^2},0,\dots,k_m)$ for some $j_1 \neq j_2$. For the notation simplicity, we consider that $j_1 = 1, j_2 = 2$. Since $\bar{f}$ is permutation-invariant, this does not lose the generality. Let $r := \sqrt{k_1^2+k_2^2} > 0$, then

$$\begin{aligned}
&\bar{f}(k_1,\dots,k_{j_1},\dots,k_{j_2},\dots,k_m)-\bar{f}(k_1,\dots,\sqrt{k_{j_1}^2+k_{j_1}^2},0,\dots,k_m)\\
&=(a+b)\left(\frac{\pi-\arccos(k_1)}{\pi}k_1+\frac{\sqrt{1-k_1^2}}{\pi}+\frac{\pi-\arccos(k_2)}{\pi}k_2+\frac{\sqrt{1-k_2^2}}{\pi}\right.\\
&\left.-\frac{\pi-\arccos(r)}{\pi}r+\frac{\sqrt{1-r^2}}{\pi}-\frac{1}{\pi}\right)+b\left(\frac{\pi-\arccos(k_1)}{\pi}k_2+\frac{\pi-\arccos(k_2)}{\pi}k_1\right)\\
&-b\frac{1}{2}k_r+b\left(\sum_{j=3}^{m}\frac{\pi-\arccos(k_j)}{\pi}\right)(k_1+k_2-r)\\
&+b\sum_{j=3}^{m}k_j\frac{-\arccos(k_1)-\arccos(k_2)+\arccos(r)+\pi/2}{\pi}.
\end{aligned}$$

By using Lemma C.5, this value is upper bounded by

$$
\begin{aligned}
&(a+b)\frac{1}{2}(k_1+k_2-r)\\
&+b\left(\frac{\pi-\arccos(k_1)}{\pi}k_2+\frac{\pi-\arccos(k_2)}{\pi}k_1\right)-b\frac{1}{2}r+b\left(\sum_{j=3}^{m}\frac{\pi-\arccos(k_j)}{\pi}\right)(k_1+k_2-r)\\
&+b\sum_{j=3}^{m}k_j\frac{-\arccos(k_1)-\arccos(k_2)+\arccos(r)+\pi/2}{\pi}\\
=&-b\left\{\left(m+\frac{2}{\pi}\right)\frac{1}{2}(k_1+k_2-r)\right.\\
&-\left(\frac{\pi-\arccos(k_1)}{\pi}k_2+\frac{\pi-\arccos(k_2)}{\pi}k_1\right)+\frac{1}{2}r-\left(\sum_{j=3}^{m}\frac{\pi-\arccos(k_j)}{\pi}\right)(k_1+k_2-r)\\
&\left.-\sum_{j=3}^{m}k_j\frac{-\arccos(k_1)-\arccos(k_2)+\arccos(r)+\pi/2}{\pi}\right\}.\\
=&-b\left\{\left[m+\frac{2}{\pi}-\left(\sum_{j=3}^{m}\frac{\pi-\arccos(k_j)}{\pi}\right)-\frac{1}{2}\right]\frac{1}{2}(k_1+k_2-r)\right.\\
&-\left(\frac{\pi/2-\arccos(k_1)}{\pi}k_2+\frac{\pi/2-\arccos(k_2)}{\pi}k_1\right)\\
&\left.-\sum_{j=3}^{m}k_j\frac{-\arccos(k_1)-\arccos(k_2)+\arccos(r)+\pi/2}{\pi}\right\}.
\end{aligned}
\tag{23}
$$

In the latter we ignore the multiplied constant $-b>0$. Then we consider the two cases: (i) we can take $k_1>0$,$k_2<0$ (ii) $k_j>0$ for all $j\in[m]$. Note that since $\sum_{j=1}^{m}k_j\geq 0$, there must be a integer $j\in[m]$ such that $k_j>0$. Firstly we consider the case (i). In this case it holds that $\sum_{j=3}^{m}k_j\geq -1$.

Firstly, we consider to evaluate the term

$$
\frac{1}{2}\left(m+\frac{2}{\pi}\right)-\left(\sum_{j=3}^{m}\frac{\pi-\arccos(k_j)}{\pi}\right)-\frac{1}{2}
$$

Firstly, we have an inequality for $k\geq 0$,

$$
\frac{1}{2}+\frac{1}{\pi}k\leq\frac{\pi-\arccos(k)}{\pi}\leq\frac{1}{2}+\frac{1}{\pi}k+\left(\frac{1}{2}-\frac{1}{\pi}\right)k^2
$$

and for for $k\leq 0$,

$$
\frac{1}{2}+\frac{1}{\pi}k-\left(\frac{1}{2}-\frac{1}{\pi}\right)k^2\leq\frac{\pi-\arccos(k)}{\pi}\leq\frac{1}{2}-\frac{1}{\pi}k,
\tag{24}
$$

which gives

$$
\frac{m-2}{2}+\frac{1}{\pi}\sum_{j=3}^{m}k_j-\left(\frac{1}{2}-\frac{1}{\pi}\right)\leq\sum_{j=3}^{m}\frac{\pi-\arccos(k)}{\pi}\leq\frac{m-2}{2}+\frac{1}{\pi}\sum_{j=3}^{m}k_j+\left(\frac{1}{2}-\frac{1}{\pi}\right).
$$

Then it follows that

$$
\frac{2}{\pi}-\frac{1}{\pi}\sum_{j=3}^{m}k_j\leq\frac{1}{2}\left(m+\frac{2}{\pi}\right)-\left(\sum_{j=3}^{m}\frac{\pi-\arccos(k_j)}{\pi}\right)-\frac{1}{2}
$$

and

$$\frac{1}{2}\left(m+\frac{2}{\pi}\right)-\left(\sum_{j=3}^{m}\frac{\pi-\arccos(k_j)}{\pi}\right)-\frac{1}{2}\leq 1-\frac{1}{\pi}\sum_{j=3}^{m}k_j.$$

Then by using the inequality $k_1+k_2-r<0$ and $-\arccos(k_1)-\arccos(k_2)+\arccos(r)+\pi/2<0$, we can get the inequality (23) ≤ 0. For the case (ii), we consider the case $k_1\leq k_2\leq\cdots\leq k_m$ Then using Lemma C.7, we can show that

$$\begin{aligned}(23)\leq&\left\{\frac{1}{2}\left(m+\frac{2}{\pi}\right)-\left(\sum_{j=3}^{m}\frac{\pi-\arccos(k_j)}{\pi}\right)-\frac{1}{2}\right\}(k_1+k_2-r)\\&-\left(\frac{1}{2}-\frac{\arccos(k_1)}{\pi}\right)k_2-\left(\frac{1}{2}-\frac{\arccos(k_2)}{\pi}\right)k_1\\&\leq 0.\end{aligned}$$

Thus we only need to treat the case where there exists j such that $k_j=0$, which gives the conclusion. □

Then we give the proof to the closeness of $f^*(p^\dagger)$ and $\bar{f}$. We show this in two perspective, global and local. In global we show that $\|f^*(p^\dagger)-\bar{f}\|_\infty$ will be small with sufficiently large n. We need another explanation to local, where θ close to θ_j for $j\in[m]$, because only the global discussion, there may be the point where f^* takes the value larger than 1. Firstly we give the global result.

Lemma A.11 (Global concentration). *Under the Assumption 3.1, there exists a constant C independent with m,n,d, for any $0<\delta<1$, with probability more than $1-\delta$, it holds*

$$\sup_\theta|f^*(p^\dagger)(\theta)-\bar{f}(\theta)|\leq C\left(md\sqrt{\frac{d\log n}{n}}+md\sqrt{\frac{md\log 1/\delta}{n}}\right).$$

Remark that $\bar{f}$ is defined as the expected function of $f^*(p^\dagger)$, in the sense that we take expected value of two matrices, $\frac{1}{n}X(\theta)X_0^\top$ and $\frac{1}{n}X_0X_0^\top$. Therefore we aim to concentration inequalities for these respectively.

Lemma A.12 (The concentration of $\frac{1}{n}X_0X_0^\top$). *We assume the Assumption 3.1 holds. Let $\hat{K}_0=\frac{1}{n}X_0X_0^\top$, then for any $0<t<1/12d$, we have*

$$\Pr\left(\left\|\hat{K}_0-\mathbb{E}_{D_n}\left[\hat{K}_0\right]\right\|_{\mathrm{op}}\geq t\right)\leq 2md\exp\left(-\frac{nt^2}{4(2m^2+2mt/3)}\right), \tag{25}$$

$$\Pr\left(\left\|\hat{K}_0^{-1}-\mathbb{E}_{D_n}\left[\hat{K}_0\right]^{-1}\right\|_{\mathrm{op}}\geq t\right)\leq 2md\exp\left(-\frac{nt^2}{600(300d^4m^2+2d^2mt/3)}\right). \tag{26}$$

Proof. Note that $\hat{K}_0$ is decomposed as

$$\hat{K}_0=\frac{1}{n}\sum_{i=1}^{n}A_i:=\frac{1}{n}\sum_{i=1}^{n}\left(x_ix_i^\top\mathbb{1}\{\langle\theta_{j_1},x_i\rangle\geq 0\cap\langle\theta_{j_2},x_i\rangle\geq 0\}\right)_{\substack{(j_1-1)d\leq i\leq j_1d\\(j_2-1)d\leq j\leq j_2d}}.$$

For each component $A_i\in\mathbb{R}^{md\times md}$, it holds that $\mathbb{E}_{D_n}[\frac{1}{n}(A_i-\mathbb{E}_{D_n}[A_i])]=0_d$ and

$$\left|\frac{1}{n}(A_i-\mathbb{E}_{D_n}[A_i])\right|\leq\frac{2}{n}m,$$

which is obtained by

$$\begin{aligned}\left\|\frac{1}{n}(A_i-\mathbb{E}_{D_n}[A_i])\right\|_{\mathrm{op}}&\leq\left\|\frac{1}{n}A_i\right\|_{\mathrm{op}}+\left\|\frac{1}{n}\mathbb{E}_{D_n}[A_i]\right\|_{\mathrm{op}}\\&\leq\left\|\frac{1}{n}A_i\right\|_{\mathrm{F}}+\frac{1}{n}\mathbb{E}_{D_n}[\|A_i\|_{\mathrm{F}}]\leq\frac{2}{n}m,\end{aligned}$$

where we use Jensen's inequality,$\|A\|_{\mathrm{op}} \leq \|A\|_{\mathrm{F}}$ for a matrix A and $\|x_i\| = 1$. Therefore we can apply Lemma C.8 with $X_i = \frac{1}{n}(A_i - \mathbb{E}_{D_n}[A_i])$. As a consequence, it holds that for any $t > 0$,

$$\Pr\left(\|\hat{K}_0 - \mathbb{E}_{D_n}[\hat{K}_0]\|_{\mathrm{op}} \geq t\right) \leq 2md \exp\left(-\frac{nt^2}{4(2m^2 + 2mt/3)}\right).$$

This gives Eq. (25). Next we consider Eq. (26). Since it holds that $E_{D_n}[\hat{K}_0] \succeq (1/4 - 1/2\pi)1/dI_{md}$ by Eq. (20), if $\left\|\hat{K}_0 - \mathbb{E}_{D_n}\left[\hat{K}_0\right]\right\|_2 \leq t$ holds, we have $\hat{K}_0 \succeq ((1/4 - 1/2\pi)1/d - t)I_{md}$, which gives

$$\begin{aligned}\left\|\hat{K}_0^{-1} - \mathbb{E}_{D_n}[\hat{K}_0]^{-1}\right\|_{\mathrm{op}} \leq \left\|\hat{K}_0\right\|_{\mathrm{op}}^{-1} \left\|\hat{K}_0 - \mathbb{E}_{D_n}[\hat{K}_0]\right\|_{\mathrm{op}} \left\|\mathbb{E}_{D_n}[\hat{K}_0]\right\|_{\mathrm{op}}^{-1} &\leq \frac{d^2 t}{(1/4 - 1/2\pi)(1/4 - 1/2\pi - dt)} \\ &\leq \frac{144 d^2 t}{1 - 12dt},\end{aligned}$$

where we use $1/6 > 1/2\pi$. This leads to

$$\Pr\left(\left\|\hat{K_0}^{-1} - \mathbb{E}_{D_n}\left[\hat{K_0}\right]^{-1}\right\|_{\mathrm{op}} \geq \frac{144d^2 t}{1 - 12dt}\right) \leq 2md \exp\left(-\frac{nt^2}{4(2m^2 + 2mt/3)}\right).$$

By replacing t by $\frac{t}{144d^2 + 12dt} \leq \frac{t}{150d^2}$, we get the conclusion. □

Lemma A.13 (The concentration of $\frac{1}{n}X(\theta)X_0^\top$). *Let $\hat{K}(\theta) = \frac{1}{n}X(\theta)X_0^\top \in \mathbb{R}^{d\times md}$. Then there exists a constant $C > 0$ independent of n and d, for any $\delta > 0$, with probability at least $1 - \delta$, it holds that*

$$\sup_\theta \left\|\hat{K}(\theta) - \mathbb{E}_{D_n}\left[\hat{K}(\theta)\right]\right\|_{\mathrm{op}} \leq C\sqrt{\frac{md\log n}{n}} + 2\sqrt{\frac{m\log(1/\delta)}{2n}}.$$

Proof. At first, we remark that

$$\begin{aligned}\sup_\theta \left\|\hat{K}(\theta) - \mathbb{E}_{D_n}\left[\hat{K}(\theta)\right]\right\|_{\mathrm{op}} &= \sup_{\substack{\theta, \|w\|\leq 1, \|v\|\leq 1 \\ w\in\mathbb{R}^{md}, v\in\mathbb{R}^d}} \left\langle v, \left(\hat{K}(\theta) - \mathbb{E}_{D_n}\left[\hat{K}(\theta)\right]\right) w\right\rangle \\ &= \sup_{\substack{\theta, \|w\|\leq 1, \|v\|\leq 1 \\ w\in\mathbb{R}^{md}, v\in\mathbb{R}^d}} \left\{\left\langle v, \hat{K}(\theta)w\right\rangle - \mathbb{E}_{D_n}\left[\left\langle v, \hat{K}(\theta)w\right\rangle\right]\right\}.\end{aligned}$$

In the above equation, it holds that

$$\left\langle v, \hat{K}(\theta)w\right\rangle = \frac{1}{n}\sum_{i=1}^n \langle v, x_i\rangle\langle w, \tilde{x}_i\rangle \mathbb{1}\{\langle\theta, x_i\rangle \geq 0\},$$

where $\tilde{x}_i := (x_i^\top \mathbb{1}\{\langle\theta_1^*, x_i\rangle \geq 0\}, \ldots, x_i^\top \mathbb{1}\{\langle\theta_m^*, x_i\rangle \geq 0\}) \in \mathbb{R}^{md}$. Therefore we consider to bound the Rademacher complexity of

$$\mathcal{F} := \left\{r(\theta, v, w) = \frac{1}{n}\sum_{i=1}^n \langle v, x_i\rangle\langle w, \tilde{x}_i\rangle \mathbb{1}\{\langle\theta, x_i\rangle \geq 0\} \mid (\theta, v, w) \in \mathbb{S}^{d-1}\times\mathbb{R}^d\times\mathbb{R}^{md}, \|v\|\leq 1, \|w\|\leq 1\right\}.$$

Let $\|f\|_n = \sqrt{\frac{1}{n} f^2(x_i)}$. For two pairs (θ, v, w), $(\theta', v', w') \in (\mathbb{S}^{d-1} \times \mathbb{R}^d \times \mathbb{R}^{md})^2$, we have

$$
\begin{aligned}
\|r(\theta, v, w) - r(\theta', v', w')\|_n^2 &= \frac{1}{n}\sum_{i=1}^n \left(\langle v, x_i\rangle\langle w, \tilde{x}_i\rangle \mathbb{1}\{\langle \theta, x_i\rangle \geq 0\} - \langle v', x_i\rangle\langle w', \tilde{x}_i\rangle \mathbb{1}\{\langle \theta', x_i\rangle \geq 0\}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Bigg(\langle v - v', x_i\rangle\langle w, \tilde{x}_i\rangle \mathbb{1}\{\langle \theta, x_i\rangle \geq 0\} \\
&\quad + \langle v', x_i\rangle\langle w - w', \tilde{x}_i\rangle \mathbb{1}\{\langle \theta, x_i\rangle \geq 0\} \\
&\quad + \langle v', x_i\rangle\langle w - w', \tilde{x}_i\rangle (\mathbb{1}\{\langle \theta, x_i\rangle \geq 0\} - \mathbb{1}\{\langle \theta', x_i\rangle \geq 0\}) \Bigg)^2 \\
&\leq 3\left(\|v - v'\|^2 m + \|w - w'\|^2 m + \left|\mathbb{1}\{\langle \theta', x_i\rangle \geq 0\} - \mathbb{1}\{\langle \theta, x_i\rangle \geq 0\}\right|^2 m \right),
\end{aligned}
$$

where we use $\|x_i\| = 1$ and $\|\tilde{x}_i\| \leq \sqrt{m}$ for any $i \in [n]$, and $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$. By this inequality, we get an upper bound of the covering number of $\mathcal{F}$ as

$$
\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_n) \leq \begin{cases} C_0 n^{d+1} \left(\frac{1}{\epsilon}\sqrt{\frac{1}{m}}\right)^{md+d} & \epsilon < \frac{3}{\sqrt{m}}, \\ 1 & otherwise, \end{cases}
$$

for a some constant $C_0 > 0$ which is independent of the other parameters. Note that the term n^{d+1} is derived by the covering over $\theta \in \mathbb{S}^{d-1}$ and the term $\frac{1}{\epsilon}\left(\sqrt{\frac{n}{m}}\right)^{md+d}$ is derived by the covering over $(v, w) \in \mathbb{R}^d \times \mathbb{R}^{md}$. Therefore by using Dudley integral argument, we get an upper bound of Rademacher complexity as

$$
R_n(\mathcal{F}) \leq \frac{c}{\sqrt{n}} \int_0^{\frac{3}{\sqrt{m}}} \sqrt{(d+1)\log n + \frac{md+d}{2}\log\frac{1}{m} - (md+d)\log\epsilon + \log C_0 \mathrm{d}\epsilon} \leq C\sqrt{\frac{md\log n}{n}} \tag{27}
$$

for a some constant $C > 0$. Finally by using the standard Rademacher complexity bound, we get the conclusion. □

Combining these concentration inequalities, we give the proof to global concentration.

