Proceedings of Machine Learning Research vol 134:1–46, 2021 34th Annual Conference on Learning Theory

The Effects of Mild Over-parameterization on the Optimization Landscape of Shallow ReLU Neural Networks

Itay Safran* SAFRAN.ITAY@GMAIL.COM
Gilad Yehudai* GILAD.YEHUDAI@WEIZMANN.AC.IL
Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the effects of mild over-parameterization on the optimization landscape of a simple ReLU neural network of the form $\mathbf{x} \mapsto \sum_{i=1}^{k} \max\{0, \mathbf{w}_i^\top \mathbf{x}\}$, in a well-studied teacher-student setting where the target values are generated by the same architecture, and when directly optimizing over the population squared loss with respect to Gaussian inputs. We prove that while the objective is strongly convex around the global minima when the teacher and student networks possess the same number of neurons, it is not even *locally convex* after any amount of over-parameterization. Moreover, related desirable properties (e.g., one-point strong convexity and the Polyak-Łojasiewicz condition) also do not hold even locally. On the other hand, we establish that the objective remains one-point strongly convex in *most* directions (suitably defined), and show an optimization guarantee under this property. For the non-global minima, we prove that adding even just a single neuron will turn a non-global minimum into a saddle point. This holds under some technical conditions which we validate empirically. These results provide a possible explanation for why recovering a global minimum becomes significantly easier when we over-parameterize, even if the amount of over-parameterization is very moderate.

1. Introduction

In recent years, a spur of theoretical papers studied how the training of neural networks benefits from over-parameterization, namely the use of more neurons than needed to express a good predictor (e.g., (Safran and Shamir, 2016; Du et al., 2018; Safran and Shamir, 2017; Allen-Zhu et al., 2018; Daniely, 2017; Li and Liang, 2018; Cao and Gu, 2019; Andoni et al., 2014; Jacot et al., 2018)). The vast majority of these papers focus on settings where a large amount of over-parameterization is needed (e.g., polynomial in some natural problem parameters). However, empirical studies such as in (Livni et al., 2014; Safran and Shamir, 2017) indicate that in many cases, very mild over-parameterization is required to successfully reach a global optimum, and sometimes adding even one or two neurons is enough. The aim of this paper is to theoretically study the effect of such mild over-parameterization.

Specifically, we focus on a simple and well-studied student-teacher setting, where the labels are generated by a teacher network composed of a sum of k neurons, and learned by a student network of the same architecture with n neurons, using the squared loss with respect to some input distribution $\mathcal{D}$:

$$F(\mathbf{w}) \;:=\; F(\mathbf{w}_1,\ldots,\mathbf{w}_n) \;=\; \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\sum_{i=1}^{n}\sigma(\langle\mathbf{w}_i,\mathbf{x}\rangle)-\sum_{i=1}^{k}\sigma(\langle\mathbf{v}_i,\mathbf{x}\rangle)\right)^2\right]. \qquad (1)$$

* equal contribution

In the above, $\sigma : \mathbb{R} \to \mathbb{R}$ is some univariate activation function. This objective has been studied in quite a few recent works (e.g., (Zhong et al., 2017; Tian, 2017; Soltanolkotabi et al., 2019; Yehudai and Shamir, 2020; Li and Yuan, 2017; Arjevani and Field, 2019, 2020)), perhaps most commonly when $\mathcal{D}$ is a standard Gaussian, and σ is the ReLU function. This will also be the setting we focus on in this paper.

Our paper is motivated by the empirical findings in (Safran and Shamir, 2017). In that paper, the authors prove that the objective above possesses local minima which are not global, and empirically show that gradient descent with standard initialization does tend to get stuck in them when $n = k$. However, significantly fewer local minima are encountered already when $n = k + 1$, and when $n = k + 2$, no local minima were encountered at all for values of $n \leq 20$ (see Table 2 in (Safran and Shamir, 2017)). Despite the progress made in understanding the loss surface and the dynamics of optimization techniques on the objective in Eq. (1), to the best of our knowledge, current deep learning theory is unable to explain why such mild over-parameterization helps gradient methods to recover the global minimum in this setting. This leads us to the following question:

> *What are the geometrical effects of mild over-parameterization on the objective function, which facilitate the use of common optimization techniques for recovering the global minimum?*

In this paper, we take a few steps in understanding the above question in the context of Eq. (1), under the standard setting where $\mathcal{D}$ is a standard Gaussian distribution, the $\mathbf{v}_i$'s are orthogonal and of unit norm, and σ is the popular ReLU activation function. Our contributions are as follows:

- First, we provide a full characterization of all twice differentiable points and all global minima of the objective (Thm. 2 and Lemma 1). We then formally prove that *without* over-parameterization ($n = k$), the objective is strongly convex in a neighborhood of every global minimum (Thm. 3). This property ensures that initializing close enough to the global minima (e.g. using a tensor initialization (Zhong et al., 2017)), gradient descent with small enough step sizes will converge to it. We note that this in itself is not too surprising, and that a similar result was shown in (Zhong et al., 2017; Li and Yuan, 2017) for a slightly different setting.

- Next, we prove that perhaps surprisingly, in the over-parameterization regime ($n > k$) the local geometry around global minima changes significantly: The objective is not even locally *convex* around global minima (Thm. 4). Moreover, we study other commonly used geometrical properties such as one-point strong convexity (also known as strong star-convexity) and Polyak-Łojasiewicz (PL) condition (see (Karimi et al., 2016)) and show that these also do not hold, even locally, around global minima (Thm. 7 and Thm. 9).

- On the flip side, we show that our objective is one-point strongly convex *in most directions* – that is, there is a significant set of points around the global minima that satisfy one-point strong convexity (Thm. 10). This allows us to prove an optimization guarantee using gradient descent with small perturbations, for functions that satisfy this property in a simplified setting and under a certain technical assumption (Thm. 13).

- Turning to the non-global minima, we prove that for *any* such point, a slight over-parameterization consisting of 'splitting' a neuron into two neurons (having the same angle and summing to the original neuron) results in turning the non-global minimum into a saddle point with a direction of descent (Thm. 14). This holds under a technical condition on the norm of the neurons at the local minima, which we justify empirically. This result demonstrates how even a tiny amount of over-parameterization helps eliminate non-global minima.

The remainder of the paper is structured as follows: After discussing related work, we formalize our setting in Sec. 2, and introduce relevant definitions and notations. Next, Sec. 3 investigates properties relating to the global minima of our objective, which includes their general form and the geometry of the objective function around them. Lastly, Sec. 4 studies the non-global minima of our objective, showing when can we guarantee that splitting local minima will become a saddle points.

1.1. Related Work

Over-Parameterization. It was shown empirically that over-parameterized networks are easier to train, e.g. in (Livni et al., 2014; Safran and Shamir, 2017). Over-parameterization was extensively studied theoretically in several contexts and architectures, such as (Du et al., 2018; Allen-Zhu et al., 2018; Daniely, 2017; Li and Liang, 2018; Cao and Gu, 2019; Andoni et al., 2014; Yehudai and Shamir, 2019; Ghorbani et al., 2019,; Kamath et al., 2020; Allen-Zhu and Li, 2019). In particular, one very popular line of works argue that sufficiently over-parameterized networks behave similarly to kernel methods (in particular, the neural tangent kernel) or random feature methods. However, these approaches only apply for a very large amount of over-parameterization, as shown in several recent papers Yehudai and Shamir (2019); Allen-Zhu and Li (2019); Kamath et al. (2020). Thus, they cannot be used to explain why adding just a few neurons can significantly increase the probability of converging to a global minimum. In contrast, our results hold for *any* amount of over-parameterization. Notably, in Yehudai and Shamir (2019) it was shown that kernel methods (such as the NTK) cannot explain learnability of even a single ReLU neuron. This means that NTK cannot explain learnability of gradient descent on Eq. (1), even for the simple case of $k = 1$, unless n is exponential in the input dimension.

Over-Parameterization beyond NTK regimes. Several papers considered theoretical analysis of over-parameterized models beyond the NTK regime. Li et al. (2020) provide recovery and generalization guarantees for an objective similar to Eq. (1), however their result only guarantees convergences to a solution with loss of about $1/d$ (d being the input dimension) and not to arbitrarily small loss, and their analysis strongly relies on the symmetry of the teacher network, and therefore cannot be generalized to cases where this symmetry breaks. Allen-Zhu and Li (2019) show an analysis that goes beyond NTK, where the target network is a one layer ResNet. Daniely and Malach (2020) provide an optimization guarantee on the problem of learning parity functions under some specific distribution using a 2-layer neural network. With that said, providing optimization guarantees for Eq. (1) for general n and k largely remains an open question.

Previous works on Eq. (1) Several works studied Eq. (1) under different assumptions such as Tian (2017); Soltanolkotabi (2017); Zhong et al. (2017); Yehudai and Shamir (2020); Li et al. (2020). In Yehudai and Shamir (2020) the authors study the case of $n = k = 1$, and show that even in this simple regime there exists distributions and activations in which gradient methods are unable to learn. On the other hand, they show that under mild assumptions on the activation and distribution it is possible to guarantee convergence to the global optimum, although in this simple case there are no non-global minima (there is a non-differentiable saddle point at the origin). This analysis does not generalize even to the case of $n = k = 2$. In Zhong et al. (2017) the authors give optimization guarantees for the case of $n = k$ for general k, where $\mathcal{D}$ is standard Gaussian and some assumptions on σ (which includes ReLU). Their method is to show that locally around global minima the objective is strongly convex, and use tensor initialization to initialize close enough to the global minimum. We prove a similar theorem (Thm. 3), although there are a couple of small differences: The objective is a bit different, because in Zhong et al. (2017) the authors consider an empirical loss over a finite set of examples drawn i.i.d from $\mathcal{N}(0, I)$, whereas we consider the population loss. Moreover, we state an explicit numerical lower bound on the minimal eigenvalue of the Hessian at the minimum. On the other hand, Zhong et al. (2017) show the result for a general class of activation functions (including ReLU) and we show it specifically for the ReLU activation. They also specify how large the open neighborhood for which the objective is strongly convex, while we only state that there exists an open neighborhood without guarantees on its size. In any case, we note that this is not a main result of our paper, as we focus more on the over-parameterized case and this theorem is given mainly as a comparison to how over-parameterization significantly changes the optimization landscape.

A similar analysis for the case of $n = k$ is done in Li and Yuan (2017) where the authors consider an architecture where the target neurons are close to unit vectors, and they show that the objective is one-point strongly convex (as opposed to strongly convex) around the global minimum. In Arjevani and Field (2019, 2020,) the authors study the properties of local minima of Eq. (1) in the case of $n = k$, standard Gaussian distribution and ReLU activation. They identify certain symmetries of the local minima and utilize them to characterize a certain family of local minima.

In Jin et al. (2017) the authors show how perturbed gradient descent can help in escaping saddle points. In our paper we also analyze perturbed gradient descent, and show that it can help to ensure convergence to a global minima, even when standard convexity-like properties (e.g. one-point strong convexity and PL) do not apply to the optimization landscape.

2. Preliminaries

Terminology and Notation. We use $[n]$ as shorthand for $\{1, \dots, n\}$. We denote the ReLU function ($z \mapsto \max\{0, z\}$) by $[\cdot]_+$. We denote vectors using bold-faced letters (e.g. $\mathbf{w}$). We let barred bold-faced letters denote vectors normalized to unit length (i.e. $\bar{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$). Given two non-zero vectors $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, we denote the angle between them using $\theta_{\mathbf{w},\mathbf{v}} = \arccos\left(\bar{\mathbf{w}}^\top \bar{\mathbf{v}}\right)$. Unless stated otherwise, we denote by $\|\cdot\|$ the standard Euclidean norm. We denote the matrix with all zero entries of size $m \times n$ by $\mathbf{0}_{m\times n}$. For $\mathbf{w}_1, \dots, \mathbf{w}_n \in \mathbb{R}^d$ denote by $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n) \in \mathbb{R}^{n\cdot d}$ their concatenation. For symmetric matrices A, B we say that $A \succeq B$ if $A - B$ is positive semi-definite (PSD). Recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ that is twice continuously differentiable is said to be strongly convex in $A \subseteq \mathbb{R}^d$ iff there is a constant $\lambda > 0$ such that $\nabla^2 f(\mathbf{x}) \succeq \lambda I$ for any $\mathbf{x} \in A$. It is convex if the above holds for $\lambda = 0$.

Setting. In this paper we study a simple network in a student-teacher setting, assuming our data have a standard Gaussian distribution. In more detail, we fix the vectors in the teacher network $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^d$, and the population objective is:

$$F(\mathbf{w}_1^n) = \mathbb{E}_{x\sim\mathcal{N}(0,I)}\left[\left(\sum_{i=1}^n [\langle \mathbf{w}_i, \mathbf{x}\rangle]_+ - \sum_{i=1}^k [\langle \mathbf{v}_i, \mathbf{x}\rangle]_+\right)^2\right]. \tag{2}$$

Throughout this paper we always assume that $d \geq k$ (to model a high-dimensional setting). We also assume for simplicity that the target vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ are orthogonal with $\|\mathbf{v}_i\| = 1$ for

$i \in [k]$. This assumption is also made in Safran and Shamir (2017), and approximately holds if $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are chosen uniformly at random from the unit sphere and the dimension is high enough. We conjecture that all the results in the paper can be extended to general target vectors, and leave it to future work.

Basic Properties of the Objective Function. For a standard Gaussian distribution, the objective function in Eq. (2) can be written down in closed form (without expectation terms). Moreover, it is continuously differentiable if $\mathbf{w}_i \neq 0$ for every $i \in [n]$, with explicit expressions for the Gradient and Hessian at any point (see (Cho and Saul, 2009; Brutzkus and Globerson, 2017; Safran and Shamir, 2017)). In particular, we will need an explicit expression for the Hessian from (Safran and Shamir, 2017, Section 4.1.1). For completeness we include the formal statement in Thm. 17, from which we immediately get that the objective is twice continuously differentiable for every $\mathbf{w}_1^n = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$ where $\mathbf{w}_i \neq 0$ for every $i \in [n]$ and there are no two $\mathbf{w}_i, \mathbf{w}_j$ with $\theta_{\mathbf{w}_i,\mathbf{w}_j} \in \{0, \pi\}$. To complete the picture we show that even when $\theta_{\mathbf{w}_i,\mathbf{w}_j} \in \{0, \pi\}$ for some $i \neq j$ the Hessian is well defined and continuous. The formal proof can be found in Appendix A.

Lemma 1 *$F(\mathbf{w}_1^n)$ is twice continuously differentiable at any $\mathbf{w}_1^n = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$ such that $\mathbf{w}_i \neq \mathbf{0}$ for all $i \in [n]$.*

3. Effects of Over-parameterization on the Global Minima

In this section we study the local geometric properties of the global minima of the objective in Eq. (2). We first characterize all the global minima of the objective $F(\mathbf{w}_1^n)$ for any $n \geq k$.

Theorem 2 *Suppose $\mathbf{w}_1^n = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$ is a global minimum of the objective in Eq. (2). Then there exists a partition $\dot{\bigcup}_{i=1}^k I_i = [n]$ and $\alpha_1, \ldots, \alpha_n \geq 0$ satisfying $\sum_{j \in I_i} \alpha_j = 1$ and $\mathbf{w}_j = \alpha_j \mathbf{v}_i$ for all $i \in [k]$ and $j \in I_i$.*

The full proof can be found in Appendix B. Thm. 2 states that for a global minimum, each vector $\mathbf{w}_i$ must be equal to some target vector $\mathbf{v}_j$ times some positive constant α_i. In addition, the sum of all the constants, for all the $\mathbf{w}_i$ in the direction of some $\mathbf{v}_j$ must be equal to 1. In particular, for the case of $n = k$ we get that the only global minima are those that for each target vector $\mathbf{v}_j$ there is exactly one $\mathbf{w}_i$ for which $\mathbf{w}_i = \mathbf{v}_j$, hence there are exactly $n!$ isolated global minima. For the case of $n > k$ there is a manifold consisting of infinitely many global minima. For example, if $n = k + 1$, then the following is a global minimum for every $\alpha \in [0, 1]$: $\mathbf{w}_1 = \mathbf{v}_1, \ldots,\ \mathbf{w}_{n-1} = \mathbf{v}_{n-1},\ \mathbf{w}_n = \alpha \mathbf{v}_n,\ \mathbf{w}_{n+1} = (1 - \alpha)\mathbf{v}_n$.

Combining Thm. 2 and Lemma 1, we have a full characterization of all (twice continuously) differentiable global minima of the objective $F(\mathbf{w})$ for general $n \geq k$. More specifically, all minima that admit the form of Thm. 2 and in addition satisfy that $\mathbf{w}_i \neq 0$ for all $i \in [n]$ are differentiable. In this section we will study local geometric properties of the differentiable local minima, distinguishing between two cases: exact parameterization ($n = k$) and over-parameterization ($n > k$).

3.1. Exact Parameterization

We first consider the case of exact parameterization, where the labels are created by a teacher network with k neurons, and learned by a student network with k neurons. Even though the objective $F(\mathbf{w}_1^k)$ in this case is not convex (at least for $k \geq 2$, as there are $k!$ isolated global minima), we will show that locally around each global minimum it is actually strongly convex.

Theorem 3 *Suppose $n = k$. For every global minimum of the objective $F(\mathbf{w}_1^k)$ in Eq. (2) we have that $\nabla^2 F(\mathbf{w}_1^k) \succeq \left(\frac{1}{4} - \frac{1}{2\pi}\right) I$. Moreover, the objective is strongly convex around an open neighborhood of any global minimum.*

Note that in the case of $n = k$, by Thm. 2 all the global minima are differentiable. The proof idea behind Thm. 3 is straightforward. The Hessian at the global minimum can be divided into a sum of two matrices, and we lower bound the smallest eigenvalue of these two matrices. Note that since the objective is twice continuously differentiable around any global minimum (in the case of $n = k$), and that the eigenvalue of a matrix is a continuous function we immediately get that in an open neighborhood of the global minimum all the eigenvalues of the Hessian are positive, hence the objective is locally strongly convex.

As discussed in the related work section, a similar result was shown in (Zhong et al., 2017) for a slightly different setting. Although this result might give hope that such properties are also preserved when over-parameterizing, as we will show in the next subsection, the over-parameterized case has a completely different geometry. Thus, this kind of analysis is specific for exact parameterization.

3.2. Over-Parameterization

In the exact parameterization case, we showed that around the global minima the objective is strongly convex. Since empirically, over-parameterization tends to improve training performance, we might expect that it improves or at least maintains favorable geometric properties around the global minima. However, we now prove that perhaps surprisingly, under any amount of over-parameterization, the objective in Eq. (2) is not even *locally convex* around any differentiable global minimum:

Theorem 4 *Assume that $n > k$ and $d > 1$ (recall that $d \geq k$, hence this assumption is trivially true for $k > 1$). Then in every neighborhood of a differentiable global minimum of Eq. (2) there is a point at which the Hessian of the objective has a negative eigenvalue.*

Since convexity of a differentiable function requires the Hessian to be positive semidefinite, we get that no local convexity property can hold. We note that the theorem's assumptions are mild, since by Thm. 2, the objective function is typically differentiable at a global minimum and its neighborhood. To provide some intuition how a global minimum without a convex neighborhood might look like, see an example (using a different function) in Fig. 2 in the Appendix B.3.

3.3. One-Point Strong Convexity and the PL condition

Instead of having convexity with respect to all directions, it may be enough from an optimization point of view to have convexity in the direction of the global minimum. This motivates the following well-known definition (see e.g. Lee and Valiant (2016); Kleinberg et al. (2018)):

Definition 5 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. $f(\mathbf{x})$ is said to be* ***one-point strongly convex*** *(OPSC) in an open neighborhood $A \subseteq \mathbb{R}^d$ with respect to a local minimum $\mathbf{y}^* \in A$ if there exists $\lambda > 0$ such that for every $\mathbf{x} \in A$: $\frac{1}{\|\mathbf{x}-\mathbf{y}^*\|^2}\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y}^* \rangle \geq \lambda$. If we further assume that $f(\mathbf{x})$ is twice differentiable, then it is OPSC in $A \subseteq \mathbb{R}^d$ if there exists $\lambda > 0$ such that for every $\mathbf{x} \in A$: $\frac{1}{\|\mathbf{x}-\mathbf{y}^*\|^2}(\mathbf{x} - \mathbf{y}^*)^\top \nabla^2 f(\mathbf{x})(\mathbf{x} - \mathbf{y}^*) \geq \lambda$, where $\nabla^2 f(\mathbf{x})$ is the Hessian of f at $\mathbf{x}$. We call such λ the OPSC coefficient.*

The Hessian definition of one-point strong convexity can be easily derived from the gradient definition, in the same manner that the Hessian definition of strong convexity is derived from the gradient definition of strong convexity for twice continuously differentiable functions. In previous works it was shown that although an objective is not strongly-convex, it may be OPSC which is enough to show convergence to a minimum for certain local search algorithms (see e.g. Li and Yuan (2017)). Intuitively, this is because if $\mathbf{y}^*$ is a local minimum, the definition above implies that the gradient at $\mathbf{x}$ is correlated with the direction to the minimum, and increases with the distance from $\mathbf{y}^*$. We note that one point convexity (i.e., taking $\lambda = 0$) is not enough, as it may imply that the gradient is arbitrarily close to being orthogonal to the direction of the minimum (see also (Lee and Valiant, 2016) for a discussion).

Unfortunately, we cannot really hope for OPSC for the objective in Eq. (2) in the over-parameterized case. The reason is that Thm. 2 reveals that in this case there is a connected manifold of global minima (on which the function is flat), instead of isolated minima as in the exact parameterization case.

Recall that if $n > k$ then the global minima form along a line on which each point is a global minimum (recall the discussion after Thm. 2). One alternative formulation is to define OPSC on any point which is not a global minimum, but the problem of defining OPSC with respect to which point still stands. One way to overcome this problem is by considering OPSC with respect to a global minimum, only in directions which lead away from nearby global minima. This is formalized in the following definition (see Fig. B.3 in the supplementary material for an intuition):

Definition 6 *Let $\tilde{\mathbf{w}}_1^n = (\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_n)$ and $\epsilon > 0$. An **ϵ-orthogonal Neighborhood** of $\tilde{\mathbf{w}}_1^n$ is:*

$$U_\epsilon^\perp(\tilde{\mathbf{w}}_1^n) = \{\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n) : \ \forall i \in [n], \ \mathbf{w}_i - \tilde{\mathbf{w}}_i \perp \tilde{\mathbf{w}}_i, \ \|\mathbf{w}_i - \tilde{\mathbf{w}}_i\| \le \epsilon\}.$$

We refer to an ϵ-neighborhood (i.e. not orthogonal) of $\tilde{\mathbf{w}}_1^n$ as

$$U_\epsilon(\tilde{\mathbf{w}}_1^n) = \{\mathbf{w}_1^n : \forall i \in [n], \ \|\mathbf{w}_i - \tilde{\mathbf{w}}_i\| \le \epsilon\} \ .$$

Note that this is different from the "Standard" definition of a neighborhood of $\tilde{\mathbf{w}}_1^n$, since here we allow each vector $\mathbf{w}_i$ to be at distance ϵ from its corresponding $\tilde{\mathbf{w}}_i$. We could hope that the objective in Eq. (2) is OPSC at least in an ϵ-orthogonal neighborhood of a global minimum, however this is not the case as shown in the following theorem.

Theorem 7 *Assume $n > k$, let $\epsilon > 0$ and let $\tilde{\mathbf{w}}_1^n = (\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_n)$ be a differentiable global minimum of Eq. (2). Then the objective is not OPSC with respect to $\tilde{\mathbf{w}}_1^n$, even in an ϵ-orthogonal neighborhood of $\tilde{\mathbf{w}}_1^n$.*

The theorem shows that the geometrical properties of our objective, although similar in some senses to the example of $f(x, y) = x^2y^2$, are still much more complex.

The full proof of the theorem can be found in Appendix B.3. The intuition for the proof of the above theorem is the following: Assume that at the global minimum $\tilde{\mathbf{w}}_1$ and $\tilde{\mathbf{w}}_2$ are both directed in the same target vector $\mathbf{v}_1$, i.e. $\tilde{\mathbf{w}}_1 = \alpha_1\mathbf{v}_1$ and $\tilde{\mathbf{w}}_2 = \alpha_2\mathbf{v}_1$ for some $\alpha_1, \alpha_2 > 0$. We define a new point close to $\tilde{\mathbf{w}}$ by taking $\mathbf{w}_1 = \tilde{\mathbf{w}}_1 + \epsilon\mathbf{u}$ and $\tilde{\mathbf{w}}_2 = \mathbf{w}_2 - \epsilon\mathbf{u}$ where $\mathbf{u} \perp \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2$, and leave all the other vectors the same, thus $\mathbf{w}_1^n \in U_\epsilon^\perp(\mathbf{w}_1^n)$. Intuitively, in the objective there are terms that to minimize them it is needed to make the $\mathbf{w}_i$ close to the $\mathbf{v}_j$, and other terms that will be minimized if the $\mathbf{w}_i$'s are far apart. Since we haven't changed any of the vectors that are directed at the target vectors $\mathbf{v}_2, \dots, \mathbf{v}_k$, then most cancel out. Actually, the only terms that remain are the ones that are minimized when $\mathbf{w}_1, \mathbf{w}_2$ are close to $\mathbf{v}_1$, and the ones that minimized when $\mathbf{w}_1$ and $\mathbf{w}_2$ are far apart from one another. But because of the way we defined $\mathbf{w}_1^n$, these terms also *almost* cancel out - they are of magnitude $O(\epsilon)$.

Another useful property which became popular in recent years is the **Polyak- Łojasiewicz** (PL) condition (Polyak (1963); Lojasiewicz (1963)):

Definition 8 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function, and let f^* be its optimal value. We say that $f(\mathbf{x})$ satisfies the **Polyak- Łojasiewicz** (PL) condition in $\mathbb{A} \subseteq \mathbb{R}^d$ if there exists $\lambda > 0$ such that for all $\mathbf{x} \in A$: $\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \lambda(f(\mathbf{x}) - f^*)$.*

In Karimi et al. (2016) the authors show that under mild smoothness assumptions on $f(\mathbf{x})$, if it satisfies the PL condition then gradient descent with a small enough step size have linear convergence rate to a global minimum. The PL condition became popular in recent years to show convergence of gradient descent for non-convex functions. For our objective, we will show a stronger result, that the PL condition does not apply even locally around any differentiable global minimum, and even if we restrict to an ϵ-orthogonal neighborhood:

Theorem 9 *Assume $n > k$, let $\epsilon > 0$ and let $\tilde{\mathbf{w}}_1^n = (\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_n)$ be a differentiable global minimum of Eq. (2). Then the objective does not satisfy the PL condition, even in an ϵ-orthogonal neighborhood of $\tilde{\mathbf{w}}_1^n$.*

The full proof can be found in Appendix B.3. The proof idea is the same as Thm. 7, by showing that the same point chosen in the proof of that theorem also violates the PL condition.

3.4. One-Point Strong Convex in Most Directions

As we previously showed, the objective surface in Eq. (2) around any differentiable global minimum is not locally convex, and also not necessarily locally OPSC, even if we restrict to an ϵ-orthogonal neighborhood. The reason for the latter is that in this neighborhood, there are "bad" points which do not satisfy the OPSC condition. Thus, it is natural to ask how common are these "bad" points.

Here, we show that these points are fortunately rare, in the following sense: If we move away from a global minimum in some direction (inside its ϵ-orthogonal neighborhood), then in "most" directions, we will arrive at points which do satisfy some form of the OPSC condition, as formalized in the theorem below. For this theorem, we consider the case where $n = m \cdot k$ where $m \geq 1$, and for simplicity consider the global minimum that for each target vector $\mathbf{v}_i$ there are exactly m neurons, each equal to $\frac{1}{m}\mathbf{v}_i$ (however it is not too difficult to extend it to all differentiable global minima - see Remark 11). We use a slightly different notation here, namely the vectorized form $\mathbf{w}_1^n \in \mathbb{R}^{n \cdot d}$ here contains vectors $\mathbf{w}_{i,j} \in \mathbb{R}^d$ for $i \in [k], j \in [m]$, to represent the assumption that at the global minimum there are m neurons in the direction of the target $\mathbf{v}_i$ for $i \in [k]$.

Theorem 10 *Let $n = m \cdot k$ and let $\tilde{\mathbf{w}}_1^n$ be the global minimum of Eq. (2) where $\tilde{\mathbf{w}}_{i,j} = \frac{1}{m}\mathbf{v}_i$ for $j \in [m]$ and $i \in [k]$. For $\epsilon > 0$ let $U_\epsilon^\perp(\tilde{\mathbf{w}}_1^n)$ be the ϵ-orthogonal neighborhood of $\tilde{\mathbf{w}}_1^n$. Also, denote $\mathbf{g}_{i,j} = \mathbf{w}_{i,j} - \tilde{\mathbf{w}}_{i,j}$, $\mathbf{g}_i = \sum_{j=1}^m \mathbf{g}_{i,j}$, $\mathbf{g} = \sum_{j=1}^m \sum_{i=1}^k \mathbf{g}_{i,j}$, denote by $H(\mathbf{w}_1^n)$ the Hessian of the objective at $\mathbf{w}_1^n$. Then if $\mathbf{w} \in U_\epsilon^\perp(\tilde{\mathbf{w}}_1^n)$ we have that:*

$$(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n) \geq \frac{1}{4}\left(\|\mathbf{g}\|^2 + \left(1 - \frac{2}{\pi}\right)\sum_{i=1}^k \|\mathbf{g}_i\|^2\right) - O(\epsilon^{2.5}) \,, \tag{3}$$

where the $O(\cdot)$ notation hides factors polynomial in m and k.

The theorem implies that the OPSC coefficient is determined by the norms of sums of differences between each $\mathbf{w}_{i,j}$ and $\tilde{\mathbf{w}}_{i,j}$. Thus, unless these differences exactly cancel out, the right hand side will generally be positive. This means that if we move away from the global minimum $\tilde{\mathbf{w}}$ in some arbitrary direction, then the OPSC condition will generally hold w.r.t. $\tilde{\mathbf{w}}$ and the current point $\mathbf{w}$. We note that for simplicity's sake, the direction vector $\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n$ in Eq. (3) is not normalized to unit length.

We now give a few examples for different values of m and different points around the global minimum in order to give an intuition on which directions the one-point strong convexity applies:

Example 1 *In the following examples, for brevity, we divide both sides of Eq. (3) by $\|\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n\|^2$, this way the r.h.s. will have a term that is independent of ϵ in some directions, as we would see in the following examples*

- *Consider the case where $m = 1$, meaning that $n = k$. This is the exact parameterization case, in this case we get by the theorem that:*

$$\frac{1}{\|\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n\|^2} \cdot (\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n) \geq \frac{1}{4} - \frac{1}{2\pi} + \frac{\|\mathbf{g}\|^2}{4\sum_{i=1}^k \|\mathbf{g}_i\|^2} - O(\sqrt{\epsilon}) \ .$$

  This result conforms with our finding in Thm. 3 that for exact parameterization, the objective is strongly convex.

- *Assume that for every target vector $\mathbf{v}_i$ we have that $\mathbf{w}_{i,j}$ are equal for every $j \in [m]$. In this case:*

$$\frac{1}{\|\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n\|^2} \cdot (\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n) \geq m \cdot \left(\frac{1}{4} - \frac{1}{2\pi}\right) + \frac{\|\mathbf{g}\|^2}{4\sum_{i=1}^k \sum_{j=1}^m \|\mathbf{g}_{i,j}\|^2} - O(\sqrt{\epsilon}) \ .$$

  In this case the function is OPSC towards the global minimum $\tilde{\mathbf{w}}_1^n$, assuming ϵ is not too large. Note that the m term is a scaling factor that appears due to the over-parameterization.

- *Assume that for every target vector $\mathbf{v}_i$ we have that $\sum_{j=1}^m \mathbf{w}_{i,j} = \mathbf{0}$. In this case $\frac{1}{\|\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n\|^2} \cdot (\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)$ is of magnitude at most $O(\sqrt{\epsilon})$. This case is similar in nature to what was shown in Thm. 7 where the function is not OPSC.*

Remark 11 *In the theorem, we chose a specific global minimum for simplicity. The theorem can be readily extended to any differentiable global minimum $\tilde{\mathbf{w}}$, at the cost of having inside the big-O notation factors polynomial in $\min_{i,j} \|\mathbf{w}_{i,j}\|^{-1}$ (which for our global minimum reduce to factors polynomial in m). We leave an exact analysis to future work.*

3.5. Optimization Under OPSC in Most Directions

Until now we have shown that although several standard properties which guarantee convergence with gradient descent (convexity, OPSC and PL condition) are not satisfied by our objective, it does satisfy another property - OPSC in most directions. In this subsection we show that, at least in certain cases, this property is enough to ensure convergence.

First, we note that in Thm. 10 there is a negative $O(\epsilon^{2.5})$ term. In the proof the sign of this term is not clearly determined, and further analysis will be needed to do so, which we leave for future work. With that said, we conjecture that this term is actually non-negative, at least in a close enough neighborhood of the global minimum. We also conjecture that this is true in a standard neighborhood of the global minimum, instead of an ϵ-orthogonal neighborhood as stated in the theorem. We state this formally in the following:

Conjecture 12 *In the setting of Thm. 10 and under the same assumptions, we have that:*

$$(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n) \geq \frac{1}{4}\left(\|\mathbf{g}\|^2 + \left(1 - \frac{2}{\pi}\right)\sum_{i=1}^{k}\|\mathbf{g}_i\|^2\right),$$

in a standard ϵ-neighborhood of every global minimum $\tilde{\mathbf{w}}_1^n$, where $\mathbf{w}_{i,j} \neq 0$ for all i, j.

We conduct thorough experiments to verify this conjecture empirically. They can be seen in Appendix D.1.

We would like to show that under Conjecture 12, initializing close enough to the global minimum would ensure convergence using gradient methods. Using standard gradient descent will not be enough here, since there are points for which the OPSC parameter is zero (even under the above assumption). To ensure convergence we need to add random noise to the optimization process which can help to escape those "bad" points.

We use a simple form of perturbed gradient descent, for the exact algorithm, see Appendix D.2. In simple words, the algorithm receives an initialized weights $\mathbf{w}_1^n(0)$, a learning rate η and noise level α. At each iteration the algorithm updates the weights w.r.t the loss function F similarly to gradient descent, and adds a perturbation in a random direction with magnitude α. The perturbation is in the same direction for all the learned vectors $\mathbf{w}_1, \dots, \mathbf{w}_n$.

We show convergence for a general function that have the property from Thm. 10, under an assumption similar to Conjecture 12. Even under this assumption, the OPSC parameter may be zero (or arbitrarily small) at some points. Nevertheless, using perturbed gradient descent we can show the following:

Theorem 13 *Let $F : \mathbb{R}^{d \cdot n} \to \mathbb{R}$ and assume that it achieves a global minimum at $\tilde{\mathbf{w}}_1^n = (\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_n) \in \mathbb{R}^{d \cdot n}$. Assume that there is an $\epsilon \in (0, 1]$ and $\lambda > 0$ such that in an ϵ-neighborhood of $\tilde{\mathbf{w}}_1^n$ the function F is twice differentiable, has an L-Lipschitz gradient, and we have that*

$$(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n) \geq \lambda \|\mathbf{g}\|^2 \tag{4}$$

where $H(\mathbf{w}_1^n)$ is the Hessian of F at $\mathbf{w}_1^n$, $\mathbf{g}_i = \mathbf{w}_i - \tilde{\mathbf{w}}_i$ and $\mathbf{g} = \sum_{i=1}^n \mathbf{g}_i$. Let $\delta > 0$. Then, initializing $\mathbf{w}_1^n(0)$ in an ϵ-neighborhood of $\tilde{\mathbf{w}}_1^n$ and using perturbed gradient descent (Algorithm D.2) with learning rate $\eta < \frac{\lambda \delta^2}{64 L^2}$ and noise $\alpha = \frac{\delta}{4n}$, after $T > \frac{\log(\delta)}{\log\left(1 - \frac{\eta \lambda \delta^2}{64}\right)}$ iterations w.p $> 1 - T e^{-\Omega(d)}$ (over the random perturbations) we have that $\|\mathbf{w}_1^n(T) - \tilde{\mathbf{w}}_1^n\|^2 \leq \delta$.

Note that the OPSC condition in this theorem is almost the same as in Thm. 10 for the case of having the property in a standard ϵ-neighborhood of the minimum. In this case, the $\|\mathbf{g}_i\|^2$ terms can be absorbed in the $\|\mathbf{g}\|^2$ terms (by increasing the constant λ).

The full proof can be found in Appendix D.3. The idea is to split the analysis into two cases: (1) $\|\mathbf{g}\|^2$ is not too small, hence a single gradient step will get $\mathbf{w}_1^n$ closer to $\tilde{\mathbf{w}}_1^n$; (2) $\|\mathbf{g}\|^2$ is very small, but the perturbation from the algorithm will help escape from those bad points.

Thm. 13 shows that even when the function is non-convex, if it has the OPSC in most directions property, gradient descent with small perturbations converges to a global minimum.

4. Effects of Over-parameterization on Non-global Minima

Having considered the effects of over-parameterization on the global minima of Eq. (2), in this section we turn to study the effects of over-parameterization on the non-global minima. In what follows, we define $H_{i,i}(\mathbf{w}_1^n)' := H_{i,i}(\mathbf{w}_1^n) - \frac{1}{2}I$, the component of the i-th diagonal block of the Hessian at $\mathbf{w}_1^n$, without the $\frac{1}{2}I$ term (see Eq. (5)). When the point $\mathbf{w}_1^n$ is clear from context, we let $H'_{i,i}$ be shorthand for $H_{i,i}(\mathbf{w}_1^n)'$. Given a point $\mathbf{w}_1^n \in \mathbb{R}^{nd}$, we let $\mathbf{w}_1^n(\alpha, i) = (\mathbf{w}_1, \dots, \mathbf{w}_{i-1}, \alpha\mathbf{w}_i, (1-\alpha)\mathbf{w}_i, \mathbf{w}_{i+1}, \dots, \mathbf{w}_n) \in \mathbb{R}^{(n+1)d}$ denote the point obtained from splitting the i-th neuron $\mathbf{w}_1^n$ into two neurons, one with a factor of α and the other with a factor of $1-\alpha$. All proofs of theorems appearing in this section can be found in Appendix E.

4.1. Over-parameterization Turns Non-global Minima into Saddle Points

As was empirically shown in (Safran and Shamir, 2017), very mild over-parameterization (adding one or two neurons) suffices for significantly improving the probability of gradient descent to recover global minima of Eq. (2). Thus, it is interesting to understand how such minimal over-parameterization changes the optimization landscape, in a way that helps local search methods avoid non-global minima. One major obstacle for pursuing this direction is that only certain non-global minima of Eq. (2) are known to have an explicit characterization (Arjevani and Field, 2020). However, if we are already given a local minimum $\mathbf{w}_1^n$, a simple way to generate additional critical points is to split the i-th neuron to obtain a point $\mathbf{w}_1^n(\alpha, i)$, for any $\alpha \in (0, 1)$ (see Lemma 28 for a formal statement). Our main result in this section is to demonstrate that if $n \leq k$ and $\sum_{i=1}^n \|\mathbf{w}_i\| \leq k$, then there exists a neuron that when split, the critical point obtained is a saddle point:

Theorem 14 *Suppose $n \leq k$, $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ is a non-global minimum of the objective in Eq. (2) such that $\sum_{i=1}^n \|\mathbf{w}_i\| \leq k$. Then F is twice continuously differentiable at $\mathbf{w}_1^n$ and there exists a neuron $\mathbf{w}_i$ such that $\mathbf{w}_1^n(\alpha, i)$ is a saddle point for all $\alpha \in (0, 1)$. Moreover, for $\alpha \in \{0, 1\}$ we have that $\mathbf{w}_1^n(\alpha, i)$ is not a local minimum of F.*

Although we do not have a proof that the assumption $\sum_{i=1}^n \|\mathbf{w}_i\| \leq k$ holds for all minima of the objective, in Subsection 4.2 we demonstrate empirically that this appears to be the case, at least for the minima found by gradient descent. Moreover, this assumption provably holds for the global minima (see Thm. 2). Finally, we can prove the following weaker bound for any minimum:

Proposition 15 *Suppose $n \geq 1$, $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ is a local minimum of the objective in Eq. (2). Then $\sum_{i=1}^n \|\mathbf{w}_i\| \leq kn$.*

We also remark that the theorem applies to the critical points obtained when splitting local minima where $n \leq k$, and it is possible that there are new local minima formed when $n > k$ that did not exist when $n \leq k$, which our analysis does not touch upon. However, current empirical evidence (see (Safran and Shamir, 2017)) suggests that these minima are less common and pose a much less significant obstacle to optimization.

Combining Thm. 14 with Thm. 2, we see that global minima can be split arbitrarily and remain global minima, whereas non-global minima can only be split in restricted ways before turning into saddle points. This provides an indication for why over-parameterization makes the landscape more favorable to optimization, and possibly explains why recovering the global minimum becomes easier when over-parameterizing.

The key in proving Thm. 14 is the observation that when we split the i-th neuron in $\mathbf{w}_1^n$, we obtain a critical point of F, and the Hessian of this new point cannot be PSD if the $H'_{i,i}$ is not PSD. Indeed, the role of the norm sum bound assumption in Thm. 14 is to show that there must exist at least one neuron having a component $H'_{i,i}$ which is not PSD. However, if we make the stronger assumption that *for several* i's, $H'_{i,i}$ is not PSD (which based on the proof of Thm. 14, we can expect to happen when for each such i, $\mathbf{w}_i$ has roughly unit norm and $\min_{j\in[k]} \theta_{\mathbf{w}_i,\mathbf{v}_j}$ is not too small) then this implies a stronger result, that when we split any such neuron i with non-PSD $H'_{i,i}$, this would necessarily turn the local minimum into a saddle point. More formally, we have the following theorem:

Theorem 16 *Suppose $n \geq 1$, $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ is a differentiable, non-global minimum of the objective in Eq. (2). Then for all $i \in [n]$ such that $H'_{i,i}$ is not PSD, $\mathbf{w}_1^n(\alpha, i)$ is a saddle point for all $\alpha \in (0, 1)$. Moreover, for $\alpha \in \{0, 1\}$ and any such $i \in [n]$, $\mathbf{w}_1^n(\alpha, i)$ is not a local minimum of F.*

In particular, if $H'_{i,i}$ is not PSD for *all* $i \in [n]$, then splitting $\mathbf{w}_1^n$ would necessarily turn it into a saddle point, regardless of which neuron is being split. In the next subsection, we show empirically that this indeed appears to be the case in general.

4.2. An Experiment

In this subsection[1], we wish to substantiate empirically the assumption $\sum_{i=1}^n \|\mathbf{w}_i\| \leq k$ made in Thm. 14, as well as the claim that $H'_{i,i}$ tends to be a non-PSD matrix. To that end, for each $n = k$ between 6 and 100, we ran 500 instantiations of gradient descent on the objective in Eq. (2), each using an independent and standard Xavier random initialization and a fixed step size of $5/k$,[2] till the norm of the gradient was at most 10^{-12}. Moreover, we ran 100 additional instantiations where we initialized at a point having a large norm-sum of roughly $2k^2$ (note that Proposition 15 guarantees that there are no minima with norm-sum more than k^2). We identified points that were equivalent up to permutations of the neurons and their coordinates (up to Frobenius norm of at most $5 \cdot 10^{-9}$). For each group of equivalent points, we computed the spectrum of the Hessian to ensure that its minimal eigenvalue is positive (using floating point computations), which was always the case.

Once the local minima we converged to were processed, we first validated the norm sum assumption of $\sum_{i=1}^k \|\mathbf{w}_i\| \leq k$ which we made in Thm. 14. *All* local minima found in our experiment indeed satisfy this bound. Moreover, histogram plots of a few selected values for k are presented in Fig. 1, suggesting that the norm sum tends to be tightly concentrated at a value slightly below k.

Next, we computed the eigenvalues of the $H'_{i,i}$ in the Hessians of the local minima found, for all i. As it turns out, *all* block components for *all* minima found have a negative eigenvalue, which by virtue of Thm. 16 implies that for any minimum point $\mathbf{w}_1^k$ found, $\mathbf{w}_1^k(\alpha, i)$ is a saddle point for all $i \in [k]$ and any $\alpha \in (0, 1)$ (and not a minimum for $\alpha \in \{0, 1\}$).

Acknowledgements

This research is supported in part by European Research Council (ERC) grant 754705.

1. The code can be found at `https://github.com/ItaySafran/Overparameterization`
2. Empirically, this step size resulted in satisfactory convergence rates for all values of k we tested.