Proof of Lemma A.11. We consider the decomposition as

$$
\begin{aligned}
\sup_\theta |f^*(p^\dagger)(\theta) - \bar{f}(\theta)| &= \sup_\theta \Bigg| \theta^\mathsf{T} \hat{K}(\theta)\left(\hat{K_0}^{-1} - \mathbb{E}_{D_n}\left[\hat{K_0}\right]^{-1}\right)\begin{pmatrix}\theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ\end{pmatrix} \\
&\quad + \theta^\mathsf{T}\left(\hat{K}(\theta) - \mathbb{E}[\hat{K}(\theta)]\right)\mathbb{E}_{D_n}\left[\hat{K_0}\right]^{-1}\begin{pmatrix}\theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ\end{pmatrix} \Bigg| \\
&\leq \left\|\hat{K_0}^{-1} - \mathbb{E}_{D_n}\left[\hat{K_0}\right]^{-1}\right\|_{\mathrm{op}} m + \sup_\theta \left\|\hat{K}(\theta) - \mathbb{E}_{D_n}[\hat{K}(\theta)]\right\|_{\mathrm{op}} \sqrt{m}\left(\frac{1}{4} - \frac{1}{2\pi}\right)^{-1} d,
\end{aligned}
$$

By using Eq. (26) in Lemma A.12 with $t = C\sqrt{\frac{md^2\log 1/\delta}{n}}$ and Lemma A.13, we get the conclusion. □

Next we show the local evaluation. More precisely, we show that it holds that $f^*(p^\dagger)(\theta) \leq 1$ around θ_j° and equality holds only at $\theta = \theta_j^\circ$.

Lemma A.14 (Local evaluation)**.** *Let $C > 0$ be a constant and $n > \mathrm{poly}(m, d, \log 1/\delta)$. Then with probability at least $1-\delta$, for all $j \in [m]$, if $\mathrm{dist}(\theta, \theta_j) < Cn^{-1/4}$ and $\theta \neq \theta_j^\circ$ it holds that $\bar{f}(\theta) < 1$.*

To prove Lemma A.14, we focus on the gradient $\nabla f^*(p^\dagger)(\theta)$ and utilize an equality $\langle \theta, \nabla f^*(p^\dagger)(\theta)\rangle = f^*(p^\dagger)(\theta)$, which is derived by the 1-homogeneity of ReLU. For simplicity, we p as $p\dagger$ in the latter of this section. At first, we see that p is given by the form

$$p_i = \left(x_i^\top \mathbb{1}\{\langle \theta_1^\circ, x_i\rangle \geq 0\}, \dots, x_i^\top \mathbb{1}\{\langle \theta_m^\circ, x_i\rangle \geq 0\}\right) \left(\frac{1}{n} X_0 X_0^\top\right)^{-1} \begin{pmatrix} \theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ \end{pmatrix},$$

if the matrix $\frac{1}{n} X_0 X_0^\top$ is invertible, where p_i denotes the i's component of p. As a preliminary, we consider the "expected value" of p as

$$q_i = \left(x_i^\top \mathbb{1}\{\langle \theta_1^\circ, x_i\rangle \geq 0\}, \dots, x_i^\top \mathbb{1}\{\langle \theta_m^\circ, x_i\rangle \geq 0\}\right) \left(\mathbb{E}_{D_n}\left[\frac{1}{n} X_0 X_0^\top\right]\right)^{-1} \begin{pmatrix} \theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ \end{pmatrix}.$$

Lemma A.15. *For any $i \in [n]$, it holds that $0 \leq q_i \leq (a - bm)d\sqrt{m}$.*

Proof. Using Eq. (21), q_i is expressed by

$$\begin{aligned} q_i &= \left(x_i^\top \mathbb{1}\{\langle \theta_1^\circ, x_i\rangle \geq 0\}, \dots, x_i^\top \mathbb{1}\{\langle \theta_m^\circ, x_i\rangle \geq 0\}\right) \left(ad \begin{pmatrix} \theta_1^\circ \\ \theta_2^\circ \\ \vdots \\ \theta_m^\circ \end{pmatrix} + bd \begin{pmatrix} \sum_{j=1}^m \theta_j^\circ \\ \sum_{j=1}^m \theta_j^\circ \\ \cdots \\ \sum_{j=1}^m \theta_j^\circ \end{pmatrix} \right) \\ &= ad \sum_{j=1}^m \sigma(\langle \theta_j^\circ, x_i\rangle) + bd \sum_{j_1=1}^m \sum_{j_2=1}^m \langle \theta_{j_1}^\circ, x_i\rangle \mathbb{1}\{\langle \theta_{j_2}^\circ, x_i\rangle \geq 0\}. \end{aligned}$$

Then the upper bound is obtained clearly. For the lower bound, we have that

$$\begin{aligned} q_i &\geq ad \sum_{j=1}^m \sigma(\langle \theta_j, x_i\rangle) + bd \sum_{j_1=1}^m \sum_{j_2=1}^m \sigma(\langle \theta_{j_1}, x_i\rangle) \\ &= d \sum_{j=1}^m \Big(a\sigma(\langle \theta_j, x_i\rangle) + mb\sigma(\langle \theta_j, x_i\rangle)\Big) \\ &= -\left(\frac{2}{\pi} + 1\right) b \sum_{j=1}^m \sigma(\langle \theta_j, x_i\rangle) \geq 0. \end{aligned}$$

In the last inequality we use fact that $b < 0$ and $a + mb = -\left(\frac{2}{\pi} + 1\right) b$. ∎

Next we give a bound on the distance between p_i and q_i, which can be evaluated trough the concentration inequality of $\frac{1}{n} X_0 X_0^\top$ (Lemma A.12).

Lemma A.16. *On the distance between p_i and q_i, we have the following inequality:*

$$\Pr\left(\max_{i\in[n]}\{|p_i - q_i|\} \geq t\right) \leq 2md \exp\left(-\frac{nt^2}{600(300d^4m^2 + 2d^2mt/3)}\right).$$

Proof. By Lemma 26, it holds that

$$\Pr\left(\left\|\hat{K_0}^{-1} - \mathbb{E}_{D_n}\left[\hat{K_0}\right]^{-1}\right\|_{\mathrm{op}} \geq t\right) \leq 2md \exp\left(-\frac{nt^2}{600(300d^4m^3 + 2d^2m^2t/3)}\right).$$

Since $\left\|\begin{pmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_m\end{pmatrix}\right\| = \sqrt{m}$ and $\left\|\begin{pmatrix}x_i\mathbb{1}\{\langle\theta_1^\circ, x_i\rangle \geq 0\}\\ x_i\mathbb{1}\{\langle\theta_2^\circ, x_i\rangle \geq 0\}\\ \vdots\\ x_i\mathbb{1}\{\langle\theta_m^\circ, x_i\rangle \geq 0\}\end{pmatrix}\right\| \leq \sqrt{m}$, we get

$$\Pr\left(\max_{i\in[n]}\{|p_i - q_i|\} \geq mt\right) \leq 2md\exp\left(-\frac{nt^2}{600(300d^4m^2 + 2d^2mt/3)}\right).$$

By replacing t by t/m, we get the conclusion. □

We prepare another Lemma, which is needed to evaluate variation of the gradient $\nabla f^*(p)(\theta)$ around each θ_j°.

Lemma A.17. *Assume the Assumption 3.1 holds. For any $\theta \in \mathbb{S}^{d-1}$ and $\tau > 0$, let $A_\tau := \{x_i \mid \left|\mathrm{dist}(x,\theta) - \frac{\pi}{2}\right| < \tau\}$, then for $0 \leq t \leq 1$, it holds that*

$$\Pr\left(\frac{\#A_\tau}{n} \geq t + \frac{d\tau}{\sqrt{\pi}}\right) \leq \exp\left(-\frac{nt^2}{2(d\tau/\sqrt{\pi} + t/3)}\right).$$

Proof. For any given $\theta \in \mathbb{S}^{d-1}$, Cai et al. (2013, Lemma 12) shows that for each i, $\mathrm{dist}(\theta_j^\circ, x_i)$ is distributed on $[0,\pi]$ with density

$$h(\varphi) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}(\sin\varphi)^{d-2}.$$

This has a maximum value $h(\pi/2) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}$. This leads to

$$\Pr\left(\left|\mathrm{dist}(\theta_j^\circ, x_i) - \frac{\pi}{2}\right| \leq \tau\right) \leq \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}2\tau \leq \frac{d\tau}{\sqrt{\pi}}$$

for any $t \in [0, \frac{\pi}{2}]$. This gives that $\#A_\tau \sim B(n, \mathsf{prob})$ with $\mathsf{prob} \leq \frac{d\tau}{\sqrt{\pi}}$, where $B(\cdot,\cdot)$ denotes the Binomial distribution. Then by using Bernstein's inequality, we get the conclusion. □

Combining Lemma A.15–A.17, we give a proof of Lemma A.14.

Proof of Lemma A.14. At first, we have

$$\theta_j^\circ = \frac{1}{n}\sum_{i:\langle\theta_j, x_i\rangle \geq 0} p_i x_i.$$

At the place θ, let I_1(resp. I_2) be the subset of $[n]$ such that $\langle\theta_j^\circ, x_i\rangle \geq 0$ and $\langle\theta, x_i\rangle < 0$ (resp. $\langle\theta_j^\circ, x_i\rangle < 0$ and $\langle\theta, x_i\rangle \geq 0$), the gradient at θ is expressed as

$$\begin{aligned}
\langle\theta, \nabla f^*(p)\rangle &= \left\langle\theta, \theta_j - \frac{1}{n}\sum_{I_1} p_i x_i + \frac{1}{n}\sum_{I_2} p_i x_i\right\rangle \\
&= \langle\theta, \theta_j\rangle - \frac{1}{n}\sum_{I_1} p_i\langle\theta, x_i\rangle + \frac{1}{n}\sum_{I_2} p_i\langle\theta, x_i\rangle \\
&= \langle\theta, \theta_j\rangle - \frac{1}{n}\sum_{I_1} q_i\langle\theta, x_i\rangle + \frac{1}{n}\sum_{I_2} q_i\langle\theta, x_i\rangle \\
&\quad - \frac{1}{n}\sum_{I_1}(p_i - q_i)\langle\theta, x_i\rangle + \frac{1}{n}\sum_{I_2}(p_i - q_i)\langle\theta, x_i\rangle
\end{aligned}.$$

Let $\langle\theta, \theta_j^\circ\rangle := 1 - T$ for $T > 0$, then we have $\|\theta - \theta_j^\circ\| \leq T$ and it holds that

- For $i \in I_1$, $-T \le \langle \theta, x_i \rangle < 0$,
- For $i \in I_2$, $0 \le \langle \theta, x_i \rangle \le T$.

Then by using the fact $q_i \ge 0$ and Lemma A.16, we get

$$\begin{aligned}\langle \theta, \nabla f^*(p)(\theta) \rangle &\le 1 - T + T\frac{\max q_i}{n}\#\{I_1 \cup I_2\} \qquad (28)\\ &\quad + \left| \frac{1}{n}\sum_{I_1}(p_i - q_i)\langle \theta, x_i \rangle \right| + \left| \frac{1}{n}\sum_{I_2}(p_i - q_i)\langle \theta, x_i \rangle \right| \\ &\le 1 - T\left\{1 - \left(\tfrac{\max q_i}{n} + \tfrac{t}{n}\right)\#\{I_1 \cup I_2\}\right\} \qquad (29)\end{aligned}$$

with probability at least RHS of Lemma A.16. It remains to show that $1 - \left(\frac{\max q_i}{n} + \frac{t}{n}\right)\#\{I_1 \cup I_2\} > 0$ while $T \le C/n^{1/4}$. By the definition, $\#\{I_1 \cup I_2\}$ is upper bounded by $\#A_T$ associated with θ_j°. Thus we can show that $\left(\frac{\max q_i}{n} + \frac{t}{n}\right)\#\{I_1 \cup I_2\} = \mathrm{O}(T)$ w.h.p. under $n > poly(m, d, \log 1/\delta)$ and this gives the conclusion. □

A.3. Proof of Theorem 3.5

Combining the discussion in the previous section, we give the proof of Theorem 3.5. At first, we show that NDSC holds in the teacher student setting w.h.p. (Proposition A.7).

proof of Proposition A.7. At first, by Eq. (25) in Lemma A.12, $\frac{1}{n}X_0X_0^\top$ is positive definite with probability at least $1 - Cm\sqrt{\log(md)/n}$ for a constant $C > 0$. Suppose that this holds, $p^\dagger$ exists and is written by Eq. (18). In this case, $f^*(p^\dagger)(\theta_j^\circ) = 1$ holds clearly by the construction for any $j \in [m]$.

Next we show the concavity around θ_j° for each j. Note that $\nabla f^*(p^\dagger)(\theta_j^\circ) = \theta_j^*$. Therefore it holds that $\nabla f^*(p^\dagger)(\theta) = \theta_j^*$ for θ sufficiently close to θ_j° to satisfy $\mathrm{sgn}(\langle \theta, x_i \rangle) = \mathrm{sgn}(\langle \theta_j^\circ, x_i \rangle)$ for all $i \in [n]$, since

$$\nabla f^*(p^\dagger)(\theta) = \frac{1}{n}\sum_{i:\langle \theta, x_i \rangle \ge 0} p^\dagger x_i = \frac{1}{n}\sum_{i:\langle \theta_j^\circ, x_i \rangle \ge 0} p_i^\dagger x_i = \nabla f^*(p_i^\dagger)(\theta_j^\circ).$$

Hence, it holds that $\langle \theta, \nabla f^*(p^\dagger)(\theta) \rangle = \langle \theta, \theta_j^\circ \rangle$ around θ_j° and this is clearly a concave. Finally we show the second condition, i.e., $|f^*(p^\dagger)(\theta)| < 1$ for any $\theta \ne \theta_j^\circ (\forall j \in [m])$. By Lemma A.11, we know for sufficiently large n, if there exists a point where $|f^*(p^\dagger)(\theta)| \ge 1$, it must be around θ_j°. Moreover, by Lemma A.14, we can ensure that there must be no point other than θ_j° until the function value decreases to $1 - \mathrm{O}(T) = 1 - \mathrm{O}(n^{-1/4})$. Combining these results, we get the conclusion. □

proof of Proposition A.1. Assume that the NDSC holds, which is ensured by Proposition A.7. Let p_λ be the unique solution of (P_λ). By Lemma A.4, for sufficiently small $\lambda > 0$, it holds that $f^*(p_\lambda)(\theta)$ only takes value 1 at $\theta = \theta_j^*$ ($j \in [m]$) which satisfies $\mathrm{sgn}(\langle \theta_j^*, x_i \rangle) = \mathrm{sgn}(\langle \theta_j^\circ, x_i \rangle)$ for all $i \in [n]$. Moreover, since $\|\theta_j^\circ - \theta_j^*\|$ can be arbitrary small as $\lambda \to +0$, we get the first conclusion. To complete the proof, we discuss $(r_j^*)_{j=1}^m$.