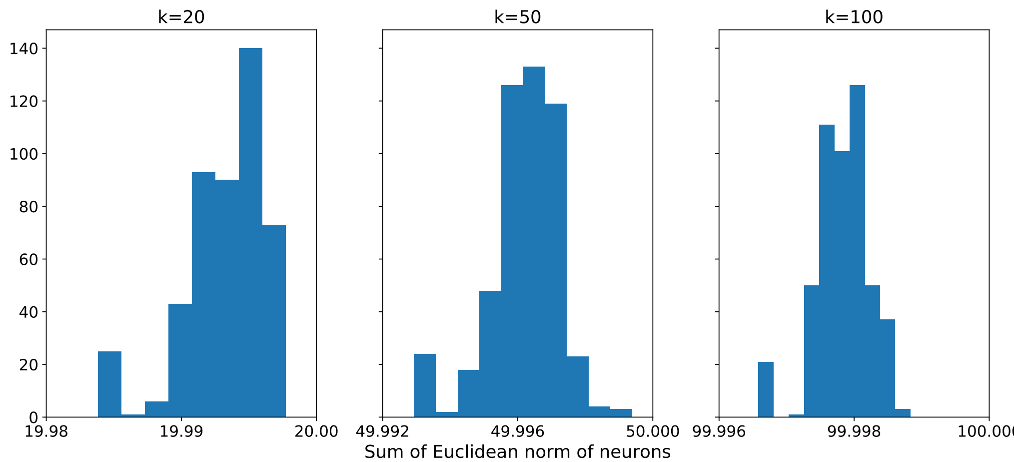

Figure 1: Histograms of the distributions of the sum of Euclidean norms of the neurons in the points converged to in the experiment, for $k = 20, 50, 100$.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? In *Advances in Neural Information Processing Systems*, pages 9017–9028, 2019.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Can SGD learn recurrent neural networks with provable generalization? *arXiv preprint arXiv:1902.01028*, 2019.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

[4] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.

[5] Yossi Arjevani and Michael Field. Spurious local minima of shallow relu networks conform with the symmetry of the target model. *arXiv preprint arXiv:1912.11939*, 2019.

[6] Yossi Arjevani and Michael Field. Analytic characterization of the hessian in shallow relu models: A tale of symmetry. *arXiv preprint arXiv:2008.01805*, 2020.

[7] Yossi Arjevani and Michael Field. Symmetry & critical points for a model shallow neural network. *arXiv preprint arXiv:2003.10576*, 2020.

[8] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

[9] Yuan Cao and Quanquan Gu. A generalization theory of gradient descent for learning overparameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*, 2019.

[10] Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009.

[11] Amit Daniely. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

[12] Amit Daniely and Eran Malach. Learning parities with neural networks. *arXiv preprint arXiv:2002.07400*, 2020.

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

[14] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9108–9118, 2019.

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

[16] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[18] Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. *arXiv preprint arXiv:2003.04180*, 2020.

[19] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[20] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pages 2698–2707, 2018.

[21] Jasper CH Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614. IEEE, 2016.

[22] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8168–8177, 2018.

[23] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in neural information processing systems*, pages 597–607, 2017.

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

[25] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

[26] Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89, 1963.

[27] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[28] Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pages 774–782, 2016.

[29] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[30] Mahdi Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, pages 2007–2017, 2017.

[31] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

[32] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3404–3413. JMLR. org, 2017.

[33] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2019.

[34] Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. *arXiv preprint arXiv:2001.05205*, 2020.

[35] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149. JMLR. org, 2017.

Appendices

Appendix A. Proof Of Lemma 1

Theorem 17 *Let $n \geq k$, and let $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ such that $\mathbf{w}_i \neq 0$ for every $i \in [n]$. Denote by $H(\mathbf{w}_1^n)$ the Hessian of $F(\mathbf{w}_1^n)$ (the objective in Eq. (2)). It is an $(n \cdot d) \times (n \cdot d)$ matrix, where for ease of notations we view $H(\mathbf{w}_1^n)$ as a $n \times n$ block matrix where each entry is a block of size $d \times d$. For every $i \in [n]$ the diagonal block entry of the Hessian is:*

$$H_{i,i}(\mathbf{w}_1^n) = \frac{1}{2}I + \sum_{j \neq i} h_1(\mathbf{w}_i, \mathbf{w}_j) - \sum_{j \in [k]} h_1(\mathbf{w}_i, \mathbf{v}_j) \tag{5}$$

where

$$h_1(\mathbf{w},\mathbf{v}) = \frac{\sin(\theta_{\mathbf{w},\mathbf{v}})\|\mathbf{v}\|}{2\pi\|\mathbf{w}\|}\left(I - \bar{\mathbf{w}}\bar{\mathbf{w}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}^\top\right) \tag{6}$$

and $\mathbf{n}_{\mathbf{v},\mathbf{w}} = \bar{\mathbf{v}} - \cos(\theta_{\mathbf{w},\mathbf{v}})\bar{\mathbf{w}}$. *For every* $i,j \in [n]$ *with* $i \neq j$ *the off-diagonal entry of the Hessian is* $H_{i,j}(\mathbf{w}_1^n) = h_2(\mathbf{w}_i,\mathbf{w}_j)$ *where*

$$h_2(\mathbf{w},\mathbf{v}) = \frac{1}{2\pi}\left((\pi - \theta_{\mathbf{w},\mathbf{v}})I + \bar{\mathbf{n}}_{\mathbf{w},\mathbf{v}}\bar{\mathbf{v}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{w}}^\top\right)\ . \tag{7}$$

We will need the following auxiliary lemma which calculates $\lim_{\mathbf{w},\mathbf{v}\to\mathbf{u}} h_2(\mathbf{w},\mathbf{v})$ for $\mathbf{u} \neq \mathbf{0}$

Lemma 18 *Suppose* $\mathbf{u} \neq \mathbf{0} \in \mathbb{R}^d$. *Then* $\lim_{\mathbf{w},\mathbf{v}\to\mathbf{u}} h_2(\mathbf{w},\mathbf{v}) = \frac{1}{2}I$.

Proof By Thm. 17 we have that:

$$h_2(\mathbf{w},\mathbf{v}) = \frac{1}{2\pi}\left((\pi - \theta_{\mathbf{w},\mathbf{v}})I + \bar{\mathbf{n}}_{\mathbf{w},\mathbf{v}}\bar{\mathbf{v}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{w}}^\top\right) = \frac{1}{2}I + \bar{\mathbf{n}}_{\mathbf{w},\mathbf{v}}\bar{\mathbf{v}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{w}}^\top. \tag{8}$$

We will show that the second and third terms approach zero if $\mathbf{w},\mathbf{v} \to \mathbf{u}$. Define the shorthand $\theta := \theta_{\mathbf{w},\mathbf{v}}$, then we have:

$$\begin{aligned}
\bar{\mathbf{n}}_{\mathbf{w},\mathbf{v}}\bar{\mathbf{v}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{w}}^\top &= \frac{\bar{\mathbf{w}}\bar{\mathbf{v}}^\top - \cos(\theta)\bar{\mathbf{v}}\bar{\mathbf{v}}^\top}{\sin(\theta)} + \frac{\bar{\mathbf{v}}\bar{\mathbf{w}}^\top - \cos(\theta)\bar{\mathbf{w}}\bar{\mathbf{w}}^\top}{\sin(\theta)} \\
&= \frac{\bar{\mathbf{w}}\bar{\mathbf{v}}^\top - \cos(\theta)\bar{\mathbf{v}}\bar{\mathbf{v}}^\top + \bar{\mathbf{w}}\bar{\mathbf{v}}^\top\cos(\theta) - \bar{\mathbf{w}}\bar{\mathbf{v}}^\top\cos(\theta)}{\sin(\theta)} + \\
&+ \frac{\bar{\mathbf{v}}\bar{\mathbf{w}}^\top - \cos(\theta)\bar{\mathbf{w}}\bar{\mathbf{w}}^\top + \bar{\mathbf{v}}\bar{\mathbf{w}}^\top\cos(\theta) - \bar{\mathbf{v}}\bar{\mathbf{w}}^\top\cos(\theta)}{\sin(\theta)} \\
&= \frac{\bar{\mathbf{w}}\bar{\mathbf{v}}^\top(1-\cos(\theta))}{\sin(\theta)} + \frac{\bar{\mathbf{v}}\bar{\mathbf{w}}^\top(1-\cos(\theta))}{\sin(\theta)} + \frac{(\bar{\mathbf{w}}-\bar{\mathbf{v}})\cos(\theta)\bar{\mathbf{v}}^\top}{\sin(\theta)} + \frac{(\bar{\mathbf{v}}-\bar{\mathbf{w}})\cos(\theta)\bar{\mathbf{w}}^\top}{\sin(\theta)} \\
&= \frac{(\bar{\mathbf{w}}-\bar{\mathbf{v}})(\bar{\mathbf{v}}^\top - \bar{\mathbf{w}}^\top)\cos(\theta)}{\sin(\theta)} + \frac{\bar{\mathbf{w}}\bar{\mathbf{v}}^\top(1-\cos(\theta))}{\sin(\theta)} + \frac{\bar{\mathbf{v}}\bar{\mathbf{w}}^\top(1-\cos(\theta))}{\sin(\theta)}.
\end{aligned} \tag{9}$$

If $\mathbf{w},\mathbf{v} \to \mathbf{u}$ the last two terms of Eq. (9) go to zero, since the outer product results in a matrix of bounded norm that is multiplied by $(1-\cos(\theta))/\sin(\theta)$ which tends to zero (can be seen using L'Hôpital's rule). For the first term, we will prove it is the zero matrix by showing that multiplying the term by any unit vector from the right yields the zero vector. Letting $\mathbf{z} \in \mathbb{R}^d$ with $\|\mathbf{z}\| = 1$, we have:

$$\begin{aligned}
\left\|\frac{(\bar{\mathbf{w}}-\bar{\mathbf{v}})(\bar{\mathbf{v}}^\top - \bar{\mathbf{w}}^\top)\mathbf{z}\cos(\theta)}{\sin(\theta)}\right\| &= \frac{\|\bar{\mathbf{w}}-\bar{\mathbf{v}}\| \cdot |\langle \bar{\mathbf{v}}-\bar{\mathbf{w}}, \mathbf{z}\rangle|\cos(\theta)}{\sin(\theta)} \\
&\leq \frac{\|\bar{\mathbf{w}}-\bar{\mathbf{v}}\|^2\|\mathbf{z}\|\cos(\theta)}{\sin(\theta)} = \frac{(2-2\cos(\theta))\cos(\theta)}{\sin(\theta)} \underset{\theta\to 0}{\to} 0\ ,
\end{aligned}$$

where the inequality is from Cauchy-Schwarz. This is true for every unit vector $\mathbf{z}$, hence this is the zero matrix. Combining the above shows that $\bar{\mathbf{n}}_{\mathbf{w},\mathbf{v}}\bar{\mathbf{v}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{w}}^\top = \mathbf{0}_{d\times d}$ ■

Proof [Proof of Lemma 1] First, recall the gradient of Eq. (2) at $\mathbf{w}_1^n$ which is defined and continuous as long as $\mathbf{w}_i \neq \mathbf{0}$ for all $i \in [n]$, as computed in (8; 29), where the coordinates with indices $(i-1)d+1$ to $i \cdot d$ are given by

$$\frac{1}{2}\mathbf{w}_i + \sum_{j \neq i} g\left(\mathbf{w}_i, \mathbf{w}_j\right) - \sum_{j=1}^{k} g\left(\mathbf{w}_i, \mathbf{v}_j\right), \tag{10}$$

where

$$g\left(\mathbf{w}, \mathbf{v}\right) = \frac{1}{2\pi}\left(\|\mathbf{v}\| \sin\left(\theta_{\mathbf{w},\mathbf{v}}\right) \bar{\mathbf{w}} + \left(\pi - \theta_{\mathbf{w},\mathbf{v}}\right) \mathbf{v}\right). \tag{11}$$

Clearly, by Thm. 17 the gradient is continuously differentiable for any $\mathbf{w}_1^n$ where the angle between any two vectors $\theta_{\mathbf{w}_i,\mathbf{w}_j} \neq 0, \pi$ for $i \neq j$. We will show that the partial derivatives of $h_1(\mathbf{w},\mathbf{v})$ and $h_2(\mathbf{w},\mathbf{v})$ are continuous for all $\mathbf{w},\mathbf{v} \neq \mathbf{0}$, by showing that they coincide with the derivative of g whenever $\theta_{\mathbf{w}_i,\mathbf{w}_j}$ tends to 0 or π.

We begin with computing the limits of $h_1(\mathbf{w},\mathbf{v})$ and $h_2(\mathbf{w},\mathbf{v})$ when $\theta_{\mathbf{w},\mathbf{v}} \to 0$ and $\theta_{\mathbf{w},\mathbf{v}} \to \pi$. First, we have

$$\lim_{\mathbf{w} \to \mathbf{u}} h_1(\mathbf{w},\mathbf{u}) = \mathbf{0}_{d \times d}$$

and

$$\lim_{\mathbf{w} \to -\mathbf{u}} h_1(\mathbf{w},\mathbf{u}) = \mathbf{0}_{d \times d}.$$

This holds since in both cases $\sin(\theta_{\mathbf{w},\mathbf{u}}) \to 0$ and since that for any unit vector $\mathbf{x}$, $\|\mathbf{x}\mathbf{x}^\top\|$ is uniformly bounded. Next, we have from Lemma 18 that

$$\lim_{\mathbf{v} \to \mathbf{u}} h_2(\mathbf{u},\mathbf{v}) = \frac{1}{2}I,$$

and from a straightforward calculation that

$$\lim_{\mathbf{v} \to -\mathbf{u}} h_2(\mathbf{u},\mathbf{v}) = \mathbf{0}_{d \times d}.$$

Assume $\mathbf{w} \to \mathbf{u}$ and $\mathbf{v} = t\mathbf{u}$ for some $t > 0$ and non-zero vector $\mathbf{u}$ and let $\mathbf{e}_i$ denote the unit vector with all-zero coordinates except for the i-th coordinate, we compute the partial derivative of $g(\mathbf{w},\mathbf{v})$ with respect to coordinate i of $\mathbf{w}$:

$$\begin{aligned}
\frac{\partial g}{\partial w_i}(\mathbf{w},\mathbf{v}) &= \lim_{\epsilon \to 0} \frac{\|t\mathbf{u}\| \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right) \frac{\mathbf{u}+\epsilon\mathbf{e}_i}{\|\mathbf{u}+\epsilon\mathbf{e}_i\|} + \left(\pi - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right) t\mathbf{u} - \left(\|t\mathbf{u}\| \sin\left(\theta_{\mathbf{u},\mathbf{u}}\right) \frac{\mathbf{u}}{\|\mathbf{u}\|} + \left(\pi - \theta_{\mathbf{u},\mathbf{u}}\right) t\mathbf{u}\right)}{2\pi\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\|t\mathbf{u}\| \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right) \frac{\mathbf{u}+\epsilon\mathbf{e}_i}{\|\mathbf{u}+\epsilon\mathbf{e}_i\|} + \left(\pi - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right) t\mathbf{u} - \pi t\mathbf{u}}{2\pi\epsilon} \qquad (12) \\
&= \lim_{\epsilon \to 0} \frac{t}{2\pi} \frac{\|\mathbf{u}\| \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right) \frac{\mathbf{u}+\epsilon\mathbf{e}_i}{\|\mathbf{u}+\epsilon\mathbf{e}_i\|} - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\mathbf{u}}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{t}{2\pi} \frac{\frac{\|\mathbf{u}\|}{\|\mathbf{u}+\epsilon\mathbf{e}_i\|} \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right) \mathbf{u} - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\mathbf{u}}{\epsilon}, \qquad (13)
\end{aligned}$$

where equality (12) is due to $\sin\left(\theta_{\mathbf{u},\mathbf{u}}\right) = 0$ and equality (13) is due to $\frac{\sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right)\epsilon\mathbf{e}_i}{\epsilon} \to 0$. Assume w.l.o.g. that $\epsilon \to 0^+$ (the following arguments are reversed in order if $\epsilon \to 0^-$), we have by using the inequality $\sin(x) \leq x$ which holds for all $x \geq 0$ that Eq. (13) is upper bounded by

$$\lim_{\epsilon \to 0} \frac{t}{2\pi} \frac{\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}}}{\epsilon} \left(\frac{\|\mathbf{u}\|}{\|\mathbf{u} + \epsilon \mathbf{e}_i\|} - 1 \right) \mathbf{u}. \tag{14}$$

Next, we have by the law of sines that

$$\frac{\epsilon}{\sin(\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}})} = \frac{\|\mathbf{u}\|}{\sin(\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \epsilon \mathbf{e}_i})} \geq \|\mathbf{u}\|,$$

which entails

$$\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}} \leq \arccos\left(\sqrt{1 - \frac{\epsilon^2}{\|\mathbf{u}\|^2}} \right),$$

therefore by L'Hôpital's rule

$$\lim_{\epsilon \to 0} \left| \frac{\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}}}{\epsilon} \right| \leq \frac{1}{\|\mathbf{u}\|}, \tag{15}$$

and since $\frac{\|\mathbf{u}\|}{\|\mathbf{u}+\epsilon \mathbf{e}_i\|} \to 1$, this implies that Eq. (14) converges to $\mathbf{0}$. Using the inequality $\sin(x) \geq x - \frac{x^3}{6}$ which holds for all $x \geq 0$ we lower bound Eq. (13) by

$$\lim_{\epsilon \to 0} \frac{t}{2\pi} \frac{\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}}}{\epsilon} \left(1 - \frac{\theta^2_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}}}{6} \right) \left(\frac{\|\mathbf{u}\|}{\|\mathbf{u} + \epsilon \mathbf{e}_i\|} - 1 \right) \mathbf{u}. \tag{16}$$

We have $1 - \theta^2_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}}/6 \to 1$ and $\frac{\|\mathbf{u}\|}{\|\mathbf{u}+\epsilon \mathbf{e}_i\|} \to 1$, and from Eq. (15) we have that the above converges to $\mathbf{0}$. Combining Eq. (14) and Eq. (16) and using the squeeze theorem, we have that $\frac{\partial g}{\partial w_i}(\mathbf{w}, \mathbf{v}) = \mathbf{0}$, from which it follows that the Hessian at $(\mathbf{w}, \mathbf{v})$ is the zero matrix $\mathbf{0}_{d \times d}$.

Now, assume $\mathbf{w} \to \mathbf{u}$ and $\mathbf{v} = -t\mathbf{u}$ for some $t > 0$, and compute

$$\begin{aligned}
\frac{\partial g}{\partial w_i}(\mathbf{w}, \mathbf{v}) &= \lim_{\epsilon \to 0} \frac{\|t\mathbf{u}\| \sin\left(\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, -\mathbf{u}}\right) \frac{\mathbf{u}+\epsilon \mathbf{e}_i}{\|\mathbf{u}+\epsilon \mathbf{e}_i\|} - \left(\pi - \theta_{\mathbf{u}+\epsilon \mathbf{e}_i, -\mathbf{u}}\right) t\mathbf{u} - \left(\|t\mathbf{u}\| \sin\left(\theta_{\mathbf{u}, -\mathbf{u}}\right) \frac{\mathbf{u}}{\|\mathbf{u}\|} - \left(\pi - \theta_{\mathbf{u}, -\mathbf{u}}\right) t\mathbf{u} \right)}{2\pi\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\|t\mathbf{u}\| \sin\left(\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, -\mathbf{u}}\right) \frac{\mathbf{u}+\epsilon \mathbf{e}_i}{\|\mathbf{u}+\epsilon \mathbf{e}_i\|} - \left(\pi - \theta_{\mathbf{u}+\epsilon \mathbf{e}_i, -\mathbf{u}}\right) t\mathbf{u}}{2\pi\epsilon} \qquad (17) \\
&= \lim_{\epsilon \to 0} \frac{t}{2\pi} \frac{\|\mathbf{u}\| \sin\left(\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}}\right) \frac{\mathbf{u}+\epsilon \mathbf{e}_i}{\|\mathbf{u}+\epsilon \mathbf{e}_i\|} - \theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}} \mathbf{u}}{\epsilon} = \mathbf{0}, \qquad (18)
\end{aligned}$$

where equality (17) is due to $\theta_{\mathbf{u}, -\mathbf{u}} = \pi$ and equality (18) is due to $\theta_{\mathbf{u}+\epsilon \mathbf{e}_i, -\mathbf{u}} = \pi - \theta_{\mathbf{u}+\epsilon \mathbf{e}_i, \mathbf{u}}$, and since we get the same limit as we did in the previous case. This implies $\lim_{\mathbf{w} \to -\mathbf{u}} h_1(\mathbf{w}, \mathbf{u}) = \mathbf{0}_{d \times d}$, and concludes the derivation for h_1.

Moving on to h_2, assume $\mathbf{v} \to \mathbf{u}$ and $\mathbf{w} = t\mathbf{u}$ for some $t > 0$, and compute

$$\begin{aligned}
\frac{\partial g}{\partial v_i}(\mathbf{w},\mathbf{v}) &= \lim_{\epsilon\to 0} \frac{\|\mathbf{u}+\epsilon\mathbf{e}_i\| \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right) \frac{t\mathbf{u}}{\|t\mathbf{u}\|} + \left(\pi - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right)(\mathbf{u}+\epsilon\mathbf{e}_i) - \left(\|\mathbf{u}\| \sin\left(\theta_{\mathbf{u},\mathbf{u}}\right) \frac{t\mathbf{u}}{\|t\mathbf{u}\|} + \left(\pi - \theta_{\mathbf{u},\mathbf{u}}\right)\mathbf{u}\right)}{2\pi\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{\|\mathbf{u}+\epsilon\mathbf{e}_i\| \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right) \frac{\mathbf{u}}{\|\mathbf{u}\|} + \left(\pi - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right)(\mathbf{u}+\epsilon\mathbf{e}_i) - \pi\mathbf{u}}{2\pi\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{1}{2\pi} \frac{\frac{\|\mathbf{u}+\epsilon\mathbf{e}_i\|}{\|\mathbf{u}\|} \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right)\mathbf{u} - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\mathbf{u}}{\epsilon} + \frac{\mathbf{e}_i}{2} - \lim_{\epsilon\to 0} \frac{\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\mathbf{e}_i}{2\pi} \\
&= \lim_{\epsilon\to 0} \frac{1}{2\pi} \frac{\frac{\|\mathbf{u}+\epsilon\mathbf{e}_i\|}{\|\mathbf{u}\|} \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right)\mathbf{u} - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\mathbf{u}}{\epsilon} + \frac{\mathbf{e}_i}{2},
\end{aligned}$$

and following the same reasoning as in the proof for h_1 we have that the above limit is $\mathbf{0}$, which implies that

$$\frac{\partial g}{\partial v_i}(\mathbf{u},\mathbf{u}) = \frac{1}{2}I.$$

Now assume $\mathbf{v} \to \mathbf{u}$ and $\mathbf{w} = -t\mathbf{u}$ for some $t > 0$, and compute

$$\begin{aligned}
&\frac{\partial g}{\partial v_i}(\mathbf{w},\mathbf{v}) \\
&= \lim_{\epsilon\to 0} \frac{\|\mathbf{u}+\epsilon\mathbf{e}_i\| \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,-\mathbf{u}}\right) \frac{-t\mathbf{u}}{\|-t\mathbf{u}\|} + \left(\pi - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,-\mathbf{u}}\right)(\mathbf{u}+\epsilon\mathbf{e}_i) - \left(\|\mathbf{u}\| \sin\left(\theta_{\mathbf{u},-\mathbf{u}}\right) \frac{-t\mathbf{u}}{\|-t\mathbf{u}\|} + \left(\pi - \theta_{\mathbf{u},-\mathbf{u}}\right)\mathbf{u}\right)}{2\pi\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{-\|\mathbf{u}+\epsilon\mathbf{e}_i\| \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,-\mathbf{u}}\right) \frac{\mathbf{u}}{\|\mathbf{u}\|} + \left(\pi - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,-\mathbf{u}}\right)(\mathbf{u}+\epsilon\mathbf{e}_i)}{2\pi\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{1}{2\pi} \frac{-\frac{\|\mathbf{u}+\epsilon\mathbf{e}_i\|}{\|\mathbf{u}\|} \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right)\mathbf{u} + \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\mathbf{u}}{\epsilon} - \lim_{\epsilon\to 0} \frac{\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\mathbf{e}_i}{2\pi} \\
&= -\frac{1}{2\pi} \lim_{\epsilon\to 0} \frac{\frac{\|\mathbf{u}+\epsilon\mathbf{e}_i\|}{\|\mathbf{u}\|} \sin\left(\theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\right)\mathbf{u} - \theta_{\mathbf{u}+\epsilon\mathbf{e}_i,\mathbf{u}}\mathbf{u}}{\epsilon}.
\end{aligned}$$

From the previous case we have that the above limit is $\mathbf{0}$, implying that

$$\frac{\partial g}{\partial v_i}(\mathbf{u},-\mathbf{u}) = \mathbf{0}_{d\times d},$$

and concluding the proof of the lemma. ■

Appendix B. Proofs from Sec. 3

B.1. Proof of Thm. 2

To prove the theorem we will need the following lemma, which essentially asserts that misclassifying a single instance will result in a strictly positive loss in expectation.

Lemma 19 *Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be continuous functions, and suppose exists $\mathbf{x}_0 \in \mathbb{R}^d$ s.t. $f(\mathbf{x}_0) \neq g(\mathbf{x}_0)$. Then*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{1}{2}(f(\mathbf{x}) - g(\mathbf{x}))^2\right] > 0.$$

Proof Assume w.l.o.g. $f(\mathbf{x}_0) - g(\mathbf{x}_0) = c > 0$. Since f and g are continuous, there exists an open neighborhood $U \ni \mathbf{x}_0$ s.t.

$$|f(\mathbf{z}) - g(\mathbf{z})| > c, \ \ \forall \mathbf{z} \in U. \tag{19}$$

Let A denote the event where a point $\mathbf{z}$ sampled from a multivariate normal random variable belongs to U, then by the law of total expectation

$$\begin{aligned}\mathbb{E}\left[\frac{1}{2}(f(\mathbf{x}) - g(\mathbf{x}))^2\right] &= \mathbb{E}\left[\frac{1}{2}(f(\mathbf{x}) - g(\mathbf{x}))^2|A\right]\Pr[A] + \mathbb{E}\left[\frac{1}{2}(f(\mathbf{x}) - g(\mathbf{x}))^2|\bar{A}\right]\Pr[\bar{A}]\\ &\geq \mathbb{E}\left[\frac{1}{2}(f(\mathbf{x}) - g(\mathbf{x}))^2|A\right]\Pr[A] > 0,\end{aligned}$$

where the strict inequality is due to $\Pr[A] > 0$ since U is open and a multivariate normal random variable has a measure which is strictly positive on all of $\mathbb{R}^d$, and due to $\mathbb{E}\left[\frac{1}{2}(f(\mathbf{x}) - g(\mathbf{x}))^2|A\right] > 0$ by virtue of Eq. (19) holding whenever A occurs. ■

Proof [Proof of Thm. 2] First, assume w.l.o.g. $v_j = e_j$ for all $j \in [k]$. This is justified since an orthonormal change of bases does not change the geometry of our objective. By virtue of Lemma 19 and the continuity of ReLU networks, it suffices to find a single point $\mathbf{x}$ s.t. any network with a different structure than in the theorem statement disagrees on $\mathbf{x}$ with $f(\mathbf{x}) = \sum_{i=1}^{k}[x_i]_+$. To this end, we shall divide the proof into several different cases, based on the set of weights $\mathbf{w}_1, \ldots, \mathbf{w}_n$ of the approximating network N.

- If $w_{i,j} < 0$ for some i, j, then w.l.o.g. $i = j = 1$ and

$$f(-e_1) = 0 < [w_{1,1} \cdot -1]_+ \leq N(-e_1).$$

- Otherwise, for $\mathbf{x} = e_1$ we have

$$f(e_1) = 1 = N(e_1) = \sum_{i=1}^{n}[w_{i,1}]_+,$$

  and thus

$$\sum_{i=1}^{n} w_{i,1} = 1.$$

- Suppose that exist two coordinates in the same neuron that are not 0, w.l.o.g $w_{1,1}, w_{1,2} > 0$. Then for $\mathbf{x} = (1, -\frac{w_{i,1}}{w_{i,2}}, 0, \ldots, 0)$, we have

$$f(\mathbf{x}) = 1 = \sum_{i=1}^{n} w_{i,1} = w_{i,1} + \sum_{i=2}^{n}[w_{i,1}]_+ > \sum_{i=2}^{n}[w_{i,1}]_+ \geq \sum_{i=1}^{n}\left[w_{i,1} - \frac{w_{i,1}}{w_{i,2}} \cdot w_{i,2}\right]_+ = N(\mathbf{x}).$$

Overall, if W^* does not have the structure as in the theorem statement then this results in a misclassified point which due to Lemma 19 implies the result. ■

B.2. Proof of Thm. 3

First we calculate the Hessian of the objective at a global minimum. Since we assume that the vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ are orthogonal the Hessian has a simple form:

Lemma 20 *Assume that $n = k$ and let $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ be a global minima. Then the Hessian $H(\mathbf{w}_1^n)$ of the objective Eq. (2) has the following block form:*

- *For $i \in [n]$:*
$$H(\mathbf{w}_1^n)_{i,i} = \frac{1}{2}I$$

- *For $i, j \in [n]$ with $i \neq j$:*
$$H_{i,j}(\mathbf{w}_1^n) = \frac{1}{4}I + \frac{1}{2\pi}\left(\bar{\mathbf{w}}_i\bar{\mathbf{w}}_j^\top + \bar{\mathbf{w}}_j\bar{\mathbf{w}}_i^\top\right)$$

where we look at the Hessian as a $k \times k$ block matrix, each block of size $d \times d$.

Proof Assume w.l.o.g that at this global minimum $\mathbf{w}_i = \mathbf{v}_i$ for every i. For the first item let $i \in [n]$, we have that:

$$H(\mathbf{w}_1^n)_{i,i} = \frac{1}{2}I + \sum_{j \neq i} h_1(\mathbf{w}_i, \mathbf{w}_j) - \sum_{l=1}^{k} h_1(\mathbf{w}_i, \mathbf{v}_l) = \frac{1}{2}I - h_1(\mathbf{w}_i, \mathbf{v}_i)\ . \tag{20}$$

We will show that if $\mathbf{w}, \mathbf{v}$ are parallel then $h_1(\mathbf{w}, \mathbf{v}) = \mathbf{0}_{d\times d}$. Let $\mathbf{w}, \mathbf{v}$ be two parallel non-zero vectors and $\mathbf{u}$ be some vector not parallel to them. We have by definition of $h_1(\mathbf{u}, \mathbf{v})$ that:

$$\begin{aligned}\lim_{\mathbf{u}\to\mathbf{w}} h_1(\mathbf{u}, \mathbf{v}) &= \lim_{\mathbf{u}\to\mathbf{w}} \frac{\sin(\theta_{\mathbf{u},\mathbf{v}})\|\mathbf{v}\|}{2\pi\|\mathbf{u}\|}\left(I - \bar{\mathbf{u}}\bar{\mathbf{u}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}^\top\right) \\ &= \lim_{\mathbf{u}\to\mathbf{w}} \frac{\|\mathbf{v}\|}{2\pi\|\mathbf{u}\|}\sin(\theta_{\mathbf{u},\mathbf{v}})\bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}^\top\ . \end{aligned}\tag{21}$$

We will show that the second term above is the zero matrix. Note that:

$$\begin{aligned}\|\mathbf{n}_{\mathbf{w},\mathbf{v}}\| &= \sqrt{\langle \bar{\mathbf{v}} - \cos(\theta_{\mathbf{w},\mathbf{v}})\bar{\mathbf{w}}, \bar{\mathbf{v}} - \cos(\theta_{\mathbf{w},\mathbf{v}})\bar{\mathbf{w}}\rangle} \\ &= \sqrt{1 - \cos^2(\theta_{\mathbf{w},\mathbf{v}})} = \sin(\theta_{\mathbf{w},\mathbf{v}})\ ,\end{aligned}$$

hence we have that $\sin(\theta_{\mathbf{u},\mathbf{v}})\bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}^\top = \mathbf{n}_{\mathbf{v},\mathbf{u}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}^\top$. Letting $\mathbf{x}$ be some vectors with norm 1, we have:

$$\lim_{\mathbf{u}\to\mathbf{w}} \|\mathbf{n}_{\mathbf{v},\mathbf{u}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}^\top\mathbf{x}\| \leq \lim_{\mathbf{u}\to\mathbf{w}} \|\mathbf{n}_{\mathbf{v},\mathbf{u}}\| \cdot |\langle \bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}}, \mathbf{x}\rangle| \leq \lim_{\mathbf{u}\to\mathbf{w}} \|\mathbf{n}_{\mathbf{v},\mathbf{u}}\| = 0\ .$$

This is true for every vector $\mathbf{x}$, hence $\lim_{\mathbf{u}\to\mathbf{w}} \mathbf{n}_{\mathbf{v},\mathbf{u}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{u}} = \mathbf{0}_{d\times d}$. Combining this with Eq. (21) and that $\mathbf{w} \neq \mathbf{0}$ we have that $\lim_{\mathbf{u}\to\mathbf{w}} h_1(\mathbf{u}, \mathbf{v}) = \mathbf{0}_{d\times d}$. This proves the first item of the lemma.

For the second item, recall that by our assumption the target vectors are orthogonal. Hence we have for $i \neq j$:

$$\begin{aligned} h_2(\mathbf{w}_i\mathbf{w}_j) &= \frac{1}{2\pi}\left((\pi-\theta_{\mathbf{w}_i,\mathbf{w}_j})I + \bar{\mathbf{n}}_{\mathbf{w}_i,\mathbf{w}_j}\bar{\mathbf{w}}_j^\top + \bar{\mathbf{n}}_{\mathbf{w}_j,\mathbf{w}_i}\bar{\mathbf{w}}_i^\top\right) \\ &= \frac{1}{2\pi}\left(\left(\pi-\frac{\pi}{2}\right)I + \frac{\bar{\mathbf{w}}_i\bar{\mathbf{w}}_j^\top - \cos(\theta_{\mathbf{w}_i,\mathbf{w}_j})\bar{\mathbf{w}}_j\bar{\mathbf{w}}_j^\top}{\sin(\theta_{\mathbf{w}_i,\mathbf{w}_j})} + \frac{\bar{\mathbf{w}}_j\bar{\mathbf{w}}_i^\top - \cos(\theta_{\mathbf{w}_i,\mathbf{w}_j})\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top}{\sin(\theta_{\mathbf{w}_i,\mathbf{w}_j})}\right) \\ &= \frac{1}{4}I + \frac{1}{2\pi}\left(\bar{\mathbf{w}}_i\bar{\mathbf{w}}_j^\top + \bar{\mathbf{w}}_j\bar{\mathbf{w}}_i^\top\right)\ . \end{aligned}$$

■

We are now ready to prove the theorem:

Proof [Proof of Thm. 3] Let $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_k)$ be some global minimum, by Lemma 20 the Hessian at $\mathbf{w}$ is equal to the sum of the following matrices:

$$H(\mathbf{w}) = \begin{pmatrix} \frac{1}{2}I_d & \dots & \frac{1}{4}I_d \\ \vdots & \ddots & \vdots \\ \frac{1}{4}I_d & \dots & \frac{1}{2}I_d \end{pmatrix} + \frac{1}{2\pi}\begin{pmatrix} 0_d & E_{1,2} & \dots & E_{1,n} \\ E_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & E_{1,n-1} \\ E_{n,1} & \dots & E_{n-1,n} & 0_d \end{pmatrix}, \tag{22}$$

where $E_{i,j} = \bar{\mathbf{w}}_i\bar{\mathbf{w}}_j^\top + \bar{\mathbf{w}}_j\bar{\mathbf{w}}_i^\top$ is a $d \times d$ matrix. Recall that the Hessian can be viewed as a $k \times k$ block matrix with blocks of size $d \times d$. We will calculate the smallest eigenvalue of the two matrices in Eq. (22), thus bounding the smallest eigenvalue of $H(\mathbf{w})$.

For the first matrix, the vectors $\begin{pmatrix} \mathbf{e}_i \\ \vdots \\ \mathbf{e}_i \end{pmatrix}$ is an eigenvector for every $i \in [d]$ with eigenvalue $\frac{k+1}{4}$.

Also, the vectors $\begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{e}_i \\ -\mathbf{e}_i \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}$ where the $\mathbf{e}_i$ can be at any two consecutive coordinates, are eigenvectors with eigenvalue $\frac{1}{4}$. There are d eigenvectors of the first kind, and $(k-1)\cdot d$ of the second kind. All of these vectors are linearly independent, thus we found $k \cdot d$ independent eigenvectors. This proves that the smallest eigenvalue of the first matrix is $\frac{1}{4}$.

For the second matrix we define a block vector $\tilde{\boldsymbol{\alpha}}$ of size $k \cdot d$ as a vectors with k coordinates, each coordinate is a vector of size d. Let $i, j \in [k]$ with $i \neq j$ and define the following block vectors:

- $\tilde{\boldsymbol{\alpha}}_i = \bar{\mathbf{w}}_i$, $\tilde{\boldsymbol{\alpha}}_j = \bar{\mathbf{w}}_j$, and the rest of the coordinates of $\tilde{\boldsymbol{\alpha}}$ are the zero vector.
- $\tilde{\boldsymbol{\beta}}_i = \bar{\mathbf{w}}_j$, $\tilde{\boldsymbol{\beta}}_j = -\bar{\mathbf{w}}_i$, and the rest of the coordinates of $\tilde{\boldsymbol{\beta}}$ are the zero vector.

- $\tilde{\boldsymbol{\gamma}}_i = \bar{\mathbf{w}}_i$, $\tilde{\boldsymbol{\gamma}}_j = \bar{\mathbf{w}}_j$, and the rest of the coordinates of $\tilde{\boldsymbol{\gamma}}$ are the zero vector.
- $\tilde{\boldsymbol{\delta}}_i = \bar{\mathbf{w}}_i$ for every $i \in [k]$

Note that $E_{i,j}\bar{\mathbf{w}}_i = \bar{\mathbf{w}}_i\bar{\mathbf{w}}_j^\top\bar{\mathbf{w}}_i + \bar{\mathbf{w}}_j\bar{\mathbf{w}}_i^\top\bar{\mathbf{w}}_i = \bar{\mathbf{w}}_j$, in the same manner $E_{i,j}\bar{\mathbf{w}}_j = \bar{\mathbf{w}}_i$ and for $l \in [k]$ with $l \neq i,\ l \neq j$ we have $E_{i,j}\bar{\mathbf{w}}_l = \mathbf{0}$. Denoting the second matrix in Eq. (22) as A, we have that $A\tilde{\boldsymbol{\alpha}} = -\tilde{\alpha},\ A\tilde{\boldsymbol{\beta}} = -\tilde{\boldsymbol{\beta}},\ A\tilde{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}},\ A\tilde{\boldsymbol{\delta}} = k\tilde{\boldsymbol{\delta}}$. Hence the vectors $\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}$ are eigenvectors for every $i \neq j$ with eigenvalue -1, the vectors $\tilde{\boldsymbol{\gamma}}$ are eigenvectors for every $i \neq j$ with eigenvalue 1, and $\tilde{\boldsymbol{\delta}}$ is an eigenvector with eigenvalue k. If $d = k$ then these eigenvectors span the entire space, hence the smaller eigenvalue is -1. If $d > k$ we complete $\mathbf{w}_{k+1}, \dots, \mathbf{w}_d$ to an orthogonal basis of the entire space, and add the eigenvectors which corresponds to the eigenvalue 0. In both cases the smallest eigenvalue of A is -1.

Combining the above with Eq. (22) and letting $\mathbf{v}$ be any vector with norm 1, we have:

$$\mathbf{v}^\top H(\mathbf{w})\mathbf{v} = \mathbf{v}^\top \begin{pmatrix} \frac{1}{2}I_d & \dots & \frac{1}{4}I_d \\ \vdots & \ddots & \vdots \\ \frac{1}{4}I_d & \dots & \frac{1}{2}I_d \end{pmatrix} \mathbf{v} + \frac{1}{2\pi}\mathbf{v}^\top \begin{pmatrix} 0_d & E_{1,2} & \dots & E_{1,n} \\ E_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & E_{n-1,n} \\ E_{n,1} & \dots & E_{n,n-1} & 0_d \end{pmatrix} \mathbf{v} \geq \frac{1}{4} - \frac{1}{2\pi}\,.$$

This proves that the Hessian is positive definite with minimal eigenvalue strictly larger than $\frac{1}{4} - \frac{1}{2\pi}$. Since the objective is twice differentiable, and the eigenvalue of a matrix is a continuous function, we have that the Hessian is positive definite in an open neighborhood of the global minimum. In particular, for any $0 < \lambda < \frac{1}{4} - \frac{1}{2\pi}$ there is an open neighborhood of the global minimum for which the objective is λ-strongly convex. ■

B.3. Proofs from Subsection 3.2 and Subsection 3.3

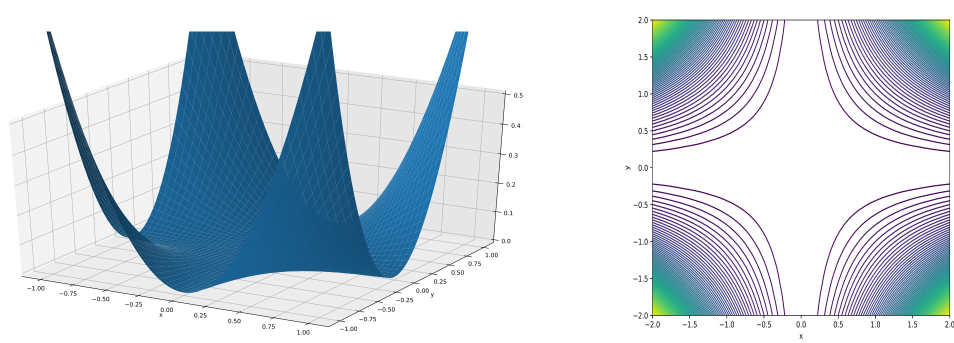

Figure 2: Plot and contour of the function $f(x, y) = x^2y^2$. This function is continuously differentiable and has a manifold of global minima $\{(x, y) : x = 0 \text{ or } y = 0\}$. $f(x, y)$ is not locally convex almost anywhere (as the determinant of its Hessian is negative almost everywhere). It is also not one-point strongly convex w.r.t. any global minimum.

See Fig. 2 for a plot of the function $f(x,y) = x^2y^2$. Note that although this function is neither locally convex nor OPSC, it is OPSC in an ϵ-orthogonal neighborhood of every global minima. Indeed, take some point $\tilde{\mathbf{w}} = (0, y)$ with any $y \neq 0$, then $\tilde{\mathbf{w}}$ is a global minimum of $f(x,y)$ for which $U_\epsilon^\perp(\tilde{\mathbf{w}})$ is not empty. It can be easily seen that for any $\epsilon > 0$ the function $f(x,y)$ is OPSC in $U_\epsilon^\perp(\tilde{\mathbf{w}})$ with respect to $\tilde{\mathbf{w}}$, with a strong convexity parameter of $\lambda = 2$. This is actually true for any global minimum $(x,y) \neq (0,0)$ of $f(x,y)$. Thus, although this function is not convex and also not OPSC, it is OPSC in an ϵ-orthogonal neighborhood of any global minimum except $(0,0)$. This function also does not satisfy the PL condition. Indeed, its global minimal value is $f^* = 0$, and we have

$$\|\nabla f(x,y)\|^2 = 4\left(f(x,y) - f^*\right) \cdot \left\| \begin{pmatrix} x \\ y \end{pmatrix} \right\|^2 ,$$

where it is easy to see that there is no global constant $\lambda > 0$ that satisfies the PL condition.

See Fig. B.3 for an intuition on an ϵ-orthogonal neighborhood.

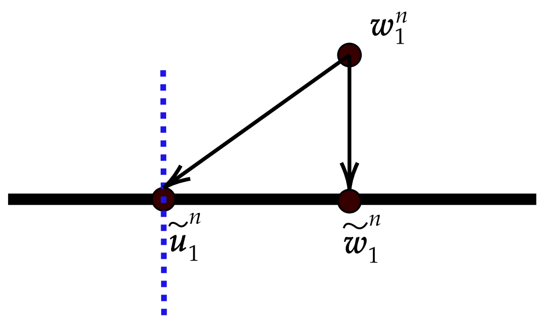

Figure 3: Suppose that the bold black line represents the manifold of global minima, $\tilde{\mathbf{w}}_1^n$ and $\tilde{\mathbf{u}}_1^n$ are two different global minima and $\mathbf{w}_1^n$ is some point. The blue dashed line represents the ϵ-orthogonal neighborhood of $\tilde{\mathbf{u}}_1^n$. In order to verify that the objective is OPSC we would have to decide with respect to which point it is OPSC. For example, from $\mathbf{w}_1^n$ it would make more sense to check its OPSC parameter with respect to $\tilde{\mathbf{w}}_1^n$ since it is its closest global minimum, rather than with respect to $\tilde{\mathbf{u}}_1^n$.