Firstly, we show that $(r_j^*)_{j=1}^m$ are uniquely determined. By the optimality condition (14), it holds that

$$(p_\lambda)_i = -\frac{1}{\lambda}(f(x_i;\nu^*) - y_i) \quad (\forall i \in [n]).$$

Remind that p_λ is uniquely determined. Let $\nu^* = \sum_{j=1}^m r_j^*\delta_{\theta_j^*}$ and rearranging this equation, we have

$$\sum_{j=1}^m r_j^*\sigma(\langle \theta_j^*, x_i \rangle) = -\lambda(p_\lambda)_i + y_i \quad (\forall i \in [n]). \qquad (30)$$

This can be seen as a linear equation about $r^* := (r_1^*, \ldots, r_m^*)^\top$, that is,

$$Ar^* = -\lambda p_\lambda + y,$$

where $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ and $A = (\sigma(\langle \theta_j^*, x_i \rangle))_{i,j} \in \mathbb{R}^{n \times m}$. We can show that $n \geq poly(m, d, \log 1/\delta)$ and sufficiently small λ, A has column full rank with probability at least $1 - \delta$. Indeed, A can be decomposed as

$$A = X_0^\top \begin{pmatrix} \theta_1^* & 0 & \cdots & 0 \\ 0 & \theta_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \theta_m^* \end{pmatrix},$$

where the second matrix has column full rank and we have already shown that X_0 has column full rank w.h.p.. Consequently, we can show the uniqueness of $(r_j^*)_{j=1}^m$. Moreover, taking the limit $\lambda \to +0$ in Eq. (30), we have $\sum_{j=1}^m r_j^* \sigma(\langle \theta_j^*, x_i \rangle) \to y_i = \sum_{j=1}^m r_j^\circ \sigma(\langle \theta_j^\circ, x_i \rangle)$ $(i \in [n])$. Then, by using $\theta_j^* \to \theta_j^\circ$ and linear independent of $(\sigma(\langle \theta_j^\circ, \cdot \rangle))_{j=1}^m$ which holds w.h.p., we get $(r_j^*)_{j=1}^m \to (r_j^\circ)_{j=1}^m$ as $\lambda \to +0$. This gives the conclusion. □

To complete the proof, we need to evaluate how close $(r_j^*, \theta_j^*)_{j=1}^m$ and teacher parameters $(r_j^\circ, \theta_j^\circ)_{j=1}^m$ will be. This quantitative evaluation is obtained by using the form $\nu^* = \sum_{j=1}^m r_j^* \delta_{\theta_j^*}$ and strong convexity of the empirical risk, as we see in the proof below.

proof of Theorem 3.5. For sufficiently large n and small $\lambda > 0$, we can assume that the optimal solution is written by a form $\nu^* = \sum_{j=1}^m r_j^* \delta_{\theta_j^*}$, as we have shown in Proposition A.1. Then, by the optimality of ν^*, it holds that

$$\frac{1}{2n} \sum_{i=1}^n (f(x_i; \nu^*) - f(x_i; \nu^\circ))^2 + \lambda \sum_{j=1}^m |r_j^*| \leq \frac{1}{2n} \sum_{i=1}^n (f(x_i; \nu^\circ) - f(x_i; \nu^\circ))^2 + \lambda \sum_{j=1}^m |r_j^\circ|.$$

This yields that

$$\frac{1}{2n} \sum_{i=1}^n (f(x_i; \nu^*) - f(x_i; \nu^\circ))^2 \leq \lambda \sum_{j=1}^m (|r_j^\circ| - |r_j^*|) \leq \lambda \sum_{j=1}^m |r_j^\circ - r_j^*|.$$

To get the lower bound on the left side, we evaluate its expected value over $(x_i)_{i=1}^m$, i.e., $\frac{1}{2}\|f(\cdot; \nu^*) - f(\cdot; \nu^\circ)\|_{L_2(P_\mathcal{X})}^2$. Now we have

$$\begin{aligned} \|f(\cdot; \nu^*) - f(\cdot; \nu^\circ)\|_{L_2(P_\mathcal{X})}^2 = & \sum_{j=1}^m \|r_j^\circ \sigma(\langle \theta_j^\circ, \cdot \rangle) - r_j^* \sigma(\langle \theta_j^*, \cdot \rangle)\|_{L_2(P_\mathcal{X})}^2 \\ & + \sum_{j \neq j'} \left\langle r_j^\circ \sigma(\langle \theta_j^\circ, \cdot \rangle) - r_j^* \sigma(\langle \theta_j^*, \cdot \rangle), r_{j'}^\circ \sigma(\langle \theta_{j'}^\circ, \cdot \rangle) - r_{j'}^* \sigma(\langle \theta_{j'}^*, \cdot \rangle) \right\rangle_{L_2(P_\mathcal{X})}. \end{aligned}$$

Then we evaluate the each term. For $j \in [m]$, let $\phi = \mathrm{dist}(\theta_j^\circ, \theta_j^*)$. Then, we obtain

$$\begin{aligned} & \|r_j^\circ \sigma(\langle \theta_j^\circ, \cdot \rangle) - r_j^* \sigma(\langle \theta_j^*, \cdot \rangle)\|_{L_2(P_\mathcal{X})}^2 \\ = & r_j^{\circ 2} \mathbb{E}_X[\sigma(\langle \theta_j^\circ, X \rangle)^2] - 2 r_j^\circ r_j^* \mathbb{E}_X[\sigma(\langle \theta_j^\circ, X \rangle) \sigma(\langle \theta_j^*, X \rangle)] + r_j^{*2} \mathbb{E}_X[\sigma(\langle \theta_j^*, X \rangle)^2] \\ = & \frac{1}{2d} r_j^{\circ 2} + \frac{1}{2d} r_j^{*2} - 2 r_j^\circ r_j^* \frac{1}{2d} \left(\frac{\pi - \phi}{\pi} \langle \theta_j^*, \theta_j^\circ \rangle + \frac{\sin \phi}{\pi} \right) \\ = & \frac{1}{2d} \left[r_j^{\circ 2} - 2 r_j^\circ r_j^* + r_j^{*2} + 2 r_j^\circ r_j^* \left(1 - \frac{\pi - \phi}{\pi} \langle \theta_j^*, \theta_j^\circ \rangle - \frac{\sin \phi}{\pi} \right) \right] \\ = & \frac{1}{2d} \left[(r_j^\circ - r_j^*)^2 + 2 r_j^\circ r_j^* \left(1 - \langle \theta_j^*, \theta_j^\circ \rangle - \frac{\phi}{\pi} (1 - \langle \theta_j^*, \theta_j^\circ \rangle) + \frac{\phi - \sin \phi}{\pi} \right) \right] \\ = & \frac{1}{2d} \left[(r_j^\circ - r_j^*)^2 + 2 r_j^\circ r_j^* \left(1 - \frac{\phi}{\pi} \right) (1 - \langle \theta_j^*, \theta_j^\circ \rangle) \right] + O(\phi^3), \end{aligned}$$

where $\mathbb{E}_X$ denotes the expectation over $P_\mathcal{X}$. Since $1 - \langle \theta_j^*, \theta_j^\circ \rangle = \Theta(\phi^2)$, the higher order term $O(\phi^3)$ is negligible for sufficiently small $\epsilon > 0$, which is the same as Proposition A.1.

For $j \neq j'$ and x,y $\in \{\circ, *\}$, let $\phi^{\mathrm{xy}}_{j,j'} = \mathrm{dist}(\theta^{\mathrm{x}}_j, \theta^{\mathrm{y}}_{j'})$. Then, we have that

$$
\begin{aligned}
&\left\langle r_j^\circ \sigma(\langle \theta_j^\circ, \cdot\rangle) - r_j^* \sigma(\langle \theta_j^*, \cdot\rangle), r_{j'}^\circ \sigma(\langle \theta_{j'}^\circ, \cdot\rangle) - r_{j'}^* \sigma(\langle \theta_{j'}^*, \cdot\rangle) \right\rangle_{L_2(P_{\mathcal{X}})} \\
=&\frac{1}{2d}\left[r_j^\circ r_{j'}^\circ \left(\frac{\pi - \phi^{\circ\circ}_{j,j'}}{\pi}\langle \theta_j^\circ, \theta_{j'}^\circ\rangle + \frac{\sin \phi^{\circ\circ}_{j,j'}}{\pi}\right) - r_j^\circ r_{j'}^* \left(\frac{\pi - \phi^{\circ *}_{j,j'}}{\pi}\langle \theta_j^\circ, \theta_{j'}^*\rangle + \frac{\sin \phi^{\circ *}_{j,j'}}{\pi}\right) \right.\\
&\qquad \left. - r_j^* r_{j'}^\circ \left(\frac{\pi - \phi^{*\circ}_{j,j'}}{\pi}\langle \theta_j^*, \theta_{j'}^\circ\rangle + \frac{\sin \phi^{*\circ}_{j,j'}}{\pi}\right) + r_j^* r_{j'}^* \left(\frac{\pi - \phi^{**}_{j,j'}}{\pi}\langle \theta_j^*, \theta_{j'}^*\rangle + \frac{\sin \phi^{**}_{j,j'}}{\pi}\right)\right] \\
=&\frac{1}{2\pi d}\left\{ (r_j^\circ - r_j^*)(r_{j'}^\circ - r_{j'}^*) \right.\\
&\quad + r_j^\circ r_{j'}^\circ \left[(\pi - \phi^{\circ\circ}_{j,j'})\langle \theta_j^\circ, \theta_{j'}^\circ\rangle + \sin\phi^{\circ\circ}_{j,j'} - 1\right] - r_j^\circ r_{j'}^* \left[(\pi - \phi^{\circ *}_{j,j'})\langle \theta_j^\circ, \theta_{j'}^*\rangle + \sin\phi^{\circ *}_{j,j'} - 1\right] \\
&\qquad \left. - r_j^* r_{j'}^\circ \left[(\pi - \phi^{*\circ}_{j,j'})\langle \theta_j^*, \theta_{j'}^\circ\rangle + \sin\phi^{*\circ}_{j,j'} - 1\right] + r_j^* r_{j'}^* \left[(\pi - \phi^{**}_{j,j'})\langle \theta_j^*, \theta_{j'}^*\rangle + \sin\phi^{**}_{j,j'} - 1\right] \right\}.
\end{aligned}
$$

Here, we note that $\langle \theta_j^\circ, \theta_{j'}^*\rangle = \cos\phi^{\circ *}_{j,j'} = -(\phi^{\circ *}_{j,j'} - \pi/2) + \mathrm{O}((\phi^{\circ *}_{j,j'} - \pi/2)^3)$ and $\sin\phi^{\circ *}_{j,j'} = 1 - (\phi^{\circ *}_{j,j'} - \pi/2)^2 + \mathrm{O}((\phi^{\circ *}_{j,j'} - \pi/2)^4)$. Therefore, it holds that

$$
\begin{aligned}
&r_j^\circ r_{j'}^* \left[(\pi - \phi^{\circ *}_{j,j'})\langle \theta_j^\circ, \theta_{j'}^*\rangle + \sin\phi^{\circ *}_{j,j'} - 1\right] \\
&= r_j^\circ r_{j'}^* \left[\frac{\pi}{2}\langle \theta_j^\circ, \theta_{j'}^*\rangle + (\pi/2 - \phi^{\circ *}_{j,j'})\langle \theta_j^\circ, \theta_{j'}^*\rangle + \sin\phi^{\circ *}_{j,j'} - 1\right] \\
&= \frac{\pi}{2} r_j^\circ r_{j'}^* \langle \theta_j^\circ, \theta_{j'}^*\rangle + r_j^\circ r_{j'}^* \left[(\pi/2 - \phi^{\circ *}_{j,j'})^2 + O((\phi^{\circ *}_{j,j'} - \pi/2)^4) - (\phi^{\circ *}_{j,j'} - \pi/2)^2 + \mathrm{O}((\phi^{\circ *}_{j,j'} - \pi/2)^4)\right] \\
&= \frac{\pi}{2} r_j^\circ r_{j'}^* \langle \theta_j^\circ, \theta_{j'}^*\rangle + \mathrm{O}((\phi^{\circ *}_{j,j'} - \pi/2)^4).
\end{aligned}
$$

By applying the same argument to the all cross terms, we obtain that

$$
\begin{aligned}
&\left\langle r_j^\circ \sigma(\langle \theta_j^\circ, \cdot\rangle) - r_j^* \sigma(\langle \theta_j^*, \cdot\rangle), r_{j'}^\circ \sigma(\langle \theta_{j'}^\circ, \cdot\rangle) - r_{j'}^* \sigma(\langle \theta_{j'}^*, \cdot\rangle) \right\rangle_{L_2(P_{\mathcal{X}})} \\
&= \frac{1}{2\pi d}\left\{ (r_j^\circ - r_j^*)(r_{j'}^\circ - r_{j'}^*) + \frac{\pi}{2}\left(r_j^\circ r_{j'}^\circ \langle \theta_j^\circ, \theta_{j'}^\circ\rangle - r_j^\circ r_{j'}^* \langle \theta_j^\circ, \theta_{j'}^*\rangle - r_j^* r_{j'}^\circ \langle \theta_j^*, \theta_{j'}^\circ\rangle + r_j^* r_{j'}^* \langle \theta_j^*, \theta_{j'}^*\rangle \right) \right\} \\
&\quad + \mathrm{O}(\text{higher order}) \\
&= \frac{1}{2\pi d}\left\{ (r_j^\circ - r_j^*)(r_{j'}^\circ - r_{j'}^*) + \frac{\pi}{2}\langle r_j^\circ \theta_j^\circ - r_j^* \theta_j^*, r_{j'}^\circ \theta_{j'}^\circ - r_{j'}^* \theta_{j'}^*\rangle \right\} + \mathrm{O}(\text{higher order}).
\end{aligned}
$$

Combining all evaluations, we have that

$$
\begin{aligned}
&\|f(\cdot;\nu^*) - f(\cdot;\nu^\circ)\|^2_{L_2(P_{\mathcal{X}})} \\
=&\frac{1}{2d}\sum_{j=1}^m \left[(r_j^\circ - r_j^*)^2 + 2 r_j^\circ r_j^* \left(1 - \frac{\phi}{\pi}\right)(1 - \langle \theta_j^*, \theta_j^\circ\rangle)\right] \\
&+ \sum_{j\neq j'} \frac{1}{2\pi d}\left\{ (r_j^\circ - r_j^*)(r_{j'}^\circ - r_{j'}^*) + \frac{\pi}{2}\langle r_j^\circ \theta_j^\circ - r_j^* \theta_j^*, r_{j'}^\circ \theta_{j'}^\circ - r_{j'}^* \theta_{j'}^*\rangle \right\} + \mathrm{O}(\text{higher order}) \\
=&\sum_{j=1}^m \sum_{j'=1}^m \left[\frac{1}{2\pi d}(r_j^\circ - r_j^*)(r_{j'}^\circ - r_{j'}^*) + \frac{1}{4d}\langle r_j^\circ \theta_j^\circ - r_j^* \theta_j^*, r_{j'}^\circ \theta_{j'}^\circ - r_{j'}^* \theta_{j'}^*\rangle \right] \\
&+ \sum_{j=1}^m \left[\left(\frac{1}{2d} - \frac{1}{2\pi d} - \frac{1}{4d}\right)(r_j^\circ - r_j^*)^2 + \left(\frac{1}{d}\left(1 - \frac{\phi^{\circ *}_{j,j}}{\pi}\right) - \frac{1}{2d}\right) r_j^\circ r_j^* (1 - \langle \theta_j^*, \theta_j^\circ\rangle)\right] + \mathrm{O}(\text{higher order}).
\end{aligned}
$$

Note that the second term in the right hand side can be lower bounded by

$$\begin{cases} \frac{1}{d}\left(\frac{1}{4}-\frac{1}{2\pi}\right)\sum_{j=1}^m (r_j^\circ - r_j^*)^2, \\ \min\left\{\frac{1}{12d}, \frac{1}{2d}\left(\frac{1}{2}-\frac{\phi_{j,j}^{\circ *}}{\pi}\right)\right\}\min_j\left(r_j^\circ r_j^*\right)\sum_{j=1}^m \mathrm{dist}^2(\theta_j^\circ, \theta_j^*). \end{cases}$$

In addition to this evaluation, by noticing

$$\|f(\cdot;\nu^*) - f(\cdot;\nu^\circ)\|_{L_2(P_\mathcal{X})}^2 - \|f(\cdot;\nu^*) - f(\cdot;\nu^\circ)\|_n^2 = \mathrm{O}_p\left(\sum_{j=1}^m \frac{(r_j^* - r_j^\circ)^2 + \mathrm{dist}^2(\theta_j^\circ, \theta_j^*)}{\sqrt{n}}\right),$$

and

$$\|f(\cdot;\nu^*) - f(\cdot;\nu^\circ)\|_n^2 \le \lambda\sum_{j=1}^m |r_j^\circ - r_j^*| \le \frac{1}{2\mu}m\lambda^2 + \frac{\mu}{2}\sum_{j=1}^m (r_j^\circ - r_j^*)^2,$$

for $\mu = \frac{1}{d}\left(\frac{1}{4}-\frac{1}{2\pi}\right)$, we finally obtain that

$$\sum_{j=1}^m (r_j^\circ - r_j^*)^2 = \mathrm{O}\left(m\lambda^2\right), \quad \sum_{j=1}^m \mathrm{dist}^2(\theta_j^\circ, \theta_j^*) = \mathrm{O}\left(m\lambda^2\right),$$

with high probability. □

B. Proof of Theorem 4.8

In this section, we give the proof of Theorem 4.8.

B.1. Preliminaries

First, we ensure boundedness of the gradients during the optimization, which is required in the proof. These follow from the boundedness of the objective function (Assumption 4.6).