B.3.1. Proof of Thm. 4

Proof Suppose $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ is a global minimum. Assume w.l.o.g that $\mathbf{w}_1 = \alpha_1 \mathbf{v}_1,\ \mathbf{w}_2 = \alpha_2 \mathbf{v}_1$ for some $\alpha_1, \alpha_2 > 0$, that is there are at least two neurons that correspond to $\mathbf{v}_1$. Let $\epsilon > 0$, take $\mathbf{u}$ to be some unit vector orthogonal to $\mathbf{v}_1$ and define $\tilde{\mathbf{w}}_1 = \alpha_1 \epsilon \mathbf{u} + \alpha_1 \mathbf{v}_1,\ \tilde{\mathbf{w}}_2 = \alpha_2 \epsilon \mathbf{u} + \alpha_2 \mathbf{v}_1,\ \tilde{\mathbf{w}}_3 = \mathbf{w}_3, \dots \tilde{\mathbf{w}}_n = \mathbf{w}_n$. Note that $\tilde{\mathbf{w}}_1$ and $\tilde{\mathbf{w}}_2$ are in the same direction.

We will calculate the Hessian at $\tilde{\mathbf{w}}_1^n$. Recall that we view the Hessian as composed of $n \times n$ blocks, where each block is of size $d \times d$. By Lemma 18 we have that the blocks w.r.t. neurons $\mathbf{w}_1, \mathbf{w}_2$ are $H(\tilde{\mathbf{w}}_1^n)_{12} = H(\tilde{\mathbf{w}}_1^n)_{21} = \frac{1}{2} I$. For the diagonal components of the Hessian, note that if $\mathbf{z}_1$ and $\mathbf{z}_2$ are parallel then for every non-zero vector $\mathbf{u}$ we have that $h_1(\mathbf{u}, \mathbf{z}_1) + h_1(\mathbf{u}, \mathbf{z}_2) =$

$(\|\mathbf{z}_1\| + \|\mathbf{z}_2\|)h_1(\mathbf{u}, \bar{\mathbf{z}}_1)$. By Thm. 17 we have:

$$\begin{aligned} H(\tilde{\mathbf{w}}_1^n)_{11} &= \frac{1}{2}I + \sum_{j\neq 1} h_1(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_j) - \sum_{l=1}^{k} h_1(\tilde{\mathbf{w}}_1, \mathbf{v}_l) \\ &= \frac{1}{2}I - (\|\tilde{\mathbf{w}}_1\| + \|\tilde{\mathbf{w}}_2\|)h_1(\tilde{\mathbf{w}}_1, \mathbf{v}_1) \\ &= \frac{1}{2}I - (1+\epsilon)(\alpha_1 + \alpha_2)h_1(\tilde{\mathbf{w}}_1, \mathbf{v}_1) \,, \end{aligned} \tag{23}$$

and in the same manner $H(\tilde{\mathbf{w}})_{22} = \frac{1}{2}I - (1+\epsilon)(\alpha_1+\alpha_2)h_1(\tilde{\mathbf{w}}_2, \mathbf{v}_1)$. Also since $\tilde{\mathbf{w}}_1$ and $\tilde{\mathbf{w}}_2$ are parallel we have that $h_1(\tilde{\mathbf{w}}_2, \mathbf{v}_1) = \frac{\|\tilde{\mathbf{w}}_1\|}{\|\tilde{\mathbf{w}}_2\|}h_1(\tilde{\mathbf{w}}_1, \mathbf{v}_1)$. The matrix $h_1(\tilde{\mathbf{w}}_2, \mathbf{v}_1)$ has an eigenvalue equal to $\frac{\sin(\theta_{\tilde{\mathbf{w}}_1,\mathbf{v}_1})\|\mathbf{v}_1\|}{\pi\|\tilde{\mathbf{w}}_1\|}$. Note that this eigenvalue is positive since we define the angle to be $\theta_{\tilde{\mathbf{w}}_1,\mathbf{v}_1} \in [0, \pi]$. Taking $\mathbf{z}$ to be a unit eigenvector corresponding to this eigenvalue, we have:

$$\begin{aligned} &\begin{pmatrix}\mathbf{z}^\top & -\mathbf{z}^\top & 0 & \dots & 0\end{pmatrix} H(\tilde{\mathbf{w}}_1^n) \begin{pmatrix}\mathbf{z} \\ -\mathbf{z} \\ 0 \\ \vdots \\ 0\end{pmatrix} = \mathbf{z}^\top H(\tilde{\mathbf{w}}_1^n)_{11}\mathbf{z}^\top + \mathbf{z}^\top H(\tilde{\mathbf{w}}_1^n)_{22}\mathbf{z} - \mathbf{z}^\top H(\tilde{\mathbf{w}}_1^n)_{12}\mathbf{z} - \mathbf{z}^\top H(\tilde{\mathbf{w}}_1^n)_{21}\mathbf{z} \\ =& \frac{1}{2} - (1+\epsilon)(\alpha_1+\alpha_2)\frac{\sin(\theta_{\tilde{\mathbf{w}}_1,\mathbf{v}_1})\|\mathbf{v}_1\|}{\pi\|\tilde{\mathbf{w}}_1\|} + \frac{1}{2} - \frac{\alpha_1}{\alpha_2}\cdot(1+\epsilon)(\alpha_1+\alpha_2)\frac{\sin(\theta_{\tilde{\mathbf{w}}_1,\mathbf{v}_1})\|\mathbf{v}_1\|}{\pi\|\tilde{\mathbf{w}}_1\|} - \frac{1}{2} - \frac{1}{2} \\ =& -(1+\epsilon)\frac{\sin(\theta_{\tilde{\mathbf{w}}_1,\mathbf{v}_1})}{\pi}\cdot\frac{(\alpha_1+\alpha_2)^2}{\alpha_1\alpha_2} < 0 \,. \end{aligned}$$

This is true for every $\epsilon > 0$, hence in every neighborhood of the global minimum we found a point where the Hessian is not PSD, meaning that the loss is not locally convex. ■

B.3.2. Proof of Thm. 7

Proof Let $\epsilon > 0$. The idea of the proof is to show that there is $\mathbf{w}_1^n \in U_\epsilon^\perp(\tilde{\mathbf{w}}_1^n)$ such that the Hessian of the objective, projected in the direction $\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n$ is of magnitude $O(\epsilon)$. This means that there is no $\lambda > 0$ such that the objective is λ-OPSC in an ϵ-orthogonal neighborhood of the global minimum.

From the assumption that $n > k$, and by Thm. 2 we know that there are at least two vectors $\tilde{\mathbf{w}}_i, \tilde{\mathbf{w}}_j$ which are parallel. In particular, assume w.l.o.g that $\tilde{\mathbf{w}}_1 = \alpha_1\mathbf{v}_1$, $\tilde{\mathbf{w}}_2 = \alpha_2\mathbf{v}_1$ where $\alpha_1, \alpha_2 > 0$. We look at the following point:

$$\mathbf{w}_1 = \alpha_1\mathbf{v}_1 + \epsilon\mathbf{v}_2,\ \mathbf{w}_2 = \alpha_2\mathbf{v}_1 - \epsilon\mathbf{v}_2,\ \mathbf{w}_3 = \tilde{\mathbf{w}}_3, \dots, \mathbf{w}_n = \tilde{\mathbf{w}}_n \,.$$

Recall that the target vectors are orthogonal, hence $\mathbf{w}_1^n \in U_\epsilon^\perp(\tilde{\mathbf{w}}_1^n)$. Using Thm. 17 we can calculate the Hessian at the above point in the direction of the global minimum:

$$\frac{1}{\|\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n\|^2}(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n) = \frac{1}{2}\begin{pmatrix} -\mathbf{v}_2 \\ \mathbf{v}_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}^\top H(\mathbf{w}_1^n)\begin{pmatrix} -\mathbf{v}_2 \\ \mathbf{v}_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= \frac{1}{2}\Bigg(1 + \mathbf{v}_2^\top h_1(\mathbf{w}_1, \mathbf{w}_2)\mathbf{v}_2 + \mathbf{v}_2^\top h_1(\mathbf{w}_2, \mathbf{w}_1)\mathbf{v}_2 - \mathbf{v}_2^\top h_1(\mathbf{w}_1, \tilde{\mathbf{v}}_1)\mathbf{v}_2 - \mathbf{v}_2^\top h_1(\mathbf{w}_2, \tilde{\mathbf{v}}_1)\mathbf{v}_2$$

$$- 2\mathbf{v}_2 h_2(\mathbf{w}_1, \mathbf{w}_2)\mathbf{v}_2\Bigg) . \tag{24}$$

The largest eigenvalue of $h_1(\mathbf{w}_1, \mathbf{w}_2)$ (see Lemma 9 in (29)) is:

$$\frac{\sin(\theta_{\mathbf{w}_1,\mathbf{w}_2})\|\tilde{w}_2\|}{\pi\|\mathbf{w}_1\|} = \frac{\epsilon(\alpha_1 + \alpha_2)}{\pi\|\mathbf{w}_1\|^2} = \frac{\epsilon(\alpha_1 + \alpha_2)}{\pi\alpha_1^2 + \pi\epsilon^2} = O(\epsilon)$$

Hence $\mathbf{v}_2^\top h_1(\mathbf{w}_1, \mathbf{w}_2)\mathbf{v}_2 = O(\epsilon)$, and for the same reasoning we get that

$$\mathbf{v}_2^\top h_1(\mathbf{w}_2, \mathbf{w}_1)\mathbf{v}_2 = O(\epsilon),\ \mathbf{v}_2^\top h_1(\mathbf{w}_1, \mathbf{v}_1)\mathbf{v}_2 = O(\epsilon),\ \mathbf{v}_2^\top h_1(\mathbf{w}_2, \mathbf{v}_1)\mathbf{v}_2 = O(\epsilon) .$$

For the last term of the Hessian we will need the following:

$$\cos(\theta_{\mathbf{w}_1,\mathbf{w}_2}) = \frac{\langle \alpha_1\mathbf{v}_1 + \epsilon\mathbf{v}_2, \alpha_2\mathbf{v}_1 - \epsilon\mathbf{v}_2\rangle}{\sqrt{(\alpha_1^2 + \epsilon^2)(\alpha_2^2 + \epsilon^2)}} = \frac{\alpha_1\alpha_2 - \epsilon^2}{\sqrt{(\alpha_1^2 + \epsilon^2)(\alpha_2^2 + \epsilon^2)}}$$

$$\sin(\theta_{\mathbf{w}_1,\mathbf{w}_2}) = \sqrt{1 - \frac{(\alpha_1\alpha_2 - \epsilon^2)^2}{(\alpha_1^2 + \epsilon^2)(\alpha_2^2 + \epsilon^2)}} = \frac{\epsilon(\alpha_1 + \alpha_2)}{\sqrt{(\alpha_1^2 + \epsilon^2)(\alpha_2^2 + \epsilon^2)}}$$

$$\theta_{\mathbf{w}_1,\mathbf{w}_2} = \arccos(\cos(\theta_{\mathbf{w}_1,\mathbf{w}_2})) = O(\epsilon)$$

Using the above we can calculate last term in Eq. (24):

$$\mathbf{v}_2^\top h_2(\mathbf{w}_1, \mathbf{w}_2)\mathbf{v}_2 = \frac{1}{2\pi}\Bigg((\pi - \theta_{\mathbf{w}_1,\mathbf{w}_2})\langle \mathbf{v}_2, \mathbf{v}_2\rangle + \frac{\langle \bar{\mathbf{w}}_2, \mathbf{v}_2\rangle(\langle \bar{\mathbf{w}}_1, \mathbf{v}_2\rangle - \cos(\theta_{\mathbf{w}_1,\mathbf{w}_2})\langle \bar{\mathbf{w}}_2, \mathbf{v}_2\rangle)}{\sin(\theta_{\mathbf{w}_1,\mathbf{w}_2})} +$$

$$+ \frac{\langle \bar{\mathbf{w}}_1, \mathbf{v}_2\rangle(\langle \bar{\mathbf{w}}_2, \mathbf{v}_2\rangle - \cos(\theta_{\mathbf{w}_1,\mathbf{w}_2})\langle \bar{\mathbf{w}}_1, \mathbf{v}_2\rangle)}{\sin(\theta_{\mathbf{w}_1,\mathbf{w}_2})}\Bigg) = \frac{1}{2} + O(\epsilon).$$

Hence in total we have:

$$\frac{1}{\|\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n\|^2}(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n) = \frac{1}{2}\left(1 + O(\epsilon) - 1 - O(\epsilon)\right) = O(\epsilon) .$$

■

B.3.3. PROOF OF THM. 9

Proof The proof method is similar to that of the proof of Thm. 7. We use the same point $\mathbf{w}_1^n$ as in Thm. 7 which is in an ϵ-orthogonal neighborhood of the relevant global minima. For ease of notation let $\theta := \theta_{\mathbf{w}_1,\mathbf{w}_2}$ and $\gamma_1 := \theta_{\mathbf{w}_1,\mathbf{v}},\ \gamma_2 = \theta_{\mathbf{w}_2,\mathbf{v}}$.

We first calculate the objective of Eq. (2) using the closed form in (29) Section 4.1.1. Set $\alpha = \alpha_1 + \alpha_2$, and note that all the terms cancel out, except for those which include $\mathbf{w}_1$ and $\mathbf{w}_2$:

$$\begin{aligned} F(\mathbf{w}_1^n) &= f(\mathbf{w}_1,\mathbf{w}_1) + f(\mathbf{w}_2,\mathbf{w}_2) + f(\mathbf{w}_1,\mathbf{w}_2) + \\ &+ f(\mathbf{w}_2,\mathbf{w}_1) - 2f(\mathbf{w}_1,\alpha\mathbf{v}_1) - 2f(\mathbf{w}_2,\alpha\mathbf{v}_2) + f(\alpha\mathbf{v}_1,\alpha\mathbf{v}_1) \end{aligned} \tag{25}$$

where

$$f(\mathbf{w},\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I)}[\left[\left[\mathbf{w}^\top\mathbf{x}\right]_+\left[\mathbf{v}^\top\mathbf{x}\right]_+\right] = \frac{1}{2\pi}\|\mathbf{w}\|\|\mathbf{v}\|\left(\sin(\theta_{\mathbf{w},\mathbf{v}}) + (\pi-\theta_{\mathbf{w},\mathbf{v}})\cos(\theta_{\mathbf{w},\mathbf{v}})\right)\ .$$

To calculate this term we will need to the following expressions (calculated the same way as in Thm. 7):

$$\begin{aligned} \cos(\theta) &= \frac{\alpha_1\alpha_2 - \epsilon^2}{\sqrt{(\alpha_1^2+\epsilon^2)(\alpha_2^2+\epsilon^2)}}, & \sin(\theta) &= \frac{\epsilon(\alpha_1+\alpha_2)}{\sqrt{(\alpha_1^2+\epsilon^2)(\alpha_2^2+\epsilon^2)}} \\ \cos(\gamma_1) &= \frac{\alpha_1}{\sqrt{\alpha_1^2+\epsilon^2}}, & \sin(\gamma_1) &= \frac{\epsilon}{\sqrt{\alpha_1+\epsilon^2}} \\ \cos(\gamma_2) &= \frac{\alpha_2}{\sqrt{\alpha_2^2+\epsilon^2}}, & \sin(\gamma_2) &= \frac{\epsilon}{\sqrt{\alpha_1+\epsilon^2}} \\ \|\mathbf{w}_1\|^2 &= \alpha_1^2+\epsilon^2, & \|\mathbf{w}_2\|^2 &= \alpha_2^2+\epsilon^2 \end{aligned}$$

Also note, that using the taylor series of $\arccos$ in the same manner as the proof of Thm. 7 we get that: $\theta = O(\epsilon),\ \gamma_1 = O(\epsilon),\ \gamma_2 = O(\epsilon)$. The expression $f(\mathbf{w},\mathbf{v})$ depends only on the norms of $\mathbf{w}$ and $\mathbf{v}$ and on the angle between them, and also $f(\mathbf{w},\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$. Thus, returning to Eq. (25) we get:

$$\begin{aligned} F(\mathbf{w}_1^n) &= \frac{\|\mathbf{w}_1\|^2}{2} + \frac{\|\mathbf{w}_2\|^2}{2} + \frac{\|\mathbf{w}_1\|\|\mathbf{w}_2\|}{\pi}\left(\sin(\theta) + (\pi-\theta)\cos(\theta)\right) - \\ &- \frac{\alpha\|\mathbf{w}_1\|}{\pi}\left(\sin(\gamma_1) + (\pi-\gamma_1)\cos(\gamma_1)\right) - \frac{\alpha\|\mathbf{w}_2\|}{\pi}\left(\sin(\gamma_2) + (\pi-\gamma_2)\cos(\gamma_2)\right) + \frac{\alpha^2}{2} \\ &= \frac{1}{\pi}\left(\epsilon\alpha_1 + \epsilon\alpha_2 + \theta\epsilon^2 - \theta\alpha_1\alpha_2 - 2\alpha\epsilon + \alpha\alpha_1\gamma_1 + \alpha\alpha_2\gamma_2\right) = \Omega(\epsilon)\ . \end{aligned} \tag{26}$$

Next, we calculate the gradient of the objective using the closed form in in (29) Section 4.1.1.

$$(\nabla F(\mathbf{w}_1^n)))_1 = \frac{1}{2}\mathbf{w}_1 + g(\mathbf{w}_1,\mathbf{w}_2) - g(\mathbf{w}_1,\alpha\mathbf{v}_1)$$

where :

$$g(\mathbf{w},\mathbf{v}) = \frac{1}{2\pi}\left(\|\mathbf{v}\|\sin(\theta_{\mathbf{w},\mathbf{v}}) + (\pi-\theta_{\mathbf{w},\mathbf{v}})\mathbf{v}\right)\ .$$

Hence the norm of the gradient of $\mathbf{w}_1$ is:

$$\| \left(\nabla F(\mathbf{w}_1^n)\right)_1 \|^2 = \left\| \frac{1}{2}\mathbf{w}_1 + \frac{1}{2\pi}\left(\frac{\|\mathbf{w}_2\|}{\|\mathbf{w}_1\|}\sin(\theta)\mathbf{w}_1 + (\pi-\theta)\mathbf{w}_2\right) - \right.$$
$$\left. -\frac{1}{2\pi}\left(\frac{\alpha}{\|\mathbf{w}_1\|}\sin(\gamma_1)\mathbf{w}_1 + (\pi-\gamma_1)\alpha\mathbf{v}_1\right)\right\|^2$$
$$= \left\| \frac{\|\mathbf{w}_2\|\sin(\theta)}{2\pi\|\mathbf{w}_1\|}\mathbf{w}_1 - \frac{\theta}{2\pi}\mathbf{w}_2 - \frac{\alpha\sin(\gamma_1)}{2\pi\|\mathbf{w}_1\|}\mathbf{w}_1 + \frac{\alpha\gamma_1}{2\pi}\mathbf{v}_1\right\|^2 = O(\epsilon^2)\ . \quad (27)$$

In the same manner as in Eq. (27) we can show that also the norm of every other coordinate of the gradient of the objective is $O(\epsilon^2)$, hence we also have that $\|\nabla F(\mathbf{w}_1^n)\|^2 = O(\epsilon^2)$, where here the O notation hides a linear term in n (note that $F(\mathbf{w}_1^n)$ does not depend on n). In particular for every $\lambda > 0$ we can find $\epsilon > 0$ such that $\|\nabla F(\mathbf{w}_1^n))\|^2 < \lambda \cdot (F(\mathbf{w}_1^n) - f^*)$ (Recall that f^* is the value at the global minimum which is 0). This shows that the PL condition does not hold, even in an ϵ-orthogonal neighborhood of the global minimum. ■

Appendix C. Proofs from Subsection 3.4

The Hessian at $\mathbf{w}_1^n = (\mathbf{w}_{i,j})_{i,j=1}^{k,m}$ in the direction of global minimum $\tilde{\mathbf{w}}_1^n = (\tilde{\mathbf{w}}_{i,j})_{i,j=1}^{k,m}$ is (recall that $\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n = \mathbf{g}_1^n$):

$$\mathbf{g}_1^{n\top} H(\mathbf{w}_1^n)\mathbf{g}_1^n = \sum_{i,j=1}^{k,m}\left(\frac{1}{2}\|\mathbf{g}_{i,j}\|^2 + \sum_{\substack{a,b=1\\(a,b)\neq(i,j)}}^{k,m} \mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j},\mathbf{w}_{a,b})\mathbf{g}_{i,j} - \sum_{l=1}^{k}\mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j},\mathbf{v}_l)\mathbf{g}_{i,j} + \right.$$
$$\left. + \sum_{\substack{a,b=1\\(a,b)\neq(i,j)}}^{k,m} \mathbf{g}_{i,j}^\top h_2(\mathbf{w}_{i,j},\mathbf{w}_{a,b})\mathbf{g}_{a,b}\right) \quad (28)$$

The proof idea of Thm. 10 is to bound each term in Eq. (28) separately. Since we look at a point close to the global minimum, each $\mathbf{w}_{i,j}$ should be close to its target vector $\mathbf{v}_i$, hence most of the expressions will *almost* cancel out, up to an $O(\sqrt{\epsilon})$ factor.

For the proof we denote the following angles for ease of notations:

1. $\theta_{i,j}^{a,b}$: the angle between $\mathbf{w}_{i,j}$ and $\mathbf{w}_{a,b}$ for $i,a \in [k],\ j,b \in [m]$.

2. $\gamma_{i,j}^l$: the angle between $\mathbf{w}_{i,j}$ and v_l for $i,l \in [k],\ j \in [m]$.

For every i,j we can write $\mathbf{w}_{i,j} = \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j}$. Assume in the following that for some $\epsilon > 0$, we have that $\mathbf{w}_1^n \in U_\epsilon^\perp(\tilde{\mathbf{w}}_1^n)$, hence we have that $\|\mathbf{g}_{i,j}\| \leq \epsilon$ and $\mathbf{g}_{i,j} \perp \mathbf{v}_i$. We will need the following terms for $i \in [k],\ j,l \in [m]$:

$$\|\mathbf{w}_{i,j}\| = \left\|\frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j}\right\| = \sqrt{\frac{1}{m^2} + \langle \mathbf{v}_i, \mathbf{g}_{i,j}\rangle + \|\mathbf{g}_{i,j}\|^2} = \sqrt{\frac{1}{m^2} + \|\mathbf{g}_{i,j}\|^2} = \frac{1}{m} + O(\epsilon) \quad (29)$$

$$\cos(\theta_{i,j}^{i,l}) = \frac{\langle \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j}, \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,l}\rangle}{\|\mathbf{w}_{i,j}\|\|\mathbf{w}_{i,l}\|} = \frac{\frac{1}{m^2} + \langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle}{\frac{1}{m^2} + O(\epsilon)} = 1 + O(\epsilon) \quad (30)$$

$$\sin(\theta_{i,j}^{i,l}) = \sqrt{1 - \cos^2(\theta_{i,j}^{i,l})} = O(\sqrt{\epsilon}) \quad (31)$$

$$\cos(\gamma_{i,j}^{i}) = \frac{\langle \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j}, \mathbf{v}_i\rangle}{\|\mathbf{w}_{i,j}\|} = \frac{\frac{1}{m}}{\frac{1}{m} + O(\epsilon)} = 1 + O(\epsilon) \quad (32)$$

$$\sin(\gamma_{i,j}^{i}) = \sqrt{1 - \cos^2(\gamma_{i,j}^{i})} = O(\sqrt{\epsilon}) \quad (33)$$

For $i, a \in [k]$ and $j, b \in [m]$ with $i \neq a$ we have that:

$$\cos(\theta_{i,j}^{a,b}) = \frac{\langle \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j}, \frac{1}{m}\mathbf{v}_a + \mathbf{g}_{a,b}\rangle}{\|\mathbf{w}_{i,j}\|\|\mathbf{w}_{a,b}\|} = \frac{\langle \mathbf{g}_{i,j}, \mathbf{v}_a\rangle + \langle \mathbf{g}_{a,b}, \mathbf{v}_i\rangle + \langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle}{\frac{1}{m^2} + O(\epsilon)} = O(\epsilon) \quad (34)$$

$$\sin(\theta_{i,j}^{a,b}) = \sqrt{1 - \cos^2(\theta_{i,j}^{a,b})} = 1 + O(\sqrt{\epsilon}) \quad (35)$$

$$\cos(\gamma_{i,j}^{a}) = \frac{\langle \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j}, \mathbf{v}_a\rangle}{\|\mathbf{w}_{i,j}\|} = \frac{\langle \mathbf{g}_{i,j}, \mathbf{v}_a\rangle}{\frac{1}{m} + O(\epsilon)} = O(\epsilon) \quad (36)$$

$$\sin(\gamma_{i,j}^{a}) = \sqrt{1 - \cos^2(\gamma_{i,j}^{a})} = 1 + O(\sqrt{\epsilon}) \quad (37)$$

We will first bound the terms in Eq. (28) which are related to h_1.

Lemma 21 *For every $i \in [k]$ and $j, l \in [m]$ we have that:*

1. $\mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{w}_{i,l})\mathbf{g}_{i,j} = O\left(\epsilon^{2.5}\right)$
2. $\mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{v}_i)\mathbf{g}_{i,j} = O\left(\epsilon^{2.5}\right)$

Proof By Lemma 9 of (29) we know that the largest eigenvalue of $h_1(\mathbf{w}, \mathbf{v})$ is $\frac{\sin(\theta_{\mathbf{w},\mathbf{v}})\|\mathbf{v}\|}{\mathbf{w}}$. Hence we have that:

$$\mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{w}_{i,l})\mathbf{g}_{i,j} \leq \|\mathbf{g}_{i,j}\|^2 \frac{\sin(\theta_{i,j}^{i,l})\|\mathbf{w}_{i,l}\|}{\|\mathbf{w}_{i,j}\|} = O\left(\epsilon^{2.5}\right)$$

where we used Eq. (29), Eq. (31) and that $\|\mathbf{g}_{i,j}\| = O(\epsilon)$.

The second part is the same as the first, where we use Eq. (33). ■

Now we can bound all the terms in Eq. (28) related to h_1, leaving only on $O(\sqrt{\epsilon})$ term. Note that in the main theorem we will divide the full expression by the sum of the norms of $\mathbf{g}_{i,j}$, which is of magnitude $O(\epsilon^2)$.

Lemma 22 *For every $i \in [k], j \in [m]$ we have that:*

$$\sum_{\substack{a,b=1 \\ (a,b)\neq(i,j)}}^{k,m} \mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{w}_{a,b})\mathbf{g}_{i,j} - \sum_{l=1}^{k} \mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{v}_l)\mathbf{g}_{i,j} = O\left(\epsilon^{2.5}\right) \quad (38)$$

Proof Let $i, a \in [k]$ and $j, b \in [m]$ where $i \neq a$. First we have:

$$\frac{\langle \mathbf{g}_{i,j}, \mathbf{w}_{i,j} \rangle}{\|\mathbf{w}_{i,j}\|} \overset{(29)}{=} \frac{\langle \mathbf{g}_{i,j}, \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j} \rangle}{\frac{1}{m} + O(\epsilon)} = \frac{\|\mathbf{g}_{i,j}\|^2}{\frac{1}{m} + O(\epsilon)} = O(\epsilon^2) \tag{39}$$

$$\begin{aligned}\frac{\langle \mathbf{g}_{i,j}, \bar{\mathbf{w}}_{a,b} - \cos(\theta_{i,j}^{a,b})\bar{\mathbf{w}}_{i,j} \rangle}{\sin(\theta_{i,j}^{a,b})} &= \frac{\langle \mathbf{g}_{i,j}, \frac{1}{m}\mathbf{v}_a + \mathbf{g}_{a,b} \rangle}{\|\mathbf{w}_{a,b}\| \sin(\theta_{i,j}^{a,b})} - \frac{\cos(\theta_{i,j}^{a,b}) \langle \mathbf{g}_{i,j}, \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j} \rangle}{\|\mathbf{w}_{i,j}\| \sin(\theta_{i,j}^{a,b})} \\ &\overset{(34),(35)}{=} \frac{\frac{1}{m}\langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle + O(\epsilon^2)}{\left(\frac{1}{m} + O(\epsilon)\right) \cdot (1 + O(\sqrt{\epsilon}))} - \frac{\|\mathbf{g}_{i,j}\|^2 \cdot O(\epsilon)}{\left(\frac{1}{m} + O(\epsilon)\right) \cdot (1 + O(\sqrt{\epsilon}))} = \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle + O(\epsilon^{1.5}) \end{aligned} \tag{40}$$

Using the function h_1 we have:

$$\begin{aligned}\mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{w}_{a,b}) \mathbf{g}_{i,j} &= \frac{\sin(\theta_{i,j}^{a,b}) \|\mathbf{w}_{a,b}\|}{2\pi \|\mathbf{w}_{i,j}\|} \left(\|\mathbf{g}_{i,j}\|^2 - \frac{\langle \mathbf{g}_{i,j}, \mathbf{w}_{i,j} \rangle^2}{\|\mathbf{w}_{i,j}\|^2} + \frac{\langle \mathbf{g}_{i,j}, \bar{\mathbf{w}}_{a,b} - \cos(\theta_{i,j}^{a,b})\bar{\mathbf{w}}_{i,j} \rangle^2}{\sin^2(\theta_{i,j}^{a,b})} \right) \\ &\overset{(39),(40)}{=} \frac{\sin(\theta_{i,j}^{a,b}) \|\mathbf{w}_{a,b}\|}{2\pi \|\mathbf{w}_{i,j}\|} \left(\|\mathbf{g}_{i,j}\|^2 + O(\epsilon^4) + \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle^2 + O(\epsilon^{2.5}) \right) \\ &\overset{(29),(35)}{=} \frac{(1 + O(\sqrt{\epsilon})) \cdot \left(\frac{1}{m} + O(\epsilon)\right)}{2\pi \left(\frac{1}{m} + O(\epsilon)\right)} \left(\|\mathbf{g}_{i,j}\|^2 + \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle^2 + O(\epsilon^{2.5}) \right) \\ &= \frac{1}{2\pi} \left(\|\mathbf{g}_{i,j}\|^2 + \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle^2 \right) + O(\epsilon^{2.5}) \end{aligned} \tag{41}$$

In the same manner for $\mathbf{v}_a$ we have (recall that $\|\mathbf{v}_a\| = 1$):

$$\begin{aligned}\frac{\langle \mathbf{g}_{i,j}, \mathbf{v}_a - \cos(\gamma_{i,j}^a)\bar{\mathbf{w}}_{i,j} \rangle^2}{\sin^2(\gamma_{i,j}^a)} &= \frac{\langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle}{\sin(\gamma_{i,j}^a)} - \frac{\cos(\gamma_{i,j}^a) \langle \mathbf{g}_{i,j}, \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j} \rangle}{\|\mathbf{w}_{i,j}\| \sin(\gamma_{i,j}^a)} \\ &\overset{(36),(37)}{=} \frac{\langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle}{1 + O(\epsilon)} - \frac{\|\mathbf{g}_{i,j}\|^2 \cdot O(\epsilon)}{\left(\frac{1}{m} + O(\epsilon)\right) \cdot (1 + O(\sqrt{\epsilon}))} = \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle + O(\epsilon^{1.5}) \end{aligned} \tag{42}$$

Hence we get:

$$\begin{aligned}\mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{v}_a) \mathbf{g}_{i,j} &= \frac{\sin(\gamma_{i,j}^a)}{2\pi \|\mathbf{w}_{i,j}\|} \left(\|\mathbf{g}_{i,j}\|^2 - \frac{\langle \mathbf{g}_{i,j}, \mathbf{w}_{i,j} \rangle^2}{\|\mathbf{w}_{i,j}\|^2} + \frac{\langle \mathbf{g}_{i,j}, \mathbf{v}_a - \cos(\gamma_{i,j}^a)\bar{\mathbf{w}}_{i,j} \rangle^2}{\sin^2(\gamma_{i,j}^a)} \right) \\ &\overset{(39),(42)}{=} \frac{\sin(\gamma_{i,j}^a)}{2\pi \|\mathbf{w}_{i,j}\|} \left(\|\mathbf{g}_{i,j}\|^2 + O(\epsilon^4) + \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle^2 + O(\epsilon^{2.5}) \right) \\ &\overset{(29),(37)}{=} \frac{1 + O(\sqrt{\epsilon})}{2\pi \left(\frac{1}{m} + O(\epsilon)\right)} \left(\|\mathbf{g}_{i,j}\|^2 + \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle^2 + O(\epsilon^{2.5}) \right) \\ &= \frac{m}{2\pi} \left(\|\mathbf{g}_{i,j}\|^2 + \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle^2 \right) + O(\epsilon^{2.5}) \end{aligned} \tag{43}$$

For any $a \in [k]$ with $a \neq i$, combining Eq. (41) and Eq. (43) and summing over all $b \in [m]$ we get:

$$\sum_{b=1}^{m} \left(\mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{w}_{a,b}) \mathbf{g}_{i,j} \right) - \mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{v}_a) \mathbf{g}_{i,j}$$
$$= \sum_{b=1}^{m} \left(\frac{1}{2\pi} \left(\|\mathbf{g}_{i,j}\|^2 + \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle^2 \right) + O(\epsilon^{2.5}) \right) - \frac{m}{2\pi} \left(\|\mathbf{g}_{i,j}\|^2 + \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle^2 \right) + O(\epsilon^{2.5}) = O(\epsilon^{2.5})$$

Also, using Lemma 21 we get that $\sum_{b \neq j} \mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{w}_{i,b}) \mathbf{g}_{i,j} = O(\epsilon^{2.5})$ and that $\mathbf{g}_{i,j}^\top h_1(\mathbf{w}_{i,j}, \mathbf{v}_a) \mathbf{g}_{i,j} = O(\epsilon^{2.5})$. This finishes the proof. ■

Now we will bound terms related to h_2:

Lemma 23 *Letting $i, a \in [k]$ with $i \neq a$ and $b, j \in [m]$, we have:*

$$\mathbf{g}_{i,j}^\top h_2(\mathbf{w}_{i,j}, \mathbf{w}_{a,b}) \mathbf{g}_{a,b} = \frac{1}{4} \langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b} \rangle + \frac{1}{2\pi} \langle \mathbf{g}_{i,j}, \mathbf{v}_a \rangle \cdot \langle \mathbf{g}_{a,b}, \mathbf{v}_i \rangle + O(\epsilon^{2.5})$$

Proof We use Eq. (7) to get:

$$\mathbf{g}_{i,j}^\top h_2(\mathbf{w}_{i,j}, \mathbf{w}_{a,b}) \mathbf{g}_{a,b} = \frac{1}{2\pi} \left((\pi - \theta_{i,j}^{a,b}) \langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b} \rangle + \mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{i,j}, \mathbf{w}_{a,b}} \bar{\mathbf{w}}_{a,b}^\top \mathbf{g}_{a,b} + \mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{a,b}, \mathbf{w}_{i,j}} \bar{\mathbf{w}}_{i,j}^\top \mathbf{g}_{a,b} \right) \tag{44}$$

We will now bound each expression in Eq. (44). For the first term we will bound the angle $\theta_{i,j}^{a,b}$ using the Taylor series of arccos. The Taylor series of arccos is $\arccos(x) = \frac{\pi}{2} - \sum_{n=0}^{\infty} \frac{(2n)!}{2^{2n}(n!)^2} \frac{x^{2n+1}}{2n+1} = \frac{\pi}{2} - \sum_{n=0}^{\infty} c_n x^{2n+1}$ where $c_n \leq \frac{1}{2}$ for all $n \geq 0$. Hence we have that:

$$\theta_{i,j}^{a,b} = \arccos\left(\cos(\theta_{i,j}^{a,b}) \right) = \frac{\pi}{2} - \sum_{n=0}^{\infty} c_n \left(\cos(\theta_{i,j}^{a,b}) \right)^{2n+1} \overset{(34)}{=} \frac{\pi}{2} + O(\epsilon)$$

Hence we can bound the first term:

$$(\pi - \theta_{i,j}^{a,b}) \langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b} \rangle = \left(\pi - \frac{\pi}{2} + O(\epsilon) \right) \langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b} \rangle = \frac{\pi}{2} \langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b} \rangle + O(\epsilon^3) \tag{45}$$

For the second term we have:

$$\begin{aligned}
&\mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{i,j}, \mathbf{w}_{a,b}} \bar{\mathbf{w}}_{a,b}^\top \mathbf{g}_{a,b} = \\
&= \frac{\langle \bar{\mathbf{w}}_{a,b}, \mathbf{g}_{a,b} \rangle \cdot \left(\langle \bar{\mathbf{w}}_{i,j}, \mathbf{g}_{i,j} \rangle - \cos(\theta_{i,j}^{a,b}) \langle \bar{\mathbf{w}}_{a,b}, \mathbf{g}_{i,j} \rangle \right)}{\sin(\theta_{i,j}^{a,b})} \\
&= \frac{\langle \frac{1}{m} \mathbf{v}_a + \mathbf{g}_{a,b}, \mathbf{g}_{a,b} \rangle \cdot \left(\frac{1}{\|\mathbf{w}_{i,j}\|} \langle \frac{1}{m} \mathbf{v}_i + \mathbf{g}_{i,j}, \mathbf{g}_{i,j} \rangle - \frac{\cos(\theta_{i,j}^{a,b})}{\|\mathbf{w}_{a,b}\|} \langle \frac{1}{m} \mathbf{v}_a + \mathbf{g}_{a,b}, \mathbf{g}_{i,j} \rangle \right)}{\sin(\theta_{i,j}^{a,b}) \|\mathbf{w}_{a,b}\|} \\
&= \frac{\|\mathbf{g}_{a,b}^2\| \left(\frac{\|\mathbf{g}_{i,j}\|^2}{\|\mathbf{w}_{i,j}\|} - \frac{\cos(\theta_{i,j}^{a,b})}{\|\mathbf{w}_{a,b}\|} \left(\frac{1}{m} \langle \mathbf{v}_a, \mathbf{g}_{i,j} \rangle + \langle \mathbf{g}_{a,b}, \mathbf{g}_{i,j} \rangle \right) \right)}{\sin(\theta_{i,j}^{a,b}) \|\mathbf{w}_{a,b}\|} \\
&\overset{(29),(35),(34)}{=} \frac{\|\mathbf{g}_{a,b}\|^2 \left(\frac{\|\mathbf{g}_{i,j}\|^2}{\frac{1}{m} + O(\epsilon)} - \frac{O(\epsilon)}{\frac{1}{m} + O(\epsilon)} \left(\frac{1}{m} \langle \mathbf{v}_a, \mathbf{g}_{i,j} \rangle + \langle \mathbf{g}_{a,b}, \mathbf{g}_{i,j} \rangle \right) \right)}{(1 + O(\sqrt{\epsilon})) \cdot \left(\frac{1}{m} + O(\epsilon) \right)} = O(\epsilon^3)
\end{aligned} \tag{46}$$

For the third expression we have:

$$
\begin{aligned}
&\mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{a,b},\mathbf{w}_{i,j}} \bar{\mathbf{w}}_{i,j}^\top \mathbf{g}_{a,b} = \\
&= \frac{\langle \bar{\mathbf{w}}_{i,j}, \mathbf{g}_{a,b}\rangle \cdot \left(\langle \bar{\mathbf{w}}_{a,b}, \mathbf{g}_{i,j}\rangle - \cos(\theta_{i,j}^{a,b})\langle \bar{\mathbf{w}}_{i,j}, \mathbf{g}_{i,j}\rangle\right)}{\sin(\theta_{i,j}^{a,b})} \\
&= \frac{\langle \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle \cdot \left(\frac{1}{\|\mathbf{w}_{a,b}\|}\langle \frac{1}{m}\mathbf{v}_a + \mathbf{g}_{a,b}, \mathbf{g}_{i,j}\rangle - \frac{\cos(\theta_{i,j}^{a,b})}{\|\mathbf{w}_{i,j}\|}\langle \frac{1}{m}\mathbf{v}_i + \mathbf{g}_{i,j}, \mathbf{g}_{i,j}\rangle\right)}{\sin(\theta_{i,j}^{a,b})\|\mathbf{w}_{i,j}\|} \\
&= \frac{\left(\frac{1}{m}\langle \mathbf{v}_i, \mathbf{g}_{a,b}\rangle + \langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle\right) \cdot \left(\frac{1}{\|\mathbf{w}_{a,b}\|}\left(\frac{1}{m}\langle \mathbf{v}_a, \mathbf{g}_{i,j}\rangle + \langle \mathbf{g}_{a,b}, \mathbf{g}_{i,j}\rangle\right) - \frac{\cos(\theta_{i,j}^{a,b})\|\mathbf{g}_{i,j}\|^2}{\|\mathbf{w}_{i,j}\|}\right)}{\sin(\theta_{i,j}^{a,b})\|\mathbf{w}_{i,j}\|} \\
&= \frac{\langle \mathbf{v}_i, \mathbf{g}_{a,b}\rangle \cdot \langle \mathbf{v}_a, \mathbf{g}_{i,j}\rangle}{m^2 \sin(\theta_{i,j}^{a,b})\|\mathbf{w}_{a,b}\|\|\mathbf{w}_{i,j}\|} + \frac{\langle \mathbf{v}_i, \mathbf{g}_{a,b}\rangle \cdot \left(\frac{\langle \mathbf{g}_{a,b}, \mathbf{g}_{i,j}\rangle}{\|\mathbf{w}_{a,b}\|} - \frac{\cos(\theta_{i,j}^{a,b})\|\mathbf{g}_{i,j}\|^2}{\|\mathbf{w}_{i,j}\|}\right)}{m\sin(\theta_{i,j}^{a,b})\|\mathbf{w}_{i,j}\|} \\
&+ \frac{\langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle \cdot \left(\frac{1}{\|\mathbf{w}_{a,b}\|}\left(\frac{1}{m}\langle \mathbf{v}_a, \mathbf{g}_{i,j}\rangle + \langle \mathbf{g}_{a,b}, \mathbf{g}_{i,j}\rangle\right) - \frac{\cos(\theta_{i,j}^{a,b})\|\mathbf{g}_{i,j}\|^2}{\|\mathbf{w}_{i,j}\|}\right)}{\sin(\theta_{i,j}^{a,b})\|\mathbf{w}_{i,j}\|}
\end{aligned}
$$

As in the previous expression, since $\langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle, \|\mathbf{g}_{i,j}\|^2 = O(\epsilon^2)$, we have that:

$$
\frac{\langle \mathbf{v}_i, \mathbf{g}_{a,b}\rangle \cdot \left(\frac{\langle \mathbf{g}_{a,b}, \mathbf{g}_{i,j}\rangle}{\|\mathbf{w}_{a,b}\|} - \frac{\cos(\theta_{i,j}^{a,b})\|\mathbf{g}_{i,j}\|^2}{\|\mathbf{w}_{i,j}\|}\right)}{m\sin(\theta_{i,j}^{a,b})\|\mathbf{w}_{i,j}\|} = O(\epsilon^3)
$$

$$
\frac{\langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle \cdot \left(\frac{1}{\|\mathbf{w}_{a,b}\|}\left(\frac{1}{m}\langle \mathbf{v}_a, \mathbf{g}_{i,j}\rangle + \langle \mathbf{g}_{a,b}, \mathbf{g}_{i,j}\rangle\right) - \frac{\cos(\theta_{i,j}^{a,b})\|\mathbf{g}_{i,j}\|^2}{\|\mathbf{w}_{i,j}\|}\right)}{\sin(\theta_{i,j}^{a,b})\|\mathbf{w}_{i,j}\|} = O(\epsilon^3)
$$

In total we get that:

$$
\begin{aligned}
\mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{a,b},\mathbf{w}_{i,j}} \bar{\mathbf{w}}_{i,j}^\top \mathbf{g}_{a,b} &= \frac{\langle \mathbf{v}_i, \mathbf{g}_{a,b}\rangle \cdot \langle \mathbf{v}_a, \mathbf{g}_{i,j}\rangle}{m^2 \sin(\theta_{i,j}^{a,b})\|\mathbf{w}_{a,b}\|\|\mathbf{w}_{i,j}\|} + O(\epsilon^3) \\
&\overset{(29),(35)}{=} \frac{\langle \mathbf{v}_i, \mathbf{g}_{a,b}\rangle \cdot \langle \mathbf{v}_a, \mathbf{g}_{i,j}\rangle}{m^2(1+O(\sqrt{\epsilon}))\left(\frac{1}{m}+O(\epsilon)\right)^2} + O(\epsilon^3) = \langle \mathbf{v}_i, \mathbf{g}_{a,b}\rangle \cdot \langle \mathbf{v}_a, \mathbf{g}_{i,j}\rangle + O(\epsilon^{2.5}) \qquad (47)
\end{aligned}
$$

Overall, using Eq. (45), Eq. (46), Eq. (47) we have:

$$
\begin{aligned}
&\mathbf{g}_{i,j}^\top h_2(\mathbf{w}_{i,j}, \mathbf{w}_{a,b})\mathbf{g}_{a,b} = \frac{1}{2\pi}\left(\frac{\pi}{2}\langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle + O(\epsilon^3) + O(\epsilon^3) + \langle \mathbf{v}_i, \mathbf{g}_{a,b}\rangle \cdot \langle \mathbf{v}_a, \mathbf{g}_{i,j}^{a,b}\rangle + O(\epsilon^{2.5})\right) \\
&= \frac{1}{4}\langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle + \frac{1}{2\pi}\langle \mathbf{g}_{i,j}, \mathbf{v}_a\rangle \cdot \langle \mathbf{g}_{a,b}, \mathbf{v}_i\rangle + O(\epsilon^{2.5})
\end{aligned}
$$