Lemma B.1. *Under Assumptions 4.6 and 4.7, it holds that for any $j \in [M]$ and $k = 0, 1, 2, \ldots$,*

$$\frac{1}{n}\sum_{i=1}^{n}|f(x_i;\Theta_k) - y_i| + \lambda \leq 2\sqrt{n}C_F + \lambda =: C_1, \tag{31}$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(f(x_i;\Theta_k) - y_i)x_i \mathbb{1}\{\langle w_{j,k}, x_i\rangle \geq 0\}\right\| \leq 2\sqrt{n}C_F =: C_2. \tag{32}$$

These bounds are used several times throughout the proof. From this, we can derive the following relationship between the norms of $a_{j,k}$ and $w_{j,k}$.

Lemma B.2. *Under Assumptions 4.6 and 4.7, if $\alpha < 2/C_2$, it holds that for any j, k,*

1. $|a_{j,k}| \leq \|w_{j,k}\|$,

2. $|w_{j,k}|^2 \leq a_{j,k}^2 + 1$.

Proof. We prove these inequalities by induction on k. In the case $k = 0$, it holds clearly by the initialization rule. Assume that each inequality holds for $k = k_0$, then for any j, we have

$$\begin{aligned}
|a_{j,k_0+1}|^2 - \|w_{j,k_0+1}\|^2 &= |a_{j,k_0} - \eta_{j,k_0} g_j(\Theta_{k_0})|^2 - \|w_{j,k_0} - \eta_{j,k_0} h_j(\Theta_{k_0})\|^2 \\
&= |a_{j,k_0}|^2 - \|w_{j,k_0}\|^2 \\
&+ \eta_{j,k_0}^2(\|w_{j,k_0}\|^2 - a_{j,k_0}^2)\left\|\frac{1}{n}\sum_{i=1}^{n}(f(x_i;\Theta_{k_0}) - y_i)x_i \mathbb{1}\{\langle w_{j,k_0}, x_i\rangle \geq 0\}\right\|^2 \\
&\leq \left(1 - \frac{\alpha^2}{4}\left\|\frac{1}{n}\sum_{i=1}^{n}(f(x_i;\Theta_{k_0}) - y_i)x_i \mathbb{1}\{\langle w_{j,k_0}, x_i\rangle \geq 0\}\right\|^2\right)(|a_{j,k_0}|^2 - \|w_{j,k_0}\|^2),
\end{aligned} \tag{33}$$

where we used the inequality $\eta_{j,k_0} \leq \alpha/2$. By Lemma B.1, we get the inequality of $k = k_0 + 1$ under the assumption $\alpha < 2/C_2$. $\square$

B.2. Conic Gradient Descent

In this section, we explain our proof strategy to show Theorem 4.8. The key technical tool in our proof is to fully make use of the update in the measure space. At first, we consider the update of $(r_{j,k}, \theta_{j,k}) \in \mathbb{R} \times \mathbb{S}^{d-1}$, which are amplitude and location of each Dirac measure. By the update rule of the parameters, we obtain the following recursive expression of each parameter:

$$
\begin{aligned}
r_{j,k+1} &= (a_{j.k} - \eta_{j,k} g(a_{j,k}))\|w_{j,k} - \eta_{j,k} h_j(\Theta_k)\| \\
&= (a_{j.k} - \eta_{j,k} g(a_{j,k}))\left(\|w_{j,k}\| - \eta_{j,k}\frac{\langle w_{j,k}, h_j(\Theta_k)\rangle}{\|w_{j,k}\|} + \delta w_{j,k}\right) \\
&= r_{j,k} - \eta_{j,k}\frac{a_{j,k}^2 + \|w_{j,k}\|^2}{|a_{j,k}|\|w_{j,k}\|}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i;\nu) - y_i)\sigma(\langle\theta_{j,k}, x_i\rangle\rangle) + \lambda\,\mathrm{sgn}(r_{j,k})\right) r_{j,k} + \delta r_{j,k}, \\
\theta_{j,k+1} &= \frac{w_{j,k+1}}{\|w_{j,k+1}\|} = \frac{w_{j,k} - \eta_{j,k} h_j(\Theta_k)}{\|w_{j,k} - \eta_{j,k} h_j(\Theta_k)\|} \\
&= \theta_{j,k} - \eta_{j,k}\frac{1}{\|w_{j,k}\|}(\mathrm{I}_d - \theta_{j,k}\theta_{j,k}^\top) h_j(\Theta_k) + \delta\theta_{j,k} \\
&= \theta_{j,k} - \eta_{j,k}\frac{a_{j,k}}{\|w_{j,k}\|}(\mathrm{I}_d - \theta_{j,k}\theta_{j,k}^\top)\left(\frac{1}{n}\sum_{i=1}^n (f(x_i;\Theta_k) - y_i) x_i \mathbb{1}\{\langle\theta_{j,k}, x_i\rangle \geq 0\}\right) + \delta\theta_{j,k},
\end{aligned}
$$

where $\delta w_{j,k}, \delta r_{j,k}, \delta\theta_{j,k}$ are residual higher-order terms. From the view point of the measure space, this can be expressed as

$$
\begin{aligned}
r_{j,k+1} &= r_{j,k} - \eta_{j,k}\frac{a_{j,k}^2 + \|w_{j,k}\|^2}{|a_{j,k}|\|w_{j,k}\|} G_{\nu_k}(\theta_{j,k}) r_{j,k} + \delta r_{j,k}, \\
\theta_{j,k+1} &= \theta_{j,k} - \eta_{j,k}\frac{a_{j,k}}{\|w_{j,k}\|}\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k}) + \delta\theta_{j,k},
\end{aligned}
$$

where $G_{\nu_k} \in \partial J(\nu_k)$. Here, the subdifferential $\partial J(\nu_k)$ is defined as $\partial J(\nu_k) := \{G \in \mathcal{C}(\mathbb{S}^{d-1}) \mid J(\mu) - J(\nu_k) \geq \int G(\theta)\mathrm{d}(\mu - \nu_k)\ (\forall \mu \in \mathcal{M}(\mathbb{S}^{d-1}))\}$ which is well defined because $J(\cdot)$ is a convex function on the measure space $\mathcal{M}(\mathbb{S}^{d-1})$. Furthermore, by the definition of $\eta_{j,k}$, this iteration can be rewritten as

$$
r_{j,k+1} = r_{j,k} - \alpha G_{\nu_k}(\theta_{j,k}) r_{j,k} + \delta r_{j,k}, \tag{34}
$$

$$
\theta_{j,k+1} = \theta_{j,k} - \alpha\,\mathrm{sgn}(r_{j,k})\frac{a_{j,k}^2}{a_{j,k}^2 + \|w_{j,k}\|^2}\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k}) + \delta\theta_{j,k}. \tag{35}
$$

We note that the term $\delta r_{j,k}$ and $\delta\theta_{j,k}$ can be seen as "higher order" term by the following lemma.

Lemma B.3. *Under Assumption 4.6, if $\alpha < 1/C_1$, it holds that for any j, k,*

$$
\begin{aligned}
|\delta r_{j,k}| &\leq C_1\alpha^2 |G_{\nu_k}(\theta_{j,k}) r_{j,k}|, \\
\|\delta\theta_{j,k}\| &\leq 5C_2\alpha^2\frac{a_{j,k}^2}{a_{j,k}^2 + \|w_{j,k}\|^2}\|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k})\|.
\end{aligned}
$$

Proof. At first, by the straight-forward calculation, we have that $\|G_{\nu_k}\|_\infty \leq C_1$ and $\sup_{\theta\in\mathbb{S}^{d-1}}\|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\| \leq C_2$. This gives that $\|\eta_{j,k} h_j(\Theta_k)\| \leq \alpha\|w_{j,k}\|\|G_{\nu_k}\|_\infty/2 < \|w_{j,k}\|/2$. By using Lemma B.2 and Lemma C.1, we have

$$
\begin{aligned}
|\delta r_{j,k}| = |a_{j,k} - \eta_{j,k} g_j(\Theta_k)|\|\delta w_{j,k}\| &\leq 2|a_{j,k}|\frac{\eta_{j,k}^2 h^2(w_{j,k})}{\|w_{j,k}\|} \\
&\leq C_1\alpha^2 |G_{\nu_k}(\theta_{j,k}) r_{j,k}|.
\end{aligned}
$$

Moreover, by Lemma C.2, it holds that

$$
\|\delta\theta_{j,k}\| \leq \frac{5\|\eta_{j,k} h_j(\Theta_k)\|^2}{\|w\|^2} \leq 5C_2\alpha^2\frac{a_{j,k}^2}{a_{j,k}^2 + \|w_{j,k}\|^2}\|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k})\|.
$$

These give the conclusion. ☐

By this lemma, we can see that $\delta r_{j,k}$ and $\delta\theta_{j,k}$ are $O(\alpha^2)$ which is smaller than other terms.

Remark B.4. *In the case $\exists j, j' \in [m]$ $(j \neq j')$, $\theta_{j,k} = \theta_{j,k'}$, we cannot represent the update by the subgradient in the measure space. However, we can avoid this problem almost surely by perturbing the step size infinitesimally. In the following, we assume this does not happen for any j, k.*

Chizat (2019) considered a conic gradient descent, which is represented as follows:

$$(r_{j,k+1}, \theta_{j,k+1}) = \mathrm{Ret}_{(r_{j,k},\theta_{j,k})}(-2\alpha G_{\nu_k}(\theta_{j,k}) r_{j,k}, -\beta \nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k}))$$

where $\alpha, \beta > 0$ are constants and Ret denotes a retraction mapping, which is defined on the manifold $\mathbb{R} \times \mathbb{S}^{d-1}$ and its tangent bundle (Absil et al., 2009). The retraction mapping and updates in Eq. (34) and Eq. (35) are almost equivalent in a sense that both of them represent first order approximations of the gradient descent in the manifold. Motivated by this point, we borrow the proof technique developed in Chizat (2019). They have shown that under several assumptions with sufficient over-parameterization and under the condition $\beta \lesssim \alpha^2$, convergence of the gradient descent to the global optimum is achieved through the following two phase:

Phase I: Global exploration. Objective value decreases until it reaches a threshold J_0,

Phase II: Local convergence. The solution converges linearly to the global minimum locally around the true parameter.

There are some different points between our approach and Chizat (2019). One is that the step-size in the iteration of $\theta_{j,k}$ is not a constant. Indeed, by (35), the step size of the update in the measure space is given by

$$\beta_{j,k} = \alpha \frac{a_{j,k}^2}{a_{j,k}^2 + \|w_{j,k}\|^2}. \tag{36}$$

This step size depends on $a_{j,k}^2$ and $\|w_{j,k}\|^2$ and is not constant. Note that by the initialization rule, $\beta_{j,0} = \frac{\alpha}{1+M^2} \ll \alpha$ for any $j \in [m]$, and we will show that the inequality $\beta_{j,k} \ll \alpha$ for all j, k, which means that the step size for $\theta_{j,k}$ is much smaller than that of $r_{j,k}$.

Another difference is that our analysis deals with the non-differentiable ReLU activation while Chizat (2019) analyzed differentiable activation functions. We avoid this difficulty by utilizing Assumption 4.3.

Moreover, Chizat (2019) only considered a positive measure (more precisely, their argument cannot be applied to the settings where the measure ν has both positive and negative parts). In this paper, we consider this situation and overcome this difficulty by utilizing the following lemma which states that a positive (resp. negative) part of the updated measure remains positive (resp. negative) throughout the iterations.

Lemma B.5. *Under Assumptions 4.6 and 4.7, if $\alpha < 1/C_1$, the signs of $(a_{j,k})_{j\in[M]}$ (i.e., those of $(r_{j,k})_{j\in[M]}$) do not change throughout the iteration.*

Proof. By the update rule of $a_{j,k}$, we have

$$\begin{aligned} a_{j,k+1} &= a_{j,k} - \alpha \frac{\|w_{j,k}^2\|}{a_{j,k}^2 + \|w_{j,k}\|^2} \left(\frac{1}{n} \sum_{i=1}^n (f(x_i;\nu) - y_i)\sigma(\langle \theta_{j,k}, x_i \rangle) + \lambda \,\mathrm{sgn}(a_{j,k}) \right) a_{j,k} \\ &= \left\{ 1 - \alpha \frac{\|w_{j,k}^2\|}{a_{j,k}^2 + \|w_{j,k}\|^2} \left(\frac{1}{n} \sum_{i=1}^n (f(x_i;\nu) - y_i)\sigma(\langle \theta_{j,k}, x_i \rangle) + \lambda \,\mathrm{sgn}(a_{j,k}) \right) \right\} a_{j,k}. \end{aligned}$$

By using the inequalities $\frac{\|w_{j,k}^2\|}{a_{j,k}^2 + \|w_{j,k}\|^2} \leq 1$ and $|\frac{1}{n}\sum_{i=1}^n (f(x_i;\nu) - y_i)\sigma(\langle \theta_{j,k}, x_i \rangle) + \lambda \,\mathrm{sgn}(a_{j,k})| \leq C_1$, we get the conclusion. $\square$

B.3. Proof of Phase I

In this section, we show the following inequality.

Proposition B.6 (Global exploration)**.** *Assume that Assumption 4.6 holds. Then there exists a constant $C,\ C_M > 0$ such that for any $J_0 > J^*$ and $0 < \epsilon < 1/2$, by setting M sufficiently large as $M \geq C_M \exp(\alpha^{-2})/\alpha$ for each $\alpha > 0$ and assuming the following conditions,*

$$W_\infty(\tau, \nu_0^+) \leq (J_0 - J^*)/C,\ W_\infty(\tau, \nu_0^-) \leq (J_0 - J^*)/C,\ \alpha \leq (J_0 - J^*)^{1+\epsilon/2}/C, \tag{37}$$

then it holds that

$$\min_{0 \leq k' \leq \alpha^{-2}} J(\nu_{k'}) \leq J_0. \tag{38}$$

Here we utilize the bound by Chizat (2019), which considered a positive measure, i.e., $r_{j,k} > 0$ for any j and k. By Lemma B.5, the signs of $(r_{j,k})_{j\in[M]}$ will not change throughout the iterations. Therefore, we can apply the same argument to ν_k^+ and ν_k^- separately, where $\nu_k := \nu_k^+ - \nu_k^-$ is the Hahn-Jordan decomposition. Then we get the following proposition.

Proposition B.7. *Suppose that Assumption 4.6 holds. In addition, suppose that $\beta_{\max} := \max_{j\in[M], 1\leq k \leq \alpha^{-2}} \beta_{j,k} \leq \alpha^3$. Let $B := \sup_{\nu: J(\nu) < C_F} \|\nu\|_{BL}$, then there exists a constant $C' > 0$ such that, for $\alpha < 1/C_1$, it holds that*

$$\min_{1\leq k \leq \alpha^{-2}} J(\nu_k) - J^* \leq C'(\log(4B\alpha^{-1}) + 1)\alpha + B\|\nu^*\|_{\mathrm{TV}}(W_\infty(\tau, \nu_0^+) + W_\infty(\tau, \nu_0^-)).$$

Proof. Following the essentially same argument as Lemma F.1 of Chizat (2019), it holds that

$$\min_{1\leq k \leq \alpha^{-2}} J(\nu_k) - J^* \leq C' \frac{\log(4B\alpha k')}{4B\alpha k'} + \beta_{\max} B^2 k' + C\alpha + B\|\nu^*\|_{\mathrm{TV}}(W_\infty(\tau, \nu_0^+) + W_\infty(\tau, \nu_0^-)).$$

In particular, in the case $k' = \alpha^{-2}$, we get an upper bound as

$$C' \frac{\log(4B\alpha^{-1})}{4B}\alpha + \frac{\beta_{\max}}{\alpha^2} B^2 + C'\alpha + B\|\nu^*\|_{\mathrm{TV}}(W_\infty(\tau, \nu_0^+) + W_\infty(\tau, \nu_0^-)).$$

With the condition $\beta_{\max} \leq \alpha^3$, we get the conclusion. □

Proof of Proposition B.6. For $0 < \epsilon < 1/2$, there exists a constant $C_\epsilon > 0$ such that $\log(u) \leq C_\epsilon u^\epsilon$. Then we have

$$\min_{1\leq k \leq \alpha^{-2}} J(\nu_k) - J^* \leq C'(C_\epsilon B^{-1+\epsilon}\alpha^{-\epsilon} + 1)\alpha + B\|\nu^*\|_{\mathrm{TV}}(W_\infty(\tau, \nu_0^+) + W_\infty(\tau, \nu_0^-))$$

This yields the conclusion that there exists a constant $C > 0$ which depends on $C', C_\epsilon, B, B, \|\nu^*\|$ and the inequality (38) is satisfied under the condition (37). □

In the following, we show the inequality $\beta_{\max} \leq \alpha^3$. This intuitively means that the "location" $\theta_{j,k}$ does not move compared with the "amplitude" $r_{j,k}$. We can verify this in the setting we consider, in which $a_{j,k}$ is much smaller than $w_{j,k}$. Note that $\beta_{j,k} \leq \alpha |a_{j,k}|^2/\|w_{j,k}\|^2$. Inspired by this inequality, we evaluate $|a_{j,k}|$ and $\|w_{j,k}\|$, and prove the inequality $|a_{j,k}| \ll \|w_{j,k}\|$ for $k \leq \alpha^{-2}$.

Lemma B.8. *Assume that Assumption 4.6 holds. Let $\xi_{j,k} = \left(1 + \frac{2}{M}\right)\left\{\prod_{k'=0}^{k-1}(1 + \eta_{j,k'}C_1) - 1\right\}$ $(j \in [M],\ k = 1, 2, \dots)$, it holds that*

$$|a_{j,k}| \leq \frac{2}{M} + \xi_{j,k}, \tag{39}$$

$$\|w_{j,k} - w_{j,0}\| \leq \xi_{j,k}. \tag{40}$$

Proof. By the update rules of $a_{j,k}$ and $w_{j,k}$, we have that

$$|a_{j,k+1}| \leq |a_{j,k}| + \eta_{j,k}\left(\frac{1}{n}\sum_{i=1}^n |f(x_i; \Theta_k) - y_i| + \lambda\right)\|w_{j,k}\|, \tag{41}$$

$$\|w_{j,k+1}\| \leq \|w_{j,k}\| + \eta_{j,k}\left(\frac{1}{n}\sum_{i=1}^n |f(x_i; \Theta_k) - y_i| + \lambda\right)|a_{j,k}|. \tag{42}$$

By Lemma B.1, $\frac{1}{n}\sum_{i=1}^{n}|f(x_i;\Theta_k)-y_i|+\lambda\le C_1$ for all k. By summing up the both sides, it holds that

$$|a_{j,k}|+\|w_{j,k}\|\le(1+\eta_{j,k}C_1)\left(|a_{j,k}|+\|w_{j,k}\|\right).$$

Then we have

$$\max\{|a_{j,k}|,\|w_{j,k}\|\}\le|a_{j,k}|+\|w_{j,k}\|\le\left(1+\frac{2}{M}\right)\prod_{k'=0}^{k-1}(1+\eta_{j,k}C_1).$$

Combining with (41), $|a_{j,k}|$ is bounded as

$$\begin{aligned}|a_{j,k}|&\le|a_{j,0}|+\sum_{k'=0}^{k-1}\eta_{j,k'}C_1\left(1+\frac{2}{M}\right)\prod_{k''=0}^{k'-1}(1+\eta_{j,k''}C_1)\\&=\frac{2}{M}+\left(1+\frac{2}{M}\right)\left\{\prod_{k'=0}^{k-1}(1+\eta_{j,k}C_1)-1\right\},\end{aligned}\tag{43}$$

which gives the first inequality Eq. (39). In addition, similar to Eq. (42), we have

$$\|w_{j,k+1}-w_{j,0}\|\le\|w_{j,k}-w_{j,0}\|+\eta_{j,k}C_1|a_{j,k}|.$$

Combining with the bound of $|a_{j,k}|$, we get the second inequality (40). □

From this bound, we obtain a bound on $|a_{j,k}|/\|w_{j,k}\|$ as we state as follows.

Lemma B.9. *Under Assumption 4.6, for any j,k satisfying $\xi_{j,k}<1$, it holds that*

$$\frac{|a_{j,k}|}{\|w_{j,k}\|}\le\frac{2/M+\xi_{j,k}}{1-\xi_{j,k}}.$$

Moreover, there exists a constant $C_M>0$ such that if $M\ge C_M\exp(\alpha^{-2})/\alpha$, it holds that $|a_{j,k}|/\|w_{j,k}\|\le\alpha$ for any $j\in[M]$ and k satisfying $1\le k\le\alpha^{-2}$.

Proof. The first conclusion holds clearly by Lemma B.8. Then we consider the second assertion. Let $\zeta_{j,k}:=\frac{|a_{j,k}|}{\|w_{j,k}\|}$. Suppose that $\xi_{j,k}\le1/2$ (which we verify later), it holds that

$$\zeta_{j,k}\le2\left(\frac{2}{M}+\xi_{j,k}\right).$$

In addition, since $\eta_{j,k}=\alpha|a_{j,k}|\|w_{j,k}\|/(|a_{j,k}|^2+\|w_{j,k}\|^2)\le\alpha\zeta_{j,k}$, we have

$$\xi_{j,k}\le\left(1+\frac{2}{M}\right)\left\{\prod_{k'=0}^{k-1}(1+\alpha\zeta_{j,k'}C_1)-1\right\}.$$

by the formulation of $\xi_{j,k}$. Combining these inequality, we get

$$\zeta_{j,k}\le\frac{4}{M}+2\left(1+\frac{2}{M}\right)\left\{\prod_{k'=0}^{k-1}(1+\alpha\zeta_{j,k'}C_1)-1\right\},$$

where we used $k\le\alpha^{-2}$. By this inequality, let $c\gtrsim\log(\alpha^{-1})$ and $M\gtrsim\exp(c\alpha^{-2})$, then we have $\zeta_{j,k}\le\frac{2}{M}\exp(ck)$ for any $0\le k\le\alpha^{-2}$ and we prove this by the induction. When $k=0$ this holds with equality. Suppose that for $k_0\ge1$,

$\zeta_{j,k} \leq \frac{2}{M}\exp(ck)$ is satisfied for any $k < k_0$. Then, we have

$$
\begin{aligned}
\zeta_{j,k_0} &\leq \frac{4}{M} + 2\left(1+\frac{2}{M}\right)\left\{\prod_{k'=0}^{k_0-1}(1+\alpha\zeta_{j,k'}C_1) - 1\right\} \\
&\leq \frac{4}{M} + 2\left(1+\frac{2}{M}\right)\left\{\left(1+\frac{2\alpha C_1}{M}\exp(c(k_0-1))\right)^{k_0} - 1\right\} \\
&\leq \frac{4}{M} + 2\left(1+\frac{2}{M}\right)\frac{4\alpha C_1 k_0}{M}\exp(c(k_0-1)) \\
&\leq \frac{2}{M}\Big(2 + 8\alpha C_1 k_0 \exp(c(k_0-1))\Big) \\
&\leq \frac{2}{M}\exp(ck_0),
\end{aligned}
$$

where the third inequality follows from

$$
\begin{aligned}
\left(1+\frac{2}{M}\exp(c(k_0-1))\right)^{k_0} &\leq 1 + \frac{2k_0}{M}\exp(c(k_0-1)) + 2^{k_0}\frac{4}{M^2}\exp(2c(k_0-1)) \\
&\leq 1 + \frac{4k_0}{M}\exp(c(k_0-1)),
\end{aligned}
$$

where we use $\sum_{j=2}^{k_0}\binom{k_0}{j} \leq 2^{k_0}$. Taking $M \geq 2\alpha^{-1}\exp(c\alpha^{-2})$, we get $\zeta_{j,k} \leq \alpha$. Finally, in this case the condition $\xi_{j,k} \leq 1/2$ remains and this gives the conclusion. □

By this Lemma, it holds that for sufficiently large M, $|a_{j,k}|/\|w_{j,k}\|$ will be small. By this inequality, we get a bound of $\beta_{j,k}$, which is supposed in the Proposition B.7.

Lemma B.10. *Under Assumption 4.6, there exists a constant $C_M > 0$ such that if $M \geq C_M \exp(\alpha^{-2})/\alpha$, it holds that $\beta_{j,k} \leq \alpha^3$ for any $j \in [M]$ and k satisfying $1 \leq k \leq \alpha^{-2}$.*

Proof. By the definition of $\beta_{j,k}$, it holds that

$$
\beta_{j,k} = \alpha\frac{a_{j,k}^2}{a_{j,k}^2 + \|w_{j,k}\|^2} \leq \alpha\zeta_{j,k}^2.
$$

Combining this with Lemma B.9, we get the conclusion. □

B.4. Proof of Phase II

In this section, we prove linear convergence to the optimal solution after a specific number of iterations. A key ingredient is a local analysis around the optimal parameters $(\theta_j^*)_{j=1}^{m^*}$ (remark that the global minimum is obtained by a sparse measure). We consider a local region around each θ_j^* which we define below and prove a "sharpness inequality" (Proposition B.15) through evaluating the function value and the norm of the gradient by using a distance from the optimal parameter.

We first divide $\mathbb{S}^{d-1}$ by the sign of inner product with each x_i, i.e., each division is written by the form $\{\theta \in \mathbb{S}^{d-1} \mid \mathrm{sgn}(\langle\theta, x_1\rangle) = s_1, \ldots, \mathrm{sgn}(\langle\theta, x_n\rangle) = s_n\}$ for $(s_i)_{i=1}^n \in \{-1,+1\}^n$. Let H_j be the region that contains θ_j^* and $R_j := \sup_{\theta\in H_j}\mathrm{dist}(\theta, \theta_{j^*})$, where $R_j > 0$ by Assumption 4.3. Then we take a a value ρ which satisfies $0 < \rho < \frac{\min R_j}{2}$ and define $N_j(\rho)$ to be an open ball around θ_j^* with radius ρ and $N_0 := \mathbb{S}^{d-1}\setminus \cup_j N_j(\rho)$.

To prove the linear convergence, we define a kind of distance between ν_k and the global minima ν^*. Our definition of the distance follows that of Chizat (2019) but they are different in that we deal with a singed measure and their definition did not properly deal with average on the manifold $\mathbb{S}^{d-1}$ while ours avoid such an average.

Definition B.11. *Let $\nu_k = \sum_{j=1}^M r_{j,k}\delta_{\theta_{j,k}}$ be the measure after k iterations. For each $j \in [m^*]$, we define a local mass by $\bar{r}_{j,k} = \nu_k(N_j(\rho))$, a local gap on $\mathbb{S}^{d-1}$ with $\Delta\theta_{j,k} = \sum_{\substack{\theta_{j',k}\in N_j(\rho)\\ \mathrm{sgn}(r_{j',k})=\mathrm{sgn}(r_j^*)}} |r_{j',k}|\mathrm{dist}^2(\theta_{j',k})$ and a local "different signed"* *mass with* $\Delta r_{j,k} = \sum_{\substack{\theta_k \in N_j(\rho) \\ \mathrm{sgn}(r_{j',k}) \neq \mathrm{sgn}(r_j^*)}} |r_{j',k}|$. *Furthermore, we define a mass of the remaining region with* $r_{0,k} := |\nu_k|(N_0)$. *Finally, according to these values, we define a "distance" between* ν_k *and* ν^* *by*

$$D_\rho(\nu_k) = \sum_{j=1}^{m^*} (\bar{r}_{j,k} - r_j^*)^2 + r_{0,k} + \sum_{j=1}^{m^*} (\Delta\theta_{j,k} + \Delta r_j).$$

As we see below, the term $(\bar{r}_{j,k} - r_j^*)^2$ and $r_{0,k}$ mainly affect the gap between $f(\cdot, \nu_k)$ and $f(\cdot, \nu^*)$. The term $\Delta\theta_{j,k} + \Delta r_j$ is related to the regularization term and will vanish due to the sparse regularization. In the following, we see that this distance upper-bounds the Wasserstein distance between ν_k and ν^*.

For the local evaluation, we firstly remark the optimality condition w.r.t. a measure.

Lemma B.12. *Let* $f^* := f(\cdot, \nu^*)$, *under Assumption 4.3, for each* $j \in [m^*]$, *it holds that*

$$-\frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i) x_i \mathbb{1}\{\langle \theta_j^*, x_i \rangle \geq 0\} = \lambda\, \mathrm{sgn}(r_j^*) \theta_j^*. \tag{44}$$

Furthermore, under Assumption 4.5, for any $\theta \in \mathbb{S}^{d-1}$ *satisfying* $\theta \neq \theta_j^*$ *for all* $j \in [m^*]$, *it holds that*

$$\left| \frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i) \sigma(\langle \theta, x_i \rangle) \right| < \lambda. \tag{45}$$

Remark that Eq. (44) is derived from $0 \in \partial J(\nu^*)$, where

$$\partial J(\nu^*) = \frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i)\sigma(\langle \cdot, x_i \rangle) + \lambda \partial \|\nu^*\|_{\mathrm{TV}} \subset \mathcal{C}(\mathbb{S}^{d-1}). \tag{46}$$

Indeed, the necessary condition for $0 \in \partial J(\nu^*)$ is that

$$\begin{cases} -\frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i)\sigma(\langle \theta_j^*, x_i \rangle) = \lambda\, \mathrm{sgn}(r_j^*), \\ -\frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i) x_i \mathbb{1}\{\langle \theta_j^*, x_i \rangle \geq 0\} = \lambda a_j \mathrm{sgn}(r_j^*) \theta_j^*, \end{cases}$$

for some $a_j \in \mathbb{R}$ for each $j \in [m^*]$, where we used the same argument to show Eq. (17). By putting together these equations and using the 1-homogeneity of ReLU, we obtain Eq. (44).

Now we introduce a characterization of subgradient $\partial J(\nu^*)$ for the proof. By the construction, we know that $\theta \mapsto -\frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i) x_i \mathbb{1}\{\langle \theta, x_i \rangle \geq 0\}$ takes a constant value $\lambda\, \mathrm{sgn}(r_j^*)\theta_j^*$ in each $N_j(\rho)$. This leads to

$$-\frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i)\sigma(\langle \theta, x_i \rangle) = \langle \theta, -\frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i) x_i \mathbb{1}\{\langle \theta_j^*, x_i \rangle \geq 0\} \rangle = \lambda\, \mathrm{sgn}(r_j^*) \langle \theta, \theta_j^* \rangle. \tag{47}$$

for $\theta \in N_j(\rho)$. This equality plays an important role in the proof.

By using $D_\rho(\nu_k)$, we can evaluate a gap between $J(\nu_k)$ and J^*.

Proposition B.13. *Under Assumption 4.3–4.7, there exists a constant* $c_\rho > 0$, $C_\rho > 0$ *that depends on* ρ *and a constant* J_0, *if* $J(\nu_k) < J_0$ *it holds that*

$$c_\rho D_\rho(\nu_k) \leq J(\nu_k) - J^* \leq C_\rho D_\rho(\nu_k). \tag{48}$$

Proof. First, we derive the first inequality $c_\rho D_\rho(\nu_k) \leq J(\nu_k) - J^*$. Let $G^*(\theta) \in \partial J(\nu^*)$, i.e.,

$$G^*(\cdot) \in \frac{1}{n}\sum_{i=1}^{n} (f^*(x_i) - y_i)\sigma(\langle \cdot, x_i \rangle) + \lambda \partial \|\nu^*\|_{\mathrm{TV}}.$$

Here, we take $\eta \in \partial\|\nu^*\|_{\mathrm{TV}}$ which satisfies $\eta(\theta_{j,k}) = \mathrm{sgn}(r_{j,k})$ (if necessary, we apply the modification as Remark B.4). In this case, by the straight-forward calculation and noticing $\int_{\mathbb{S}^{d-1}} G^*(\theta)\mathrm{d}\nu^* = 0$, it holds that

$$J(\nu_k) - J^* = \underbrace{\int_{\mathbb{S}^{d-1}} G^*(\theta)\mathrm{d}\nu_k}_{\text{(I)}} + \underbrace{\frac{1}{2}\|f(\cdot;\nu_k) - f^*\|_n^2}_{\text{(II)}}. \tag{49}$$

We evaluate each term. For the term (I), we have

$$\text{(I)} = \sum_{j=1}^{m^*} \int_{N_j(\rho)} G^*(\theta)\mathrm{d}\nu_k + \int_{N_0} G^*(\theta)\mathrm{d}\nu_k.$$

Now we have $C'_\rho := \inf_{\theta\in N_0} \left\{\lambda - \left|\frac{1}{n}\sum_{i=1}^n (f^*(x_i) - y_i)\sigma(\langle\theta, x_i\rangle)\right|\right\} > 0$ by the inequality (45) in Lemma B.12 and compactness of N_0. Then the second term of the right hand side is evaluated as

$$\int_{N_0} G^*(\theta)\mathrm{d}\nu_k \geq r_{0,k}C'_\rho. \tag{50}$$

For the term $\int_{N_j(\rho)} G^*(\theta)\mathrm{d}\nu_k$, Eq. (47) and the definition of η yield that

$$\begin{aligned}
\int_{N_j(\rho)} G^*(\theta)\mathrm{d}\nu_k &\geq \int_{N_j(\rho)} (\lambda\,\mathrm{sgn}(r_j^*)\langle\theta_j^*, \theta\rangle + \lambda\,\mathrm{sgn}(\eta(\theta)))\mathrm{d}\nu_k(\theta) \\
&\geq \lambda \int_{N_j(\rho)} \min\{1 - \mathrm{sgn}(r_j^*\eta(\theta))\langle\theta_j^*, \theta\rangle, 1\}\mathrm{d}|\nu_k| \\
&\geq \lambda(c\Delta\theta_{j,k} + \Delta r_j),
\end{aligned} \tag{51}$$

where $c > 0$ is a constant which does not depend on other parameters, which is derived from Lemma C.3. Combining Eq. (50) and Eq. (51), we get

$$\text{(I)} \geq \lambda \sum_{j=1}^{m^*} (c\Delta\theta_{j,k} + \Delta r_{j,k}) + r_{0,k}C'_\rho.$$