■

Lemma 24 *Letting $i \in [k]$ and $j, l \in [m]$, we have:*

$$\mathbf{g}_{i,j}^\top h_2(\mathbf{w}_{i,j}, \mathbf{w}_{i,l})\mathbf{g}_{i,l} = \frac{1}{2}\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle + O(\epsilon^3)$$

Proof We use Thm. 17 to get:

$$\mathbf{g}_{i,j}^\top h_2(\mathbf{w}_{i,j}, \mathbf{w}_{i,l})\mathbf{g}_{i,l} = \frac{1}{2\pi}\left((\pi - \theta_{i,j}^{i,l})\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle + \mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{i,j},\mathbf{w}_{i,l}}\bar{\mathbf{w}}_{i,l}^\top \mathbf{g}_{i,l} + \mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{i,l},\mathbf{w}_{i,j}}\bar{\mathbf{w}}_{i,j}^\top \mathbf{g}_{i,l}\right) \tag{48}$$

We will bound each expression of Eq. (48). For the first expression, as in the proof of Lemma 23 we use the Taylor series of $\arccos$ to get:

$$\theta_{i,j}^{a,b} = \arccos\left(\cos(\theta_{i,j}^{i,l})\right) = \frac{\pi}{2} - \sum_{n=0}^{\infty} c_n \left(\cos(\theta_{i,j}^{i,l})\right)^{2n+1} \overset{(30)}{=} O(\epsilon)\ .$$

Hence, we have that:

$$(\pi - \theta_{i,j}^{i,l})\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle = \pi\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle + O(\epsilon^3)\ . \tag{49}$$

For the second expression we get:

$$\begin{aligned}
&\mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{i,j},\mathbf{w}_{i,l}}\bar{\mathbf{w}}_{i,l}^\top \mathbf{g}_{i,l} = \frac{\langle \bar{\mathbf{w}}_{i,l}, \mathbf{g}_{i,l}\rangle \cdot \left(\langle \bar{\mathbf{w}}_{i,j}, \mathbf{g}_{i,j}\rangle - \cos(\theta_{i,j}^{i,l})\langle \bar{\mathbf{w}}_{i,l}, \mathbf{g}_{i,j}\rangle\right)}{\sin(\theta_{i,j}^{a,b})} \\
&= \frac{\frac{\|\mathbf{g}_{i,l}\|^2}{\|\mathbf{w}_{i,l}\|}\left(\frac{\|\mathbf{g}_{i,j}\|^2}{\|\mathbf{w}_{i,j}\|} - \cos(\theta_{i,j}^{i,l})\frac{\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle}{\|\mathbf{w}_{i,l}\|}\right)}{\sin(\theta_{i,j}^{i,l})} \overset{(29),(30),(31)}{=} \frac{\frac{\|\mathbf{g}_{i,l}\|^2}{\frac{1}{m}+O(\epsilon)}\left(\frac{\|\mathbf{g}_{i,j}\|^2}{\frac{1}{m}+O(\epsilon)} - (1+O(\epsilon))\frac{\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle}{\frac{1}{m}+O(\epsilon)}\right)}{O(\sqrt{\epsilon})} \\
&= O(\epsilon^{3.5})\ .
\end{aligned} \tag{50}$$

For the third expression we have:

$$\begin{aligned}
&\mathbf{g}_{i,j}^\top \bar{n}_{\mathbf{w}_{i,l},\mathbf{w}_{i,j}}\bar{\mathbf{w}}_{i,j}^\top \mathbf{g}_{i,l} = \frac{\langle \bar{\mathbf{w}}_{i,j}, \mathbf{g}_{i,l}\rangle \cdot \left(\langle \bar{\mathbf{w}}_{i,l}, \mathbf{g}_{i,j}\rangle - \cos(\theta_{i,j}^{i,l})\langle \bar{\mathbf{w}}_{i,j}, \mathbf{g}_{i,j}\rangle\right)}{\sin(\theta_{i,j}^{i,l})} \\
&= \frac{\frac{\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle}{\|\mathbf{w}_{i,j}\|} \cdot \left(\frac{\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle}{\|\mathbf{w}_{i,l}\|} - \cos(\theta_{i,j}^{i,l})\frac{\|\mathbf{g}_{i,j}\|^2}{\|\mathbf{w}_{i,j}\|}\right)}{\sin(\theta_{i,j}^{i,l})} \overset{(29),(30),(31)}{=} \frac{\frac{\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle}{\frac{1}{m}+O(\epsilon)} \cdot \left(\frac{\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle}{\frac{1}{m}+O(\epsilon)} - (1+O(\epsilon))\frac{\|\mathbf{g}_{i,j}\|^2}{\frac{1}{m}+O(\epsilon)}\right)}{O(\sqrt{\epsilon})} \\
&= O(\epsilon^{3.5})\ .
\end{aligned} \tag{51}$$

Combining Eq. (49), Eq. (50) and Eq. (51) we get:

$$\mathbf{g}_{i,j}^\top h_2(\mathbf{w}_{i,j}, \mathbf{w}_{i,l})\mathbf{g}_{i,l} = \frac{1}{2\pi}\left(\pi\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle + O(\epsilon^3) + O(\epsilon^{3.5}) + O(\epsilon^{3.5})\right) = \frac{1}{2}\langle \mathbf{g}_{i,j}, \mathbf{g}_{i,l}\rangle + O(\epsilon^3)$$

■

Before combining all the parts we will also need the following technical lemma:

Lemma 25 *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{R}^n$, and denote by u_{ij} the $j-th$ coordinate of $\mathbf{u}_i$. Then:*

$$\sum_{i=1}^{n} \sum_{j \neq i} u_{ij} u_{ji} \geq \sum_{i=1}^{n} u_{ii}^2 - \sum_{i=1}^{n} \|\mathbf{u}_i\|^2$$

Proof Let U be a matrix with columns equal to $\mathbf{u}_i$. Note that:

$$tr(U^\top U) = \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij}^2$$
$$tr(U^2) = \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij} u_{ji}$$
$$tr(UU^\top) = \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ji}^2$$

Now we have that:

$$tr(U^\top U) + tr(U^2) = \frac{1}{2} tr(U^\top U) + tr(U^2) + \frac{1}{2} tr(UU^\top) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{1}{2} u_{ij}^2 + u_{ij} u_{ji} + \frac{1}{2} u_{ji}^2 \right) \tag{52}$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (u_{ij} + u_{ji})^2 \geq \frac{1}{2} \sum_{i=1}^{n} \sum_{j \neq i} (u_{ij} + u_{ji})^2 + 2 \sum_{i=1}^{n} u_{ii}^2 \geq 2 \sum_{i=1}^{n} u_{ii}^2$$

The two terms from the first part of Eq. (52) also have the following useful forms:

$$tr(U^\top U) = \sum_{i=1}^{n} \|\mathbf{u}_i\|^2$$
$$tr(U^2) = \sum_{i=1}^{n} \sum_{j \neq i} u_{ij} u_{ji} + \sum_{i=1}^{n} u_{ii}^2$$

In total we have:

$$\sum_{i=1}^{n} \sum_{j \neq i} u_{ij} u_{ji} + \sum_{i=1}^{n} u_{ii}^2 + \sum_{i=1}^{n} \|\mathbf{u}_i\|^2 \geq 2 \sum_{i=1}^{n} u_{ii}^2,$$

hence:

$$\sum_{i=1}^{n} \sum_{j \neq i} u_{ij} u_{ji} \geq \sum_{i=1}^{n} u_{ii}^2 - \sum_{i=1}^{n} \|\mathbf{u}_i\|^2.$$

■

We are now ready to prove the main theorem of this section:

Proof [Proof of Thm. 10] By Lemma 22 we have that

$$\sum_{i,j=1}^{k,m}\left(\sum_{\substack{a,b=1\\(a,b)\neq(i,j)}}^{k,m}\mathbf{g}_{i,j}^{\top}h_1(\mathbf{w}_{i,j},\mathbf{w}_{a,b})\mathbf{g}_{i,j}-\sum_{l=1}^{k}\mathbf{g}_{i,j}^{\top}h_1(\mathbf{w}_{i,j},\mathbf{v}_l)\mathbf{g}_{i,j}\right)=\sum_{i,j=1}^{k,m}O(\epsilon^{2.5})=O(\epsilon^{2.5})\,. \tag{53}$$

Applying the above to Eq. (28) we get:

$$\mathbf{g}_1^{n\top}H(\mathbf{w}_1^n)\mathbf{g}_1^n=\sum_{i,j=1}^{k,m}\left(\frac{1}{2}\|\mathbf{g}_{i,j}\|^2+\sum_{\substack{a,b=1\\(a,b)\neq(i,j)}}^{k,m}\mathbf{g}_{i,j}^{\top}h_2(\mathbf{w}_{i,j},\mathbf{w}_{a,b})\mathbf{g}_{a,b}\right)+O(\epsilon^{2.5})\,. \tag{54}$$

First we separate the expression inside the sum of Eq. (54) for the different values of i,a where either $i=a$ or $i\neq a$:

$$\begin{aligned}&\sum_{i,j=1}^{k,m}\sum_{\substack{a,b=1\\(a,b)\neq(i,j)}}^{k,m}\mathbf{g}_{i,j}^{\top}h_2(\mathbf{w}_{i,j},\mathbf{w}_{a,b})\mathbf{g}_{a,b}\\&=\sum_{i,j=1}^{k,m}\sum_{l\neq j}\mathbf{g}_{i,j}^{\top}h_2(\mathbf{w}_{i,j},\mathbf{w}_{i,l})\mathbf{g}_{i,l}+\sum_{i,j=1}^{k,m}\sum_{\substack{a,b=1\\a\neq i}}^{k,m}\mathbf{g}_{i,j}^{\top}h_2(\mathbf{w}_{i,j},\mathbf{w}_{a,b})\mathbf{g}_{a,b}\,.\end{aligned} \tag{55}$$

Recall that $\mathbf{g}_i=\sum_{j=1}^{m}\mathbf{g}_{i,j}$ and $\mathbf{g}=\sum_{i=1}^{k}\mathbf{g}_i$. For the first sum in Eq. (55) we fix $i\in[k]$ and use Lemma 24 to get:

$$\begin{aligned}&\sum_{j=1}^{m}\sum_{l\neq j}\mathbf{g}_{i,j}^{\top}h_2(\mathbf{w}_{i,j},\mathbf{w}_{i,l})\mathbf{g}_{i,l}=\sum_{j=1}\sum_{l\neq j}\frac{1}{2}\langle\mathbf{g}_{i,j},\mathbf{g}_{i,l}\rangle+O(\epsilon^3)=\frac{1}{2}\sum_{j=1}^{m}\langle\mathbf{g}_{i,j},\sum_{l\neq j}\mathbf{g}_{i,l}\rangle+O(\epsilon^3)\\&=\frac{1}{2}\sum_{j=1}^{m}\langle\mathbf{g}_{i,j},\mathbf{g}_i-\mathbf{g}_{i,j}\rangle+O(\epsilon^3)=\frac{1}{2}\left(\sum_{j=1}^{m}\langle\mathbf{g}_{i,j},\mathbf{g}_i\rangle-\sum_{j=1}^{m}\langle\mathbf{g}_{i,j},\mathbf{g}_{i,j}\rangle\right)+O(\epsilon^3)\\&=\frac{1}{2}\|\mathbf{g}_i\|^2-\frac{1}{2}\sum_{j=1}^{m}\|\mathbf{g}_{i,j}\|^2+O(\epsilon^3)\end{aligned} \tag{56}$$

Summing for all $i\in[k]$ we get:

$$\sum_{i,j=1}^{k,m}\sum_{l\neq j}\mathbf{g}_{i,j}^{\top}h_2(\mathbf{w}_{i,j},\mathbf{w}_{i,l})\mathbf{g}_{i,l}=\frac{1}{2}\sum_{i=1}^{k}\|\mathbf{g}_i\|^2-\frac{1}{2}\sum_{i,j=1}^{k,m}\|\mathbf{g}_{i,j}\|^2+O(\epsilon^3) \tag{57}$$

For the second sum in Eq. (55) we use Lemma 23 to get:

$$\sum_{i,j=1}^{k,m}\sum_{\substack{a,b=1\\a\neq i}}^{k,m}\mathbf{g}_{i,j}^{\top}h_2(\mathbf{w}_{i,j},\mathbf{w}_{a,b})\mathbf{g}_{a,b}=\sum_{i,j=1}^{k,m}\sum_{\substack{a,b=1\\a\neq i}}^{k,m}\left(\frac{1}{4}\langle\mathbf{g}_{i,j},\mathbf{g}_{a,b}\rangle+\frac{1}{2\pi}\langle\mathbf{g}_{i,j},\mathbf{v}_a\rangle\cdot\langle\mathbf{g}_{a,b},\mathbf{v}_i\rangle\right)+O(\epsilon^{2.5}) \tag{58}$$

We will bound the two expressions in Eq. (58). For the first expression we use the same calculation as Eq. (56) to get:

$$\sum_{i,j=1}^{k,m} \sum_{\substack{a,b=1 \\ a\neq i}}^{k,m} \frac{1}{4}\langle \mathbf{g}_{i,j}, \mathbf{g}_{a,b}\rangle = \sum_{i=1}^{k}\sum_{a\neq i}\frac{1}{4}\langle \mathbf{g}_i, \mathbf{g}_a\rangle = \frac{1}{4}\|\mathbf{g}\|^2 - \frac{1}{4}\sum_{i=1}^{k}\|\mathbf{g}_i\|^2 . \tag{59}$$

For the second expression in Eq. (58) we first simplify:

$$\frac{1}{2\pi}\sum_{i,j=1}^{k,m} \sum_{\substack{a,b=1 \\ a\neq i}}^{k,m} \langle \mathbf{g}_{i,j}, \mathbf{v}_i\rangle \cdot \langle \mathbf{g}_{a,b}, \mathbf{v}_i\rangle = \frac{1}{2\pi}\sum_{i=1}^{k}\sum_{a\neq i}\langle \mathbf{g}_i, \mathbf{v}_a\rangle \cdot \langle \mathbf{g}_a, \mathbf{v}_i\rangle$$

Denote by $\mathbf{u}_1,\dots,\mathbf{u}_k \in \mathbb{R}^k$ the vectors where the $a-th$ coordinate of $\mathbf{u}_i$ is equal to $u_{ia} = \langle \mathbf{g}_i, \mathbf{v}_a\rangle$ for $a = 1,\dots,k$. Since $\mathbf{g}\in\mathbb{R}^d$ with $d \geq k$ and the vectors $\mathbf{v}_1,\dots,\mathbf{v}_k$ are orthonormal we have that $\|\mathbf{u}_i\|^2 = \sum_{a=1}^k \langle \mathbf{g}_i, \mathbf{v}_a\rangle^2 \leq \|\mathbf{g}_i\|^2$. Also, by the assumptions of the theorem we have $u_{ii} = \langle \mathbf{g}_i, \mathbf{v}_i\rangle = 0$. Using Lemma 25 on the vectors $\mathbf{u}_1,\dots,\mathbf{u}_k$ we have:

$$\frac{1}{2\pi}\sum_{i=1}^{k}\sum_{a\neq i}\langle \mathbf{g}_i, \mathbf{v}_a\rangle \cdot \langle \mathbf{g}_a, \mathbf{v}_i\rangle = \frac{1}{2\pi}\sum_{i=1}^{k}\sum_{a\neq i} u_{ia}u_{ai} \geq \frac{1}{2\pi}\left(\sum_{i=1}^{k} u_{ii}^2 - \sum_{i=1}^{k}\|\mathbf{u}_i\|^2\right) \geq -\frac{1}{2\pi}\sum_{i=1}^{k}\|\mathbf{g}_i\|^2 + O(\epsilon^{2.5}) \tag{60}$$

Returning to Eq. (58), we combine Eq. (59) and Eq. (60) to get:

$$\sum_{i,j=1}^{k,m} \sum_{\substack{a,b=1 \\ a\neq i}}^{k,m} \mathbf{g}_{i,j}^\top h_2(\mathbf{w}_{i,j}, \mathbf{w}_{a,b})\mathbf{g}_{a,b} \geq \frac{1}{4}\|\mathbf{g}\|^2 - \frac{1}{4}\sum_{i=1}^{k}\|\mathbf{g}_i\|^2 - \frac{1}{2\pi}\sum_{i=1}^{k}\|\mathbf{g}_i\|^2$$

Combining the above with Eq. (28) and Eq. (57) to get:

$$\begin{aligned}
&\mathbf{g}_1^{n\top} H(\mathbf{w}_1^n)\mathbf{g}_1^n \\
&\geq \frac{1}{2}\sum_{i,j=1}^{k,m}\|\mathbf{g}_{i,j}\|^2 + \frac{1}{2}\sum_{i=1}^{k}\|\mathbf{g}_i\|^2 - \frac{1}{2}\sum_{i,j=1}^{k,m}\|\mathbf{g}_{i,j}\|^2 + O(\epsilon^3) + \frac{1}{4}\|\mathbf{g}\|^2 - \left(\frac{1}{4}+\frac{1}{2\pi}\right)\cdot\sum_{i=1}^{k}\|\mathbf{g}_i\|^2 + O(\epsilon^{2.5}) \\
&\geq \frac{1}{4}\|\mathbf{g}\|^2 + \left(\frac{1}{2}-\frac{1}{4}-\frac{1}{2\pi}\right)\sum_{i=1}^{k}\|\mathbf{g}_i\|^2 + O(\epsilon^{2.5}) \geq \frac{1}{4}\|\mathbf{g}\|^2 + \left(\frac{1}{4}-\frac{1}{2\pi}\right)\sum_{i=1}^{k}\|\mathbf{g}_i\|^2 + O(\epsilon^{2.5})
\end{aligned}$$

■

Appendix D. Optimization Proofs

D.1. Empirical Investigation of Conjecture 12

In this section, we wish to empirically verify the correctness of Conjecture 12. Our method is to focus on a small neighborhood of a global minimum, sample the vectors $\mathbf{g}_{i,j}$ either randomly or adversarially, and verify that indeed the l.h.s is larger than the r.h.s, even without the $O(\epsilon^{2.5})$ term, and in a standard ϵ-neighborhood of the global minimum. To this end, we sampled 1,000 random Gaussian vectors independently with zero mean and variance 10^{-5}, for each pair of values $k \in \{5, 10, 20\}$ and $m \in \{2, 5, 10\}$, and computed the difference between the left-hand side and right-hand side of Conjecture 12. Additionally, we also tested the above difference in an adversarial sample, when the right-hand side cancels out, i.e. when $\sum_{j=1}^{m} \mathbf{g}_{i,j} = \mathbf{0}$ for all $i \in [k]$, by replacing the noise vector $\mathbf{g}_{i,1}$ with $-\sum_{j=2}^{m} \mathbf{g}_{i,j}$. In *all* the samples made, the difference computed was strictly positive. This means that empirically for many global minima (i.e. many k and m) the conjecture holds. Table 1 summarizes the smallest curvature (i.e. the l.h.s of Conjecture 12) found among the 1,000 random samples performed on each pair.

k	m	Gaussian $\mathbf{g}_{i,j}$	$\mathbf{g}_{i,1} = -\sum_{j=2}^{m} \mathbf{g}_{i,j}$
5	2	0.84	$2.27 \cdot 10^{-4}$
5	5	2.43	0.00754
5	10	5.43	0.0693
10	2	12.7	0.0037
10	5	43.1	0.088
10	10	62.3	0.733
20	2	88.6	0.0318
20	5	231	0.67
20	10	412	5.15

Table 1: The minimal curvature measured among 1,000 random samples, for different values of k and m. The third column is the minimal curvature measured for multivariate Normally-distributed noise $\mathbf{g}_{i,j}$, and the rightmost column is the curvature measured for noise satisfying $\sum_{j=1}^{m} \mathbf{g}_{i,j} = \mathbf{0}$. While the curvature is lower bounded by roughly 1 for Gaussian noise, we see that the more problematic points in the neighborhood of the global minimum where the right-hand side of Conjecture 12 cancels out have curvature several magnitudes smaller. In both cases, the curvature increases at least linearly with the over-parameterization factor m, which further demonstrates the benefits of over-parameterization for making the loss surface more benign.

D.2. Perturbed Gradient Descent

To show an optimization guarantee, we use the following form of the perturbed gradient descent algorithm:

Input: $\mathbf{w}_1^n(0),\ \eta, \alpha > 0,\ T \in \mathbb{N}$
for $t = 1, 2, \ldots, T$ **do**
  Sample $\xi \sim \mathcal{N}\left(\mathbf{0}_d, \frac{1}{\sqrt{d}} I\right)$
  Set $\hat{\mathbf{w}}_1^n(t) := (\mathbf{w}_1(t) + \alpha\xi, \ldots, \mathbf{w}_n(t) + \alpha\xi)$
  Update $\mathbf{w}_1^n(t+1) = \hat{\mathbf{w}}_1^n(t) - \eta\nabla F(\hat{\mathbf{w}}_1^n(t))$
end
Return $\mathbf{w}_1^n(T)$

Algorithm 1: Perturbed gradient descent

Algorithm D.2 inputs an initialized weights $\mathbf{w}_1^n(0)$, a learning rate η and noise level α. At each iteration the algorithm updates the weights w.r.t the loss function F similarly to gradient descent, and adds a perturbation in a random direction with magnitude α. Note that the same perturbation direction is the same for all the learned vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$.

We believe it is possible to achieve better optimization guarantees if in Algorithm D.2, a different noise direction would be added to each $\mathbf{w}_i$. This would also require a more careful analysis, as there is a positive (small) probability that adding up all the noise vectors would produce a vector with small norm.

D.3. Proof of Thm. 13

We will first lower bound the norm of $\mathbf{g}$:

$$\begin{aligned}
\|\mathbf{g}\|^2 &= \left\|\sum_{i=1}^n \mathbf{w}_i + \alpha\xi\right\|^2 \\
&= \left\|\sum_{i=1}^n \mathbf{w}_i\right\|^2 + \left\|\sum_{i=1}^n \alpha\xi\right\|^2 + 2\langle\sum_{i=1}^n \alpha\xi, \sum_{i=1}^n \mathbf{w}_i\rangle \\
&\geq \left\|\sum_{i=1}^n \mathbf{w}_i\right\|^2 + n^2\alpha^2\|\xi\|^2 - 2n\alpha\left|\langle\xi, \sum_{i=1}^n \mathbf{w}_i\rangle\right|
\end{aligned} \tag{61}$$

Assume that $\|\sum_{i=1}^n \mathbf{w}_i\| \geq \delta$, then we can use Cauchy-Schwartz to bound Eq. (61) by:

$$\begin{aligned}
\|\mathbf{g}\|^2 &\geq \left\|\sum_{i=1}^n \mathbf{w}_i\right\|^2 - 2n\alpha\|\xi\| \cdot \left\|\sum_{i=1}^n \mathbf{w}_i\right\| \\
&\geq \left\|\sum_{i=1}^n \mathbf{w}_i\right\| \cdot (\delta - 2n\alpha\|\xi\|) \geq \delta \cdot (\delta - 2n\alpha\|\xi\|) \,.
\end{aligned}$$

Note that $\|\xi\|^2$ has a χ_d^2 distribution. By standard concentration bound on the χ^2 distribution, w.p $> 1 - e^{-\Omega(d)}$ we have that $\|\xi\| \leq 1.5$, taking this event into account and substituting α we get that $\|\mathbf{g}\| \geq \frac{\delta^2}{4}$.