Next for the term (II), we consider the decomposition $f(\cdot;\nu_k) =: \sum_{j=1}^{m^*} f_{j,k} + f_{0,k}$, where $f_{j,k}(\cdot) = \int_{N_j(\rho)} \sigma(\langle\theta, \cdot\rangle)\mathrm{d}\nu_k$. Then we have

$$\begin{aligned}
\|f(\cdot;\nu_k) - f^*\|_n^2 &= \left\|\sum_{j=1}^{m^*}\left(r_j^*\sigma(\langle\theta_j^*, \cdot\rangle) - f_{j,k}\right) + f_{0,k}\right\|_n^2 \\
&\geq \left\|\sum_{j=1}^{m^*}\left(r_j^*\sigma(\langle\theta_j^*, \cdot\rangle) - f_{j,k}\right)\right\|_n^2 - 2\left\|\sum_{j=1}^{m^*}(r_j^*\sigma(\langle\theta_j^*, \cdot\rangle) - f_{j,k})\right\|_n \|f_{0,k}\|_n \\
&\geq \left\|\sum_{j=1}^{m^*}\left(r_j^*\sigma(\langle\theta_j^*, \cdot\rangle) - f_{j,k}\right)\right\|_n^2 - 2\left\|\sum_{j=1}^{m^*}(r_j^*\sigma(\langle\theta_j^*, \cdot\rangle) - f_{j,k})\right\|_n r_{0,k},
\end{aligned} \tag{52}$$

where the last inequality follows from $\int_A \sigma(\langle\theta, x\rangle)\mathrm{d}\nu_k \leq |\nu_k|(A)$ for $A \subset \mathbb{S}^{d-1}$ if $\|x\| = 1$. For the first term, we have

$$\begin{aligned}
r_j^*\sigma(\langle\theta_j^*, \cdot\rangle) - f_{j,k} &= r_j^*\sigma(\langle\theta_{j,k}^*, \cdot\rangle) - \sum_{j':\theta_{j'}\in N_j(\rho)} r_{j',k}\sigma(\langle\theta_{j',k}, \cdot\rangle) \\
&= (r_j^* - \bar{r}_{j,k})\sigma(\langle\theta_j^*, \cdot\rangle) - \sum_{j':\theta_{j'}\in N_j(\rho)} r_{j',k}(\sigma(\langle\theta_{j',k}, \cdot\rangle) - \sigma(\langle\theta_j^*, \cdot\rangle)),
\end{aligned}$$

which gives

$$\left\|\sum_{j=1}^{m^*}\left(r_j^*\sigma(\langle\theta_j^*,\cdot\rangle)-f_{j,k}\right)\right\|_n^2 \geq\kappa\sum_{j=1}^{m^*}(r_j^*-\bar{r}_{j,k})^2-2\left(\sum_{j=1}^{m^*}|r_j^*-\bar{r}_{j,k}|\right)\left(\sum_{j=1}^{m^*}(\Delta\theta_{j,k}+\Delta r_{j,k})\right)$$

where the first term is derived by Assumption 4.4. Combining with Eq. (52), we have a lower bound of (II) as

$$\text{(II)}\geq\kappa\sum_{j=1}^{m^*}(\bar{r}_{j,k}-r_j^*)^2-2\left\|\sum_{j=1}^{m^*}(r_j^*\sigma(\langle\theta_j^*,\cdot\rangle)-f_{j,k})\right\|_n r_{0,k}-2\left(\sum_{j=1}^{m^*}|\bar{r}_{j,k}-r_j^*|\right)\left(\sum_{j=1}^{m^*}(\Delta\theta_{j,k}+\Delta r_{j,k})\right).$$

Finally, we have $\max\{r_{0,k},\sum_{j=1}^{m^*}(\Delta\theta_{j,k}+\Delta r_{j,k})\}\leq\max\left\{1/(c\lambda),1/\lambda,1/C'_\rho\right\}(J(\nu_k)-J^*)$ by the lower bound of (I). For sufficiently small $J(\nu_k)-J^*$, by transposing the minus term and using the arithmetic-geometric mean relation, this leads to a bound

$$\sum_{j=1}^{m^*}(\bar{r}_{j,k}-r_j^*)^2\leq C(J(\nu_k)-J^*)$$

for some constant $C>0$. Combining (I) and (II), we get the conclusion.

To get the upper bound we use the equality Eq. (49) as

$$\begin{aligned}\text{(I)}&=\int_{N_0}G^*(\theta)\mathrm{d}\nu_k+\sum_{j=1}^{m^*}\int_{N_j(\rho)}G^*(\theta)\mathrm{d}\nu_k\\&\leq r_{0,k}\|G^*(\cdot)\|_\infty+\sum_{j=1}^{m^*}\int_{N_j(\rho)}\lambda(-\mathrm{sgn}(r_j^*)\langle\theta,\theta_j^*\rangle+\mathrm{sgn}(\eta(\theta)))\mathrm{d}\nu_k\\&\leq r_{0,k}\|G^*(\cdot)\|_\infty+\sum_{j=1}^{m^*}2\lambda(\Delta\theta_{j,k}+\Delta r_{j,k})\\&\leq C_1r_{0,k}+\sum_{j=1}^{m^*}2\lambda(\Delta\theta_{j,k}+\Delta r_{j,k}).\end{aligned}$$

For the term (II), we follow the similar argument to the lower bound as

$$\begin{aligned}\|f(\cdot;\nu_k)-f^*\|_n^2&=\left\|\sum_{j=1}^{m^*}\left(r_j^*\sigma(\langle\theta_j^*,\cdot\rangle)-f_{j,k}\right)+f_{0,k}\right\|_n^2\\&\leq 2\left\|\sum_{j=1}^{m^*}\left(r_j^*\sigma(\langle\theta_j^*,\cdot\rangle)+f_{j,k}\right)\right\|_n^2+2\left\|f_{0,k}\right\|_n^2\\&\leq 4\kappa_{\max}\sum_{j=1}^{m^*}(r_j^*-\bar{r}_{j,k})^2+4\left(\sum_{j=1}^{m^*}(\Delta\theta_{j,k}+\Delta r_{j,k})\right)^2+2r_{0,k}^2,\end{aligned}$$

where $\kappa_{\max}$ is the largest eigenvalue of a matrix $\left(\frac{1}{n}\sum_{j=1}^n\sigma(\langle\theta_{j_1}^*,x_i\rangle)\sigma(\langle\theta_{j_2}^*,x_i\rangle)\right)_{j_1,j_2}\in\mathbb{R}^{m^*\times m^*}$ which only depends on m^*. This gives the conclusion. □

To ensure the linear convergence in the local region, we evaluate how much $J(\nu_k)$ decrease in each iteration as in the following lemma.

Lemma B.14. *Under Assumptions 4.3, 4.6, 4.6 and 4.7, if $\alpha < \min\{1/8C_1, \rho/C_2, 1/(10C_2), (\lambda/C_F)^2/8\}$, then for any positive integer k, it holds that*

$$J(\nu_{k+1}) - J(\nu_k) \le -\frac{1}{2}g_{\nu_k}^2 + \alpha r_{0,k}\|f^*(\cdot) - f(\cdot,\nu_k)\|_n^2, \tag{53}$$

where $g_{\nu_k}^2 := \int_{\mathbb{S}^{d-1}}(\alpha G_{\nu_k}^2(\theta) + \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2\beta(\theta))\mathrm{d}|\nu_k|$ and $\beta(\theta) = \begin{cases} \beta_{j,k} & \theta = \theta_{j,k} \\ 0 & \text{o.w.}\end{cases}$.

Proof. For a continuous function $\phi : \mathbb{S}^{d-1} \to \mathbb{R}$, we have that

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} \phi(\theta)\mathrm{d}(\nu_{k+1} - \nu_k) &= \sum_{j=1}^M (r_{j,k+1}\phi(\theta_{j,k+1}) - r_{j,k}\phi(\theta_{j,k})) \\ &= \sum_{j=1}^M (r_{j,k+1}\phi(\theta_{j,k+1}) - r_{j,k}\phi(\theta_{j,k+1})) + \sum_{j=1}^M (r_{j,k}\phi(\theta_{j,k+1}) - r_{j,k}\phi(\theta_{j,k}))\end{aligned}$$

In particular if we take $\phi \equiv G_{\nu_k} = \frac{1}{n}\sum_{i=1}^n (f(x_i;\nu_k) - y_i)\sigma(\langle\cdot, x_i\rangle) + \lambda\eta_k(\cdot) \in \partial J(\nu_k)$ where η satisfies $\eta_k(\theta_{j,k+1}) = \mathrm{sgn}(r_{j,k+1})$ for any $j \in [M]$, we have

$$\begin{aligned}J(\nu_{k+1}) - J(\nu_k) &= \int_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\mathrm{d}(\nu_{k+1} - \nu_k) + \frac{1}{2}\|f(\cdot;\nu_{k+1}) - f(\cdot;\nu_k)\|_n^2 \\ &= -\sum_{j=1}^M \alpha|r_{j,k}||G_{\nu_k}(\theta_{j,k})|^2 + \sum_{j=1}^M 2C_1^2\alpha^2|r_{j,k}||G_{\nu_k}(\theta_{j,k})|^2 \\ &\quad + \sum_{j=1}^M r_{j,k}(G_{\nu_k}(\theta_{j,k+1}) - G_{\nu_k}(\theta_{j,k})) + \frac{1}{2}\|f(\cdot;\nu_{k+1}) - f(\cdot;\nu_k)\|_n^2,\end{aligned}$$

where we used Lemma B.3 and $\|G_{\nu_k}\|_\infty \le C_1$ to bound the term related to $\delta r_{j,k}$. For the term $\frac{1}{2}\|f(\cdot;\nu_{k+1}) - f(\cdot;\nu_k)\|_n^2$, by taking $\phi \equiv \sigma(\langle\cdot, x_i\rangle)$ for $(x_i)_{i=1}^n$ and using the 1-Lipschitz continuity of $\sigma(\cdot)$, we have

$$\begin{aligned}\|f(\cdot;\nu_{k+1}) - f(\cdot;\nu_k)\|_n^2 &= \left\|\int_{\mathbb{S}^{d-1}} \sigma(\langle\theta,\cdot\rangle)\mathrm{d}(\nu_{k+1} - \nu_k)\right\|_n^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left(\int_{\mathbb{S}^{d-1}} \sigma(\langle\theta, x_i\rangle)\mathrm{d}(\nu_{k+1} - \nu_k)\right)^2 \\ &\le \left(\sum_{j=1}^M (|r_{j,k+1} - r_{j,k}| + \|\theta_{j,k+1} - \theta_{j,k}\||r_{j,k}|)\right)^2 \\ &\le \left(\int_{\mathbb{S}^{d-1}} (\alpha|G_{\nu_k}(\theta_{j,k})| + \beta_{j,k}\|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k})\|)\mathrm{d}|\nu_k|\right)^2\end{aligned}$$

and this can be upper bounded by

$$\begin{aligned}&\|\nu_k\|_{\mathrm{TV}}^2 \int_{\mathbb{S}^{d-1}} (\alpha|G_{\nu_k}(\theta_{j,k})| + \beta_{j,k}\|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k})\|)^2\mathrm{d}|\nu_k|/\|\nu_k\|_{\mathrm{TV}} \\ &\le 2\left(\frac{C_F}{\lambda}\right)^2 \alpha g_{\nu_k}^2,\end{aligned}$$

by the Jensen's inequality and the inequalities $(a+b)^2 \le 2a^2 + 2b^2$, $\beta_{j,k} \le \alpha$ and $\|\nu_k\|_{\mathrm{TV}} \le C_F/\lambda$ which is derived from Assumption 4.6. Finally we consider the term $\sum_{j=1}^M r_{j,k}(G_{\nu_k}(\theta_{j,k+1}) - G_{\nu_k}(\theta_{j,k}))$. If we take $\alpha < \rho/C_2$, $\theta_{j,k} \in N_j(\rho)$

means that $\theta_{j,k+1}$ remains in H_j in which G_{ν_k} is an (locally) affine function. For $\theta_{j,k} \in N_0$, we note $\|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k})\| \leq \|f^*(\cdot) - f(\cdot, \nu_k)\|_n$ and $\|G_{\nu_k}(\cdot)\|_{\mathrm{Lip}} \leq \|f^*(\cdot) - f(\cdot, \nu_k)\|_n$. Then combining all of them, we get

$$
\begin{aligned}
J(\nu_{k+1}) - J(\nu_k) \leq & -\sum_{j=1}^{M} \alpha |r_{j,k}| |G_{\nu_k}(\theta_{j,k})|^2 + \sum_{j=1}^{M} \frac{1}{4} \alpha |r_{j,k}| |G_{\nu_k}(\theta_{j,k})|^2 \\
& - \sum_{j: \theta_{j,k} \in \cup_{j' \geq 1} N_{j'}(\rho)} \frac{1}{2} \beta_{j,k} |r_{j,k}| \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta_{j,k})\|^2 + \alpha r_{0,k} \|f^*(\cdot) - f(\cdot, \nu_k)\|_n^2 + \frac{1}{4} g_{\nu_k}^2 \\
= & -\frac{1}{2} g_{\nu_k}^2 + \alpha r_{0,k} \|f^*(\cdot) - f(\cdot, \nu_k)\|_n^2,
\end{aligned}
$$

which gives the conclusion. □

Then we give a lower bound of $g_{\nu_k}^2$, in terms of $J(\nu_k) - J^*$.

Proposition B.15 (sharpness inequality)**.** *Under Assumptions 4.3–4.7, there exist constants $J_0 > J^*$ and $\kappa_1 > 0$ such that if $J(\nu_k) \leq J_0$, it holds that*

$$
\kappa_1 (J(\nu_k) - J^*) \leq g_{\nu_k}^2. \tag{54}
$$

To prove this inequality, we prepare a lemma which ensures the sharpness of the gradient in terms of the distance $D_\rho(\nu_k)$.

Lemma B.16. *Under Assumption 4.3–4.7, there exists a constant $J_0 > J^*$ and a constant $C_g > 0$ that depends on α, if $J(\nu_k) \leq J_0$, it holds that*

$$
g_{\nu_k}^2 \geq C_g D_\rho(\nu_k). \tag{55}
$$

Proof. At first, let $0 < \beta_0 < \alpha/4$ and we consider a decomposition

$$
\begin{aligned}
g_{\nu_k}^2 &= \int_{\mathbb{S}^{d-1}} (\alpha G^2(\theta) + \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2 \beta(\theta)) \mathrm{d}|\nu_k| \\
&\geq \int_{\mathbb{S}^{d-1}} (\alpha G^2(\theta) + \beta_0 \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2) \mathrm{d}|\nu_k| - \int_{\mathbb{S}^{d-1} \cup \{\beta(\theta) \leq \beta_0\}} \beta_0 \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2 \mathrm{d}|\nu_k|.
\end{aligned} \tag{56}
$$

For the second term of the right hand side, $\beta_{j,k} \leq \beta_0$ means

$$
\alpha \frac{a_{j,k}^2}{a_{j,k}^2 + \|w_{j,k}\|^2} \leq \beta_0.
$$

By using an inequality $\|w_{j,k}\|^2 \leq |a_{j,k}|^2 + 1$, which is derived from Lemma B.2, we obtain

$$
\alpha \frac{a_{j,k}^2}{2a_{j,k}^2 + 1} \leq \beta_0.
$$

Rearranging this inequality, we get $|a_{j,k}| \leq \sqrt{\frac{2\beta_0}{\alpha - 2\beta_0}} < \sqrt{\beta_0/\alpha}$, therefore $|r_{j,k}| \leq |a_{j,k}|(|a_{j,k}| + 1) \leq 2|a_{j,k}| \leq 2\sqrt{\beta_0/\alpha}$. Then we have

$$
\int_{\mathbb{S}^{d-1} \cup \{\beta(\theta) \leq \beta_0\}} \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2 \beta_0 \mathrm{d}|\nu_k| \leq 2M\beta_0 \frac{\beta_0}{\alpha} \|f(\cdot, \nu_k) - f^*\|_n^2 \leq 4C_\rho M \beta_0 \sqrt{\frac{\beta_0}{\alpha}} D_\rho(\nu_k). \tag{57}
$$

For evaluating the first term of the right hand side of Eq. (56), we have

$$
\int_{\mathbb{S}^{d-1}} (\alpha G^2(\theta) + \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2 \beta_0) \mathrm{d}|\nu_k| \geq \min\{\alpha, \beta_0\} \int_{\mathbb{S}^{d-1}} (G^2(\theta) + \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2) \mathrm{d}|\nu_k|
$$

Now we take $\eta(\theta) \in \|\nu_k\|_{\mathrm{TV}}$, then if $\theta \in N_j(\rho)$, it holds that

$$\begin{aligned}
G_{\nu_k}(\theta) &= \frac{1}{n}\sum_{i=1}^n (f(x_i;\nu_k) - y_i)\sigma(\langle \theta, x_i\rangle) + \lambda\,\mathrm{sgn}(\eta(\theta)) \\
&= \frac{1}{n}\sum_{i=1}^n (f(x_i;\nu^*) - y_i)\sigma(\langle \theta, x_i\rangle) + \lambda\,\mathrm{sgn}(\eta(\theta)) + \sum_{i=1}^n f(x_i;\nu_k - \nu^*)\sigma(\langle \theta, x_i\rangle) \\
&= -\lambda\,\mathrm{sgn}(r_j^*)\langle \theta, \theta_j^*\rangle + \lambda\,\mathrm{sgn}(\eta(\theta)) + \sum_{i=1}^n f(x_i;\nu_k - \nu^*)\sigma(\langle \theta, x_i\rangle)
\end{aligned}$$

and

$$\begin{aligned}
\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta) &= (\mathrm{I}_d - \theta\theta^\mathsf{T})\frac{1}{n}\sum_{i=1}^n (f(x_i;\nu_k) - y_i)x_i \mathbb{1}\{\langle \theta, x_i\rangle \geq 0\} \\
&= (\mathrm{I}_d - \theta\theta^\mathsf{T})\left(\theta_j^* + \frac{1}{n}\sum_{i=1}^n (f(x_i;\nu_k - \nu^*) - y_i)x_i \mathbb{1}\{\langle \theta, x_i\rangle \geq 0\}\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
\int_{\mathbb{S}^{d-1}} (G^2(\theta) + \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2)\mathrm{d}|\nu_k| &\geq \sum_{j=1}^{m^*}\int_{N_j(\rho)} (G^2(\theta) + \|\nabla_{\mathbb{S}^{d-1}} G_{\nu_k}(\theta)\|^2)\mathrm{d}|\nu_k| \\
&= \sum_{j=1}^{m^*}\int_{N_j(\rho)} \Bigg(\underbrace{\left(-\lambda\,\mathrm{sgn}(r_j^*)\langle \theta, \theta_j^*\rangle + \lambda\,\mathrm{sgn}(\eta(\theta))\right)^2 + \|(\mathrm{I}_d - \theta\theta^\mathsf{T})\theta_j^*\|^2}_{\text{(I)}} \\
&\quad + \underbrace{2(-\lambda\,\mathrm{sgn}(r_j^*)\langle \theta, \theta_j^*\rangle + \lambda\,\mathrm{sgn}(\eta(\theta)))\left(\frac{1}{n}\sum_{i=1}^n f(x_i;\nu_k - \nu^*)\sigma(\langle \theta, x_i\rangle)\right)}_{\text{(II)}} \\
&\quad + \underbrace{2\theta_j^{*\mathrm{T}}(\mathrm{I}_d - \theta\theta^\mathsf{T})\frac{1}{n}\sum_{i=1}^n f(x_i;\nu_k - \nu^*)x_i \mathbb{1}\{\langle \theta, x_i\rangle \geq 0\}}_{\text{(III)}} \\
&\quad + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n f(x_i;\nu_k - \nu^*)x_i \mathbb{1}\{\langle \theta, x_i\rangle \geq 0\}\right|^2}_{\text{(IV)}}\Bigg)\mathrm{d}|\nu_k|.
\end{aligned}$$

Now we evaluate each term in the right hand side. The term (I) can be evaluated as

$$\begin{aligned}
\text{(I)} &= \left(-\lambda\,\mathrm{sgn}(r_j^*)\langle \theta, \theta_j^*\rangle + \lambda\,\mathrm{sgn}(\eta(\theta))\right)^2 + \|(\mathrm{I}_d - \theta\theta^\mathsf{T})\theta_j^*\|^2 \\
&= \left(-\lambda\,\mathrm{sgn}(r_j^*)\langle \theta, \theta_j^*\rangle + \lambda\,\mathrm{sgn}(\eta(\theta))\right)^2 + 1 - \langle \theta, \theta_j^*\rangle^2 \\
&\geq \begin{cases} (1+\lambda^2)(1 - \langle \theta, \theta_j^*\rangle) & (\mathrm{sgn}(\eta(\theta)) = \mathrm{sgn}(r_j^*)), \\ \lambda^2 & (\text{otherwise}), \end{cases}
\end{aligned}$$

which gives

$$\int_{N_j(\rho)} \text{(I)}\mathrm{d}|\nu_k| \geq \lambda^2(\Delta\theta_{j,k} + \Delta r_{j,k}).$$

For the terms (II) and (III), we have for any $\theta \in \mathbb{S}^{d-1}$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} f(x_i; \nu_k - \nu^*) x_i \mathbb{1}\{\langle \theta, x_i \rangle \geq 0\} \right\| \leq \|f(\cdot; \nu_k) - f^*\|_n \leq (C_\rho D_\rho(\nu_k))^{\frac{1}{2}},$$

by Lemma B.13. Then it holds that have

$$\begin{aligned}
\left| \int_{N_j(\rho)} \text{(II)} \mathrm{d}|\nu_k| \right| &= 2 \left| \int_{N_j(\rho)} \left(-\lambda \operatorname{sgn}(r_j^*) \langle \theta, \theta_j^* \rangle + \lambda \operatorname{sgn}(\eta(\theta)) \right) \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i; \nu_k - \nu^*) \sigma(\langle \theta, x_i \rangle) \right) \mathrm{d}|\nu_k| \right| \\
&\leq 2\lambda (\Delta\theta_{j,k} + \Delta r_{j,k}) (C_\rho D_\rho(\nu_k))^{\frac{1}{2}}, \\
\left| \int_{N_j(\rho)} \text{(III)} \mathrm{d}|\nu_k| \right| &= 2 \left| \int_{N_j(\rho)} \theta_j^{*\mathrm{T}} (\mathrm{I}_d - \theta\theta^{\mathsf{T}}) \frac{1}{n} \sum_{i=1}^{n} f(x_i; \nu_k - \nu^*) x_i \mathbb{1}\{\langle \theta, x_i \rangle \geq 0\} \mathrm{d}|\nu_k| \right| \\
&\leq 2(\Delta\theta_{j,k} + \Delta r_{j,k}) (C_\rho D_\rho(\nu_k))^{\frac{1}{2}}.
\end{aligned}$$

For the term (IV), we consider the decomposition $f(\cdot; \nu_k) = \sum_{j=0}^{m^*} f_{j,k}$ as Lemma B.13. Then it holds that

$$\begin{aligned}
\sum_{j=1}^{m^*} \int_{N_j(\rho)} \text{(IV)} \mathrm{d}|\nu_k| &= \sum_{j=1}^{m^*} |r_j^*| \left| \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i; \nu_k) - f^*(x_i) \right) x_i \mathbb{1}\{\langle \theta_j^*, x_i \rangle \geq 0\} \right|^2 \\
&\geq \left\| \left(f(\cdot; \nu_k) - f^* \right) \sum_{j=1}^{m^*} \left(|r_j^*|^{\frac{1}{2}} \sigma(\langle \theta_j^*, \cdot \rangle) \right) \right\|_n^2 \\
&\geq c_0 \|f(\cdot; \nu_k) - f^*\|_n^2 \geq c_0 \kappa \sum_{j=1}^{m^*} (\bar{r}_{j,k} - r_j^*)^2 + o(D_\rho(\nu_k)),
\end{aligned}$$

where $c_0 := \min_i \sum_{j=1}^{m^*} \left(|r_j^*|^{\frac{1}{2}} \sigma(\langle \theta_j^*, x_i \rangle) \right) > 0$. Combining the evaluations of (I)-(IV) and Eq. (57), we have

$$g_{\nu_k}^2 \geq \min\{\alpha, \beta_0\} C D_\rho(\nu_k) + \mathrm{o}(D_\rho(\nu_k)) - 4 C_\rho M \beta_0 \sqrt{\frac{\beta_0}{\alpha}} D_\rho(\nu_k),$$

for a some constant $C > 0$. Therefore, with taking sufficiently small β_0 to satisfy $4 C_\rho M \beta_0 \sqrt{\frac{\beta_0}{\alpha}} \leq \min\{\alpha, \beta_0\} C/2$, we have

$$g_{\nu_k}^2 \geq \frac{\min\{\alpha, \beta_0\}}{2} C D_\rho(\nu_k).$$

This gives the conclusion. □

Proof of Proposition B.15. Combining Lemma B.13 and Lemma B.16, we get the conclusion easily. □

Finally, we give the proof which ensures the linear convergence.

Proposition B.17 (Local convergence)**.** *Under Assumption 4.3–4.7, there exist constants $J_0 > J^*$ and $0 < \kappa_0 < 1$ such that if $J(\nu_k) < J_0$, it holds that*

$$J(\nu_{k+1}) - J^* \leq (1 - \kappa_0)(J(\nu_k) - J^*).$$

Proof. Combining Lemma B.14 and Lemma B.13, we have

$$J(\nu_{k+1}) - J(\nu_k) \leq -C_g (J(\nu_k) - J^*) + \mathrm{O}((J(\nu_k) - J^*)^2).$$

Then for sufficiently small $J(\nu_k) - J^*$, there exists a constant $\tilde{\kappa} > 0$ such that

$$J(\nu_{k+1}) - J(\nu_k) \le -C_g(J(\nu_k) - J^*) + \tilde{\kappa}(J(\nu_k) - J^*)^2.$$

By rearranging this inequality, we obtain

$$J(\nu_{k+1}) - J^* \le (1 - \kappa_0)(J(\nu_k) - J^*)$$

for a constant $0 < \kappa_0 < 1$. This gives the conclusion. □

B.4.1. Evaluation of ρ

In the previous section, we have considered a division of $\mathbb{S}^{d-1}$ with the parameter ρ. We have seen that the step-size parameter α needs to be as small as ρ (Lemma B.14). Therefore we need to evaluate how small ρ should be, i.e., how small $\min_j R_j$ will be, which is evaluated by the angles between the sample $(x_i)_{i=1}^n$ and the optimal parameters $(\theta_j^*)_{j=1}^{m^*}$, i.e., $\min_{i,j} \mathrm{dist}(x_i, \theta_j^*)$.

Lemma B.18 (Evaluation of ρ)**.** *Assume that $d \ge 3$, with probability at least $1-\delta$ over the sample $(x_i)_{i=1}^n$, it holds that*

$$\min_{i,j} \left|\mathrm{dist}(\theta_j^\circ, x_i) - \frac{\pi}{2}\right| > \frac{\sqrt{\pi}}{2nm^*}\frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\delta. \tag{58}$$

Proof. Lemma 12 in Cai et al. (2013) shows that for each i, j, $\mathrm{dist}(\theta_j^\circ, x_i)$ is distributed on $[0, \pi]$ with density

$$h(\varphi) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}(\sin\varphi)^{d-2}.$$

This has a maximum value $h(\pi/2) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}$. This leads to

$$\Pr\left(\left|\mathrm{dist}(\theta_j^\circ, x_i) - \frac{\pi}{2}\right| \le t\right) \le \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}2t$$

for any $t \in [0, \frac{\pi}{2}]$. Therefore we have

$$\Pr\left(\min_{i,j}\left|\mathrm{dist}(\theta_j^\circ, x_i) - \frac{\pi}{2}\right| \le t\right) \le \frac{nm^*}{\sqrt{\pi}}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}2t,$$

which gives the conclusion with taking $t = \frac{\sqrt{\pi}}{2nm^*}\frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\delta$. □

This shows that if θ_j° and θ_j^* are sufficiently close for any $j \in [m^*]$, we have $\min_j R_j = \mathrm{O}_p(1/nm^*)$.

B.5. Convergence in $\mathcal{M}(\mathbb{S}^{d-1})$

Theorem 4.8 only ensures the convergence of function value. In this section, we give a convergence in a measure space. At first, we introduce a distance in $\mathcal{M}(\mathbb{S}^{d-1})$.

Definition B.19 (Wasserstein-Fisher-Rao metric (Chizat, 2019))**.**

$$\widetilde{W}_2(\nu_1, \nu_2) := \inf\{W_2(\mu_1, \mu_2) | (\mu_1, \mu_2) \in \mathcal{P}_2(\mathbb{R}_+ \times \mathbb{S}^{d-1})^2 \text{ satisfy } (\mathsf{h}\mu_1, \mathsf{h}\mu_2) = (\nu_1, \nu_2)\}.$$

where $\mathsf{h} : \mathcal{P}_2(\mathbb{R}_+ \times \mathbb{S}^{d-1}) \to \mathcal{M}_+(\mathbb{S}^{d-1})$ is a homogeneous projection operator, i.e., $\mathsf{h}\mu$ satisfies

$$\int_{\mathbb{S}^{d-1}} \phi(\theta)\mathrm{d}(\mathsf{h}\mu)(\theta) = \int_{\mathbb{R}_+ \times \mathbb{S}^{d-1}} r\phi(\theta)\mathrm{d}\mu(r, \theta)$$

for any continuous function $\phi : \mathbb{S}^{d-1} \to \mathbb{R}$.

In above definition, $W_2(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1,\mu_2)} \int_{\mathbb{R}_+ \times \mathbb{S}^{d-1}} \widetilde{\mathrm{dist}}^2((r_1, \theta_1), (r_2, \theta_2)) \mathrm{d}\gamma$, where $\Pi(\mu_1, \mu_2)$ is a set of product measures with marginals μ_1 and μ_2, where $\widetilde{\mathrm{dist}}$ is a distance defined in $\mathbb{R}_+ \times \mathbb{S}^{d-1}$. In this section we especially consider the cone metric (Chizat, 2019), which is expressed by

$$\widetilde{\mathrm{dist}}^2((r_1, \theta_1), (r_2, \theta_2)) = (r_1 - r_2)^2 + 2r_1 r_2(1 - \langle \theta_1, \theta_2 \rangle).$$

Then we can show that a distance between ν_k and ν^* induced by this metric is upper bounded by $D_\rho(\nu_k)$, which we utilize in the proof of local convergence.

Lemma B.20. *Let $\nu^* = \nu^*_+ - \nu^*_-$, $\nu_k = \nu_{k+} - \nu_{k-}$ be Hahn-Jordan decomposition, then it holds that*

$$\max\{\widetilde{W}_2^2(\nu_{k+}, \nu^*_+), \widetilde{W}_2^2(\nu_{k-}, \nu^*_-)\} \le D_\rho(\nu_k).$$

Proof. Remark that $D_\rho(\nu_k)$ is given by

$$D_\rho(\nu_k) = \sum_{j=1}^{m^*} (\bar{r}_{j,k} - r^*_j)^2 + r_{0,k} + \sum_{j=1}^{m^*} (\Delta\theta_{j,k} + \Delta r_{j,k}).$$

Let $I_+ := \{\, j \mid r^*_j > 0\}$, $I_- := \{\, j \mid r^*_j < 0\}$ be subsets of $[m^*]$. Then it holds that $\nu^*_+ = \sum_{j \in I_+} r^*_j \delta_{\theta^*_j}$, $\nu^*_- = \sum_{j \in I_-} r^*_j \delta_{\theta^*_j}$. We only consider the bound of $\widetilde{W}_2^2(\nu_{k+}, \nu^*_+)$ since we can follow the same argument for $\widetilde{W}_2^2(\nu_{k-}, \nu^*_-)$. For each $j \in I_+$, we define a "local positive mass with" $\bar{r}_{j,k,+} = \sum_{\substack{\theta_k \in N_j(\rho) \\ \mathrm{sgn}(r_{j',k}) > 0}} r_{j',k}$. Note that by the definition, $\bar{r}_{j,k} = \bar{r}_{j,k,+} + \Delta r_{j,k}$. For $D_\rho(\nu_k)$ small enough, it holds that $\Delta r_{j,k} \le 1$ for all j. Therefore it holds that

$$\begin{aligned} D_\rho(\nu_k) &\ge \sum_{j \in I_+} ((\bar{r}_{j,k} - r^*_j)^2 + \Delta r^2_{j,k}) + r_{0,k} + \sum_{j \in I_-} \Delta r_{j,k} + \sum_{j \in I_+} \Delta\theta_{j,k} \\ &\ge \sum_{j \in I_+} (\bar{r}_{j,k,+} - r^*_j)^2 + r_{0,k} + \sum_{j \in I_-} \Delta r_{j,k} + \sum_{j \in I_+} \Delta\theta_{j,k}. \end{aligned}$$

Then by using the similar argument as Chizat (2019), we get the conclusion. $\square$

B.6. Evaluation of Estimation Error

In this section, we give a result to the estimation error $\|f(\cdot; \nu_k) - f^\circ\|_{L^2(P_\mathcal{X})}$, i.e., Corollary 4.10. This is a straightforward consequence of Theorem 4.9 and this can be verified by the following Lemmas and Lemma B.20.

Lemma B.21.

$$\|f(\cdot; \nu_k) - f^*\|_\infty \le 2\sqrt{2} \max\{\widetilde{W}_2(\nu_{k+}, \nu^*_+), \widetilde{W}_2(\nu_{k-}, \nu^*_-)\} \tag{59}$$

Proof. Firstly, for any $x \in \mathbb{S}^{d-1}$, $r, r' \in \mathbb{R}_+$ and $\theta, \theta' \in \mathbb{S}^{d-1}$, we have

$$\begin{aligned} |r\sigma(\langle \theta, x \rangle) - r'\sigma(\langle \theta', x \rangle)|^2 &\le 2(r - r')^2 + 2\min\{r, r'\}^2 \|\theta - \theta'\|^2 \\ &\le 2(r - r')^2 + 2rr'(2 - 2\langle \theta, \theta' \rangle) \le 2\widetilde{\mathrm{dist}}^2((r, \theta), (r', \theta')), \end{aligned}$$

where we use $(a + b)^2 \le 2a^2 + 2b^2$ for the first inequality. By using this inequality, let $\tilde{f}(x; (\cdot, \cdot)) : (r, \theta) \mapsto r\sigma(\langle \theta, x \rangle)$, then we have for any $x \in \mathbb{S}^{d-1}$,

$$\|\tilde{f}(x; (\cdot, \cdot))\|_{\mathrm{Lip}} \le \sqrt{2}.$$

Let (μ_{k+}, μ^*_+) and (μ_{k-}, μ^*_-) be any element of $\mathcal{P}_2(\mathbb{R}_+ \times \mathbb{S}^{d-1})^2$ which satisfy $(\mathsf{h}\mu_{k+}, \mathsf{h}\mu^*_+) = (\nu_{k+}, \nu^*_+)$ and $(\mathsf{h}\mu_{k-}, \mathsf{h}\mu^*_-) = (\nu_{k-}, \nu^*_-)$ respectively. By the above inequality, the triangle inequality and the Kantorovich-Rubinstein duality, we have

$$\begin{aligned}
\|f(\cdot;\nu_k)-f^*\|_\infty &= \sup_{x\in\mathbb{S}^{d-1}} |f(x;\nu_k)-f^*(x)| \\
&\le \sup_{x\in\mathbb{S}^{d-1}} |f(x;\nu_{k+})-f(x;\nu_+^*)| + \sup_{x\in\mathbb{S}^{d-1}} |f(x;\nu_{k-})-f(x;\nu_-^*)| \\
&= \sup_{x\in\mathbb{S}^{d-1}} \left|\int_{\mathbb{S}^{d-1}} \sigma(\langle\theta,x\rangle)(\mathrm{d}\nu_{k+}-\mathrm{d}\nu_+^*)(\theta)\right| + \sup_{x\in\mathbb{S}^{d-1}} \left|\int_{\mathbb{S}^{d-1}} \sigma(\langle\theta,x\rangle)(\mathrm{d}\nu_{k-}-\mathrm{d}\nu_-^*)(\theta)\right| \\
&= \sup_{x\in\mathbb{S}^{d-1}} \left|\int_{\mathbb{S}^{d-1}} \tilde{f}(x;(\cdot,\cdot))(\mathrm{d}\mu_{k+}-\mathrm{d}\mu_+^*)(r,\theta)\right| + \sup_{x\in\mathbb{S}^{d-1}} \left|\int_{\mathbb{S}^{d-1}} \tilde{f}(x;(\cdot,\cdot))(\mathrm{d}\mu_{k-}-\mathrm{d}\mu_-^*)(r,\theta)\right| \\
&\le \sqrt{2} \sup_{\|\check{f}\|_{\mathrm{Lip}}\le 1} \left|\int_{\mathbb{R}_+\times\mathbb{S}^{d-1}} \check{f}(\mathrm{d}\mu_{k+}-\mathrm{d}\mu_+^*)(r,\theta)\right| + \sqrt{2} \sup_{\|\check{f}\|_{\mathrm{Lip}}\le 1} \left|\int_{\mathbb{R}_+\times\mathbb{S}^{d-1}} \check{f}(\mathrm{d}\mu_{k-}-\mathrm{d}\mu_-^*)(r,\theta)\right| \\
&= \sqrt{2}(W_2(\mu_{k+},\mu_+^*)+W_2(\mu_{k-},\mu_-^*)) \\
&\le 2\sqrt{2}\max\{W_2(\mu_{k+},\mu_+^*),W_2(\mu_{k-},\mu_-^*)\},
\end{aligned}$$

which gives the conclusion. □

proof of Corollary 4.10. Firstly, we have

$$\|f(\cdot;\nu_k)-f^\circ\|_{L_2(P_\mathcal{X})}^2 \le 2\|f(\cdot;\nu_k)-f(\cdot;\nu^*)\|_{L_2(P_\mathcal{X})}^2 + 2\|f(\cdot;\nu^*)-f^\circ\|_{L_2(P_\mathcal{X})}^2.$$

For the first term, it holds that

$$\|f(\cdot;\nu_k)-f(\cdot;\nu^*)\|_{L_2(P_\mathcal{X})}^2 \le \|f(\cdot;\nu_k)-f(\cdot;\nu^*)\|_\infty^2.$$

The second term can be bounded by

$$\|f(\cdot;\nu^*)-f^\circ\|_{L_2(P_\mathcal{X})}^2 \le \mathrm{O}(m\lambda^2),$$

which is derived by $\sum_{j=1}^m |r_j^\circ - r_j^*|^2 \le \mathrm{O}(m\lambda^2)$ and $\sum_{j=1}^m \mathrm{dist}^2(\theta_j^*,\theta_j^\circ) \le \mathrm{O}(m\lambda^2)$. Then by using Theorem 4.9 for the first term and combining them, we get the conclusion. □

C. Auxiliary Lemmas

In this section we introduce some auxiliary Lemmas.

Lemma C.1. *For $w,\Delta w\in\mathbb{R}^d$, if $w\neq 0$ and $\|\Delta w\|\le\|w\|/2$, it holds that*

$$0 \le \|w-\Delta w\| - \left(\|w\| - \frac{\langle w,\Delta w\rangle}{\|w\|}\right) \le \frac{\|\Delta w\|^2}{\|w\|}$$

Proof. At first, we note that

$$\|w\| - \frac{\langle w,\Delta w\rangle}{\|w\|} \ge \|w\| - \|\Delta w\| \ge \frac{\|\Delta w\|^2}{2\|w\|}.$$

By the straightforward calculation, it holds that

$$\begin{aligned}
\|w-\Delta w\|^2 &= \|w\|^2 - 2\langle w,\Delta w\rangle + \|\Delta w\|^2 \\
&= \left(\|w\| - \frac{\langle w,\Delta w\rangle}{\|w\|}\right)^2 + \|\Delta w\|^2 - \frac{\langle w,\Delta w\rangle^2}{\|w\|^2} \\
&\le \left(\|w\| - \frac{\langle w,\Delta w\rangle}{\|w\|}\right)^2 + \|\Delta w\|^2.
\end{aligned}$$

Then we have

$$\left(\|w-\Delta w\| - \left(\|w\| - \frac{\langle w, \Delta w\rangle}{\|w\|}\right)\right)\left(\|w-\Delta w\| + \left(\|w\| - \frac{\langle w, \Delta w\rangle}{\|w\|}\right)\right) \leq \|\Delta w\|^2.$$

Furthermore, because $\|w-\Delta w\| + \left(\|w\| - \frac{\langle w,\Delta w\rangle}{\|w\|}\right) \geq \|w\|$, we get

$$\|w-\Delta w\| - \left(\|w\| - \frac{\langle w, \Delta w\rangle}{\|w\|}\right) \leq \frac{\|\Delta w\|^2}{\|w\|}$$

and this gives the conclusion. $\square$

Lemma C.2. *For $w, \Delta w \in \mathbb{R}^d$, if $w \neq 0$ and $\|\Delta w\| \leq \|w\|/2$, it holds that*

$$\left\|\frac{w-\Delta w}{\|w-\Delta w\|} - \frac{w}{\|w\|} + \frac{1}{\|w\|}\left(\mathrm{I}_d - \frac{ww^\top}{\|w\|^2}\right)\Delta w\right\| \leq \frac{5\|\Delta w\|^2}{\|w\|^2}.$$

Proof. By putting $\|w-\Delta w\| = \|w\| - \frac{\langle w,\Delta w\rangle}{\|w\|} + \delta$, we have

$$\begin{aligned}
& w - \Delta w - \frac{\|w-\Delta w\|}{\|w\|}\left(w - \left(\mathrm{I}_d - \frac{ww^\top}{\|w\|^2}\right)\Delta w\right) \\
&= w - \Delta w - \frac{\|w\| - \frac{\langle w,\Delta w\rangle}{\|w\|} + \delta}{\|w\|}\left(w - \Delta w + \frac{\langle w, \Delta w\rangle}{\|w\|^2} w\right) \\
&= -\frac{\delta}{\|w\|} w - \frac{\langle w, \Delta w\rangle - \delta\|w\|}{\|w\|^2}\left(\Delta w - \frac{\langle w, \Delta w\rangle}{\|w\|^2} w\right).
\end{aligned}$$

Then by the triangle inequality, an upper bound of the norm of this vector is obtained by

$$|\delta| + \frac{\|\Delta w\| + |\delta|}{\|w\|}\|\Delta w\|.$$

Divided by $\|w-\Delta w\|$ and by using inequalities $\delta \leq \|\Delta w\|/2$ and $\|w-\Delta w\| \geq \|w\|/2$, we get the conclusion. $\square$

Lemma C.3. *Let $\theta, \theta' \in \mathbb{S}^{d-1}$, then it holds that*

$$\frac{\mathrm{dist}^2(\theta,\theta')}{6} \leq 1 - \langle\theta,\theta'\rangle \leq \frac{\mathrm{dist}^2(\theta,\theta')}{2}$$

Proof. Let $d := \mathrm{dist}(\theta,\theta') = \arccos(\langle\theta,\theta'\rangle)$, then we have $\cos d = \langle\theta,\theta'\rangle$. By using the inequality $1 - d^2/2 \leq \cos d \leq 1 - d^2/6$ for $d \in [0,\pi]$, we get the conclusion. $\square$

Lemma C.4. *For $k \in [-1, 1]$, it holds that*

$$\frac{\pi - \arccos(k)}{\pi} k + \frac{\sqrt{1-k^2}}{\pi} \leq \frac{1}{\pi} + \frac{k}{2} + \left(\frac{1}{2} - \frac{1}{\pi}\right) k^2$$

Lemma C.5. *For k_1, k_2 satisfying $r := \sqrt{k_1^2 + k_2^2} \leq 1$, it holds that*

$$\begin{aligned}
\frac{\pi - \arccos(k_1)}{\pi} k_1 + \frac{\sqrt{1-k_1^2}}{\pi} + \frac{\pi - \arccos(k_2)}{\pi} k_2 + \frac{\sqrt{1-k_2^2}}{\pi} - \frac{\pi - \arccos(r)}{\pi} r + \frac{\sqrt{1-r^2}}{\pi} - \frac{1}{\pi} \\
\leq \frac{1}{2}(k_1 + k_2 - r).
\end{aligned}$$

Proof. Let $g(k_1, k_2) := (\mathrm{LHS}) - (\mathrm{RHS})$. Simple calculation shows that g is even w.r.t. both of k_1 and k_2. Therefore we only need to consider the case $k_1 \geq 0, k_2 \geq 0$. Let $k_1 = r\cos\theta, k_2 = r\sin\theta$ $(0 \leq r \leq 1, 0 \leq \theta \leq \pi/2)$. This gives

$$\begin{aligned}\tilde{g}(r,\theta) &:= g(k_1,k_2)\\ &= \frac{\pi - \arccos(r\cos\theta)}{\pi} r\cos\theta + \frac{\sqrt{1-r^2\cos^2\theta}}{\pi} + \frac{\pi - \arccos(r\sin\theta)}{\pi} r\sin\theta + \frac{\sqrt{1-r^2\sin^2\theta}}{\pi}\\ &\quad - \frac{\pi - \arccos(r)}{\pi} r + \frac{\sqrt{1-r^2}}{\pi} - \frac{1}{\pi} - \frac{r}{2}(\cos\theta + \sin\theta - 1).\end{aligned}$$

For any fixed $0 \leq \theta \leq \pi/2$, we have

$$\begin{aligned}\frac{\partial\tilde{g}}{\partial r}(r,\theta) &= \frac{\pi - \arccos(r\cos\theta)}{\pi}\cos\theta + \frac{\pi - \arccos(r\sin\theta)}{\pi}\sin\theta - \frac{\pi - \arccos(r)}{\pi} - \frac{1}{2}(\cos\theta + \sin\theta - 1).\\ \frac{\partial^2\tilde{g}}{\partial r^2}(r,\theta) &= \frac{\cos^2\theta}{\pi\sqrt{1-r^2\cos^2\theta}} + \frac{\sin^2\theta}{\pi\sqrt{1-r^2\sin^2\theta}} - \frac{1}{\pi\sqrt{1-r^2}}\\ &\leq \frac{\cos^2\theta + \sin^2\theta}{\pi\sqrt{1-r^2}} - \frac{1}{\pi\sqrt{1-r^2}} = 0.\end{aligned}$$

Therefore $\frac{\partial\tilde{g}}{\partial r}(r,\theta)$ is monotonically decreasing w.r.t. r and $\frac{\partial\tilde{g}}{\partial r}(0,\theta) = 0$, then $\tilde{g}$ is also monotonically decreasing. This means that g takes maximum value at $(k_1, k_2) = (0,0)$. Since $g(0,0) = 0$, we get the conclusion. ☐

Lemma C.6. *For k_1, k_2 satisfying $r := \sqrt{k_1^2 + k_2^2} \leq 1$, it holds that*

$$-\arccos(k_1) - \arccos(k_2) + \arccos(r) + \pi/2 \leq k_1 + k_2 - r.$$

Proof. Let $g(k_1, k_2) := (\mathrm{LHS}) - (\mathrm{RHS})$. It is sufficient to consider the case $k_1 \geq 0, k_2 \geq 0$ because it holds that

$$g(k_1,k_2) - g(-k_1,k_2) = \pi - 2\arccos(k_1) - 2k_1 \geq 0$$

for $k_1 \geq 0$ and arbitrary k_2. The same argument follows with swapping k_1 and k_2. Let $k_1 = r\cos\theta, k_2 = r\sin\theta$ $(0 \leq r \leq 1, 0 \leq \theta \leq \pi/2)$. We consider a function

$$\begin{aligned}\tilde{g}(r,\theta) &:= g(k_1,k_2)\\ &= -\arccos(r\cos\theta) - \arccos(r\sin\theta) + \arccos(r) - \frac{\pi}{2} - r(\cos\theta + \sin\theta - 1).\end{aligned}$$

or any fixed $0 \leq \theta \leq \pi/2$, we have

$$\begin{aligned}\frac{\partial g}{\partial r}(r,\theta) &= \frac{\cos\theta}{\sqrt{1-r^2\cos^2\theta}} + \frac{\sin\theta}{\sqrt{1-r^2\sin^2\theta}} - \frac{1}{\sqrt{1-r^2}} - (\cos\theta + \sin\theta - 1)\\ \frac{\partial^2\tilde{g}}{\partial r^2}(r,\theta) &= \frac{r\cos^2\theta}{\sqrt{1-r^2\cos^3\theta}} + \frac{r\sin^3\theta}{\sqrt{1-r^2\sin^2\theta}} - \frac{r}{\sqrt{1-r^2}}\\ &\leq \frac{r\cos^3\theta}{\sqrt{1-r^2}} + \frac{r\sin^3\theta}{\sqrt{1-r^2}} - \frac{r}{\sqrt{1-r^2}} \leq 0.\end{aligned}$$

Therefore $\frac{\partial\tilde{g}}{\partial r}(r,\theta)$ is monotonically decreasing w.r.t. r and $\frac{\partial\tilde{g}}{\partial r}(0,\theta) = 0$, then $\tilde{g}$ is also monotonically decreasing. This means that g takes maximum value at $(k_1, k_2) = (0,0)$. Since $g(0,0) = 0$, we get the conclusion. ☐

Lemma C.7. *For $k_1 \geq 0, k_2 \geq 0$ satisfying $\sqrt{k_1^2 + k_2^2} \leq 1$, it holds that*

$$\arccos(k_1) + \arccos(k_2) \leq \arccos(\sqrt{k_1^2 + k_2^2}) + \pi/2.$$

Proof. We have

$$\cos(\arccos(k_1) + \arccos(k_2)) = k_1 k_2 - \sqrt{1-k_1^2}\sqrt{1-k_2^2}.$$
$$\cos\left(\arccos(\sqrt{k_1^2+k_2^2}) + \frac{\pi}{2}\right) = -\sqrt{1-k_1^2-k_2^2}$$

Then it holds that $\cos\left(\arccos(\sqrt{k_1^2+k_2^2} + \frac{\pi}{2})\right) \leq \cos(\arccos(k_1) + \arccos(k_2))$, because $-1 + k_1^2 + k_2^2 \leq 0$ and

$$\begin{aligned} 1 - k_1^2 - k_2^2 - (k_1 k_2 - \sqrt{1-k_1^2}\sqrt{1-k_2^2})^2 &= 2k_1k_2(\sqrt{1-k_1^2}\sqrt{1-k_2^2} - k_1k_2) \\ &\leq 2k_1k_2(k_1k_2 - k_1k_2) = 0. \end{aligned}$$

By the fact $0 \leq \arccos(k_1) + \arccos(k_2) \leq \pi$ and $\pi/2 \leq \arccos(\sqrt{k_1^2+k_2^2}) + \pi/2 \leq \pi$, we get the conclusion. □

Lemma C.8 (Matrix Bernstein (Tropp, 2015))**.** *Let $A_1, \ldots, A_n$ be independent random $d \times d$ matrices with $\mathbb{E}[A_i] = 0_d$ and $\|A_i\|_{\mathrm{op}} \leq L$ for some $L > 0$. then for any $t \geq 0$,*

$$\Pr\left(\left\|\sum_i^n A_i\right\|_{\mathrm{op}} \geq t\right) \leq 2d\exp\left(\frac{-t^2}{2(V + Lt/3)}\right)$$

where $V := \left\|\sum_{i=1}^n \mathbb{E}[A_i^2]\right\|_{\mathrm{op}}$.