Now, assume that $\|\sum_{i=1}^{n}\mathbf{w}_i\| < \delta$, and denote $\mathbf{u} = \sum_{i=1}^{n}\mathbf{w}_i$, we can bound Eq. (61) by:

$$\begin{aligned}\|\mathbf{g}\|^2 &\geq n^2\alpha^2\|\xi\|^2 - 2n\alpha\left|\langle\xi, \sum_{i=1}^{n}\mathbf{w}_i\rangle\right| \\ &\geq n^2\alpha^2\|\xi\|^2 - 2n\alpha\|\mathbf{u}\|\,|\langle\xi, \bar{\mathbf{u}}\rangle| \\ &\geq n^2\alpha^2\|\xi\|^2 - 2n\alpha\delta\,|\langle\xi, \bar{\mathbf{u}}\rangle| \\ &= \frac{\delta^2}{16}\|\xi\|^2 - \frac{\delta^2}{2}\,|\langle\xi, \bar{\mathbf{u}}\rangle|\ .\end{aligned}$$

Since ξ has a spherically symmetric distribution, independent of $\mathbf{u}$, we can assume w.l.o.g that $\bar{\mathbf{u}}$ is a standard unit vector, hence $\langle\xi, \bar{\mathbf{u}}\rangle \sim \mathcal{N}\left(0, \frac{1}{\sqrt{d}}\right)$. Again, using standard concentration bounds on both the distribution of $\|\xi\|^2$ and $\langle\xi, \bar{\mathbf{u}}\rangle$, and applying union bound, we get that w.p $> 1 - e^{-\Omega(d)}$ we have that $\|\xi\|^2 > 0.5$ and $\langle\xi, \bar{\mathbf{u}}\rangle \leq \frac{1}{32}$. Taking those bounds into account we get that $\|\mathbf{g}\|^2 \geq \frac{\delta^2}{32}$. In total, from both cases we get that w.p $> 1 - e^{-\Omega(d)}$ we have that $\|\mathbf{g}\| \geq \frac{\delta^2}{32}$.

Applying the above bound to Eq. (4) we get that w.p $> 1 - e^{-\Omega(d)}$:

$$(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n)^\top H(\mathbf{w}_1^n)(\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n) \geq \lambda\|\mathbf{g}\|^2 \geq \frac{\lambda\delta^2}{32}$$

By the assumption that the function is twice differentiable, the above bound translates to the following bound on the gradient w.p $> 1 - e^{-\Omega(d)}$:

$$\langle\nabla F(\mathbf{w}_1^n), \mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n\rangle \geq \frac{\lambda\delta^2}{32}\cdot\|\mathbf{w}_1^n - \tilde{\mathbf{w}}_1^n\|^2\ . \tag{62}$$

Now we can bound the iterates of gradient descent, conditioning on the event of Eq. (62):

$$\begin{aligned}\|\mathbf{w}_1^n(t+1) - \tilde{\mathbf{w}}_1^n\|^2 &= \|\mathbf{w}_1^n(t) - \eta\nabla F(\mathbf{w}_1^n(t)) + \tilde{\mathbf{w}}_1^n\|^2 \\ &= \|\mathbf{w}_1^n(t) - \tilde{\mathbf{w}}_1^n\|^2 - \eta\langle\nabla F(\mathbf{w}_1^n(t)), \mathbf{w}_1^n(t) - \tilde{\mathbf{w}}_1^n\rangle + \eta^2\|\nabla F(\mathbf{w}_1^n(t))\|^2 \\ &\leq \|\mathbf{w}_1^n(t) - \tilde{\mathbf{w}}_1^n\|^2 - \frac{\eta\lambda\delta^2}{32}\cdot\|\mathbf{w}_1^n(t) - \tilde{\mathbf{w}}_1^n\|^2 + \eta^2L^2\|\mathbf{w}_1^n(t) - \tilde{\mathbf{w}}_1^n\|^2 \qquad (63) \\ &= \|\mathbf{w}_1^n(t) - \tilde{\mathbf{w}}_1^n\|^2\left(1 - \frac{\eta\lambda\delta^2}{32} + \eta^2L^2\right) \\ &\leq \|\mathbf{w}_1^n(t) - \tilde{\mathbf{w}}_1^n\|^2\left(1 - \frac{\eta\lambda\delta^2}{64}\right)\end{aligned}$$

where in Eq. (63) we used Eq. (62), the assumption that the gradient of F is Lipschitz and that $\tilde{\mathbf{w}}_1^n$ is a global minima (hence $\nabla F(\tilde{\mathbf{w}}_1^n) = \mathbf{0}$). By induction on t we get that:

$$\|\mathbf{w}_1^n(t) - \tilde{\mathbf{w}}_1^n\|^2 \leq \|\mathbf{w}_1^n(0) - \tilde{\mathbf{w}}_1^n\|^2\left(1 - \frac{\eta\lambda\delta^2}{64}\right)^t\ .$$

Recall that we initialized in an ϵ neighborhood of $\tilde{\mathbf{w}}_1^n$ for $\epsilon < 1$, hence $\|\mathbf{w}_1^n(0) - \tilde{\mathbf{w}}_1^n\|^2 < 1$. Using union bound, after $T > \frac{\log(\delta)}{\log\left(1 - \frac{\eta\lambda\delta^2}{64}\right)}$ iterations, w.p $> 1 - Te^{-\Omega(d)}$ we get that $\|\mathbf{w}_1^n(T) - \tilde{\mathbf{w}}_1^n\|^2 \leq \delta$.

Appendix E. Proofs from Sec. 4

Before we prove Thm. 14, we will first state and prove some auxiliary lemmas.

Lemma 26 *For any $n \geq 1$, the origin is neither a local minimum nor a local maximum of Eq. (2).*

Proof Assume $\mathbf{w}_1^n = \mathbf{0}$ is the origin. Consider the point $\tilde{\mathbf{w}}_1^n = (\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_n)$ where $\tilde{\mathbf{w}}_1 = \epsilon \mathbf{v}_1$ for some real ϵ, and $\tilde{\mathbf{w}}_i = 0$ for any $i \in \{2, \dots, n\}$. Recall the closed-form of the objective in Eq. (2), given in Safran and Shamir (29, Section 4.1.1) by

$$F(\mathbf{w}_1^n) = \frac{1}{2} \sum_{i,j=1}^{n} f(\mathbf{w}_i, \mathbf{w}_j) - \sum_{\substack{i \in [n] \\ j \in [k]}} f(\mathbf{w}_i, \mathbf{v}_j) + \frac{1}{2} \sum_{i,j=1}^{k} f(\mathbf{v}_i, \mathbf{v}_j), \tag{64}$$

where

$$f(\mathbf{w}, \mathbf{v}) = \frac{1}{2\pi} \|\mathbf{w}\| \|\mathbf{v}\| \left(\sin(\theta_{\mathbf{w},\mathbf{v}}) + (\pi - \theta_{\mathbf{w},\mathbf{v}}) \cos(\theta_{\mathbf{w},\mathbf{v}}) \right). \tag{65}$$

Next, Eq. (65) reveals that $f(\mathbf{w}, \mathbf{v}) \geq 0$ for any two vectors $\mathbf{w}, \mathbf{v}$, and from Eq. (64) we have

$$\begin{aligned} F(\mathbf{w}_1^n) - F(\tilde{\mathbf{w}}_1^n) &= \frac{1}{2} \sum_{i,j=1}^{k} f(\mathbf{v}_i, \mathbf{v}_j) - \left(\frac{1}{2} \sum_{i,j=1}^{k} f(\mathbf{v}_i, \mathbf{v}_j) + \frac{1}{2} f(\epsilon \mathbf{v}_1, \mathbf{v}_1) - \sum_{j \in [k]} f(\epsilon \mathbf{v}_1, \mathbf{v}_j) \right) \\ &= \epsilon \sum_{j=2}^{n} f(\mathbf{v}_1, \mathbf{v}_j) + \frac{1}{2} \epsilon f(\mathbf{v}_1, \mathbf{v}_1) = \epsilon \left(\sum_{j=2}^{n} f(\mathbf{v}_1, \mathbf{v}_j) + \frac{1}{2} f(\mathbf{v}_1, \mathbf{v}_1) \right), \end{aligned}$$

and since

$$c := \left(\sum_{j=2}^{n} f(\mathbf{v}_1, \mathbf{v}_j) + \frac{1}{2} f(\mathbf{v}_1, \mathbf{v}_1) \right) \geq \frac{1}{4} \|\mathbf{v}_1\|^2 = \frac{1}{4},$$

we have that $F(\mathbf{w}_1^n) - F(\tilde{\mathbf{w}}_1^n) = \epsilon c \to 0_+$ if $\epsilon \to 0_+$ and $F(\mathbf{w}_1^n) - F(\tilde{\mathbf{w}}_1^n) = \epsilon c \to 0_-$ if $\epsilon \to 0_-$, therefore we can approach $\mathbf{w}_1^n$ from two different directions where in one the objective is strictly increasing and in the other it is strictly decreasing, hence $\mathbf{w}_1^n$ is neither a local minimum nor a local maximum. ■

Lemma 27 *For any $n \geq 1$, the objective in Eq. (2) has no local maxima.*

Proof Let $\mathbf{w}_1^n = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ and for $t \geq 0$ define $\mathbf{w}_1^n(t) = (t\mathbf{w}_1, \dots, t\mathbf{w}_n)$. We have from Equations (64) and (65) that

$$F(\mathbf{w}_1^n(t)) = \frac{1}{2} t^2 \sum_{i,j=1}^{n} f(\mathbf{w}_i, \mathbf{w}_j) - t \sum_{\substack{i \in [n] \\ j \in [k]}} f(\mathbf{w}_i, \mathbf{v}_j) + \frac{1}{2} \sum_{i,j=1}^{k} f(\mathbf{v}_i, \mathbf{v}_j), \tag{66}$$

hence the objective is quadratic as a function of t.

Assuming $\mathbf{w}_1^n$ is not the origin, we have from Eq. (65) that

$$\sum_{i,j=1}^{n} f(\mathbf{w}_i, \mathbf{w}_j) > 0,$$

thus from the above and Eq. (66) the objective is strongly convex in t, and therefore cannot attain a local maximum at $\mathbf{w}_1^n = \mathbf{w}_1^n(1)$. Otherwise, if $\mathbf{w}_1^n$ is the origin, then from Lemma 26 it is not a local maximum. ■

Lemma 28 *Suppose $n \geq 1$, $\mathbf{w}_1^n = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$ is a differentiable local minimum of the objective in Eq. (2). Then for all $\alpha \in (0,1)$ and any $i \in [n]$, the point $\mathbf{w}_1^n(\alpha, i)$ is a critical point of F, and the Hessian of F at $\mathbf{w}_1^n(\alpha, i)$ is given in terms of the blocks of $H(\mathbf{w}_1^n)$ by*

$$\begin{pmatrix}
H_{1,1} & \cdots & H_{1,i-1} & H_{1,i} & H_{1,i} & H_{1,i+1} & \cdots & H_{1,n} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
H_{i-1,1} & \cdots & H_{i-1,i-1} & H_{i-1,i} & H_{i-1,i} & H_{i-1,i+1} & \cdots & H_{i-1,n} \\
H_{i,1} & \cdots & H_{i,i-1} & \frac{1}{2}I + \frac{1}{\alpha}H'_{i,i} & \frac{1}{2}I & H_{i,i+1} & \cdots & H_{i,n} \\
H_{i,1} & \cdots & H_{i,i-1} & \frac{1}{2}I & \frac{1}{2}I + \frac{1}{1-\alpha}H'_{i,i} & H_{i,i+1} & \cdots & H_{i,n} \\
H_{i+1,1} & \cdots & H_{i+1,i-1} & H_{i+1,i} & H_{i+1,i} & H_{i+1,i+1} & \cdots & H_{i+1,n} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
H_{n,1} & \cdots & H_{n,i-1} & H_{n,i} & H_{n,i} & H_{n,i+1} & \cdots & H_{n,n}
\end{pmatrix}.$$

Proof From the gradient of the objective in Eq. (10) and Eq. (11) and since $\mathbf{w}_1^n$ is a local minimum, we have for all $i \in [n]$ that

$$\frac{1}{2}\mathbf{w}_i + \sum_{j \neq i} g(\mathbf{w}_i, \mathbf{w}_j) - \sum_{j=1}^{k} g(\mathbf{w}_i, \mathbf{v}_j) = \mathbf{0}.$$

We begin with asserting that for all $\alpha \in (0,1)$, $\mathbf{w}_1^n(\alpha, i)$ is a critical point of F. First, from Brutzkus and Globerson (8, Lemma 3.2), F is differentiable for all $\alpha \in (0,1)$, therefore the gradient is well-defined. For $m \in [n+1] \setminus \{i, i+1\}$ we have

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}_m} F(\mathbf{w}_1^n(\alpha, i)) &= \frac{1}{2}\mathbf{w}_m + \sum_{j \in [n] \setminus \{i,m\}} g(\mathbf{w}_m, \mathbf{w}_j) + g(\mathbf{w}_m, \alpha \mathbf{w}_i) + g(\mathbf{w}_m, (1-\alpha)\mathbf{w}_i) - \sum_{j=1}^{k} g(\mathbf{w}_m, \mathbf{v}_j) \\
&= \frac{1}{2}\mathbf{w}_m + \sum_{j \in [n] \setminus \{i,m\}} g(\mathbf{w}_m, \mathbf{w}_j) + \alpha g(\mathbf{w}_m, \mathbf{w}_i) + (1-\alpha) g(\mathbf{w}_m, \mathbf{w}_i) - \sum_{j=1}^{k} g(\mathbf{w}_m, \mathbf{v}_j) \\
&= \frac{1}{2}\mathbf{w}_m + \sum_{j \in [n] \setminus \{i,m\}} g(\mathbf{w}_m, \mathbf{w}_j) + g(\mathbf{w}_m, \mathbf{w}_i) - \sum_{j=1}^{k} g(\mathbf{w}_m, \mathbf{v}_j) \\
&= \frac{1}{2}\mathbf{w}_m + \sum_{j \in [n] \setminus \{m\}} g(\mathbf{w}_m, \mathbf{w}_j) - \sum_{j=1}^{k} g(\mathbf{w}_m, \mathbf{v}_j) = \mathbf{0}.
\end{aligned}$$

For i we have

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}_i} F(\mathbf{w}_1^n(\alpha,i)) &= \frac{1}{2}\alpha\mathbf{w}_i + \sum_{j\in[n]\setminus\{i\}} g(\alpha\mathbf{w}_i,\mathbf{w}_j) + g(\alpha\mathbf{w}_i,(1-\alpha)\mathbf{w}_i) - \sum_{j=1}^{k} g(\mathbf{w}_i,\mathbf{v}_j)\\ &= \frac{1}{2}\alpha\mathbf{w}_i + \sum_{j\in[n]\setminus\{i\}} g(\mathbf{w}_i,\mathbf{w}_j) + (1-\alpha)g(\mathbf{w}_i,\mathbf{w}_i) - \sum_{j=1}^{k} g(\mathbf{w}_i,\mathbf{v}_j)\\ &= \frac{1}{2}\alpha\mathbf{w}_i + \sum_{j\in[n]\setminus\{i\}} g(\mathbf{w}_i,\mathbf{w}_j) + \frac{1}{2}(1-\alpha)\mathbf{w}_i - \sum_{j=1}^{k} g(\mathbf{w}_i,\mathbf{v}_j)\\ &= \frac{1}{2}\mathbf{w}_i + \sum_{j\in[n]\setminus\{i\}} g(\mathbf{w}_i,\mathbf{w}_j) - \sum_{j=1}^{k} g(\mathbf{w}_i,\mathbf{v}_j) = \mathbf{0},\end{aligned}$$

where likewise a similar computation for $\mathbf{w}_{i+1}$ shows that $\frac{\partial}{\partial \mathbf{w}_{i+1}} F(\mathbf{w}_1^n(\alpha,i)) = \mathbf{0}$. Turning to the Hessian, we have from Lemma 1 that F at $\mathbf{w}_1^n(\alpha,i)$ is twice differentiable. We then have by Thm. 17 that all off-diagonal blocks other than $H_{i,i+1}$, and $H_{i+1,i}$ remain the same as in $H(\mathbf{w}_1^n)$, since $h_2(\mathbf{u}_1,\mathbf{u}_2)$ isn't affected by linearly rescaling $\mathbf{u}_1,\mathbf{u}_2$. For the remaining two off-diagonal blocks we have from Lemma 18 that each is $\frac{1}{2}I$, and lastly we compute the diagonal blocks, starting with the m-th block $H_{m,m}$ where $m \in [n+1]\setminus\{i,i+1\}$. We have

$$\begin{aligned}H_{m,m}(\mathbf{w}_1^n(\alpha,i)) &= \frac{1}{2}I + \sum_{j\in[n]\setminus\{i,m\}} h(\mathbf{w}_m,\mathbf{w}_j) + h(\mathbf{w}_m,\alpha\mathbf{w}_i) + h(\mathbf{w}_m,(1-\alpha)\mathbf{w}_i) - \sum_{j=1}^{k} h(\mathbf{w}_m,\mathbf{v}_j)\\ &= \frac{1}{2}I + \sum_{j\in[n]\setminus\{i,m\}} h(\mathbf{w}_m,\mathbf{w}_j) + \alpha h(\mathbf{w}_m,\mathbf{w}_i) + (1-\alpha)h(\mathbf{w}_m,\mathbf{w}_i) - \sum_{j=1}^{k} h(\mathbf{w}_m,\mathbf{v}_j)\\ &= \frac{1}{2}I + \sum_{j\in[n]\setminus\{i,m\}} h(\mathbf{w}_m,\mathbf{w}_j) + h(\mathbf{w}_m,\mathbf{w}_i) - \sum_{j=1}^{k} h(\mathbf{w}_m,\mathbf{v}_j)\\ &= \frac{1}{2}I + \sum_{j\in[n]\setminus\{m\}} h(\mathbf{w}_m,\mathbf{w}_j) - \sum_{j=1}^{k} h(\mathbf{w}_m,\mathbf{v}_j).\end{aligned}$$

That is, $H_{m,m}(\mathbf{w}_1^n(\alpha,i))$ equals $H_{m,m}(\mathbf{w}_1^n)$ for $m \in [i-1]$ and equals $H_{m-1,m-1}(\mathbf{w}_1^n)$ for $m \in \{i+2,\ldots,n+1\}$. For the i-th block $H_{i,i}$ we have

$$\begin{aligned}H_{i,i}(\mathbf{w}_1^n(\alpha,i)) &= \frac{1}{2}I + \sum_{j\in[n]\setminus\{i\}} h(\alpha\mathbf{w}_i,\mathbf{w}_j) + h(\alpha\mathbf{w}_i,(1-\alpha)\mathbf{w}_i) - \sum_{j=1}^{k} h(\alpha\mathbf{w}_i,\mathbf{v}_j)\\ &= \frac{1}{2}I + \frac{1}{\alpha}\sum_{j\in[n]\setminus\{i\}} h(\mathbf{w}_i,\mathbf{w}_j) - \frac{1}{\alpha}\sum_{j=1}^{k} h(\mathbf{w}_i,\mathbf{v}_j)\\ &= \frac{1}{2}I + \frac{1}{\alpha}H'_{i,i}(\mathbf{w}_1^n),\end{aligned}$$

where the second equality is due to Lemma 1, and likewise, a similar computation reveals that

$$H_{i+1,i+1}(\mathbf{w}_1^n(\alpha,1)) = \frac{1}{1-\alpha} H'_{i,i}(\mathbf{w}_1^n),$$

concluding the proof of the lemma. ■

Lemma 29 *Suppose $n \geq 1$, $\mathbf{w}_1^n = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$ is a differentiable local minimum of the objective in Eq. (2) such that there exists $i \in [n]$ with component $H'_{i,i}$ satisfying $\mathbf{u}^\top H'_{i,i}\mathbf{u} = \lambda$ for some unit vector $\mathbf{u} \in \mathbb{R}^d$ and scalar $\lambda < 0$. Then for any $\alpha \in (0,1)$, $\mathbf{w}_1^n(\alpha, i)$ is a saddle point. Moreover, for $\alpha \in \{0,1\}$, $\mathbf{w}_1^n(\alpha, i)$ is not a local minimum of F.*

Proof The key in proving the lemma is that the i-th diagonal block of the Hessian at $\mathbf{w}_1^n$ (having dimensions $d \times d$ and given by $0.5I + H'_{i,i}$) is turned into a $2d \times 2d$ block of the following form:

$$\begin{pmatrix} \frac{1}{2}I + \frac{1}{\alpha}H'_{i,i} & \frac{1}{2}I \\ \frac{1}{2}I & \frac{1}{2}I + \frac{1}{1-\alpha}H'_{i,i} \end{pmatrix}. \tag{67}$$

Next, we show that $H'_{i,i}$ is not PSD, hence the block matrix above is not PSD, and consequentially the Hessian at $\mathbf{w}_1^n(\alpha, i)$ is not PSD, implying the lemma. For the case where $\alpha \in (0,1)$, by using Lemma 28, we have that this is a critical point, and from Lemma 27 it is not a local maximum. We multiply the $2d \times 2d$ block of the Hessian at $\mathbf{w}_1^n(\alpha, i)$ given in Eq. (67) by $(\mathbf{u}, -\mathbf{u}) \in \mathbb{R}^{2d}$ from both sides and obtain

$$(\mathbf{u}, -\mathbf{u}) \begin{pmatrix} \frac{1}{2}I + \frac{1}{\alpha}H'_{i,i} & \frac{1}{2}I \\ \frac{1}{2}I & \frac{1}{2}I + \frac{1}{1-\alpha}H'_{i,i} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ -\mathbf{u} \end{pmatrix} = \frac{1}{\alpha}\mathbf{u}^\top H'_{i,i}\mathbf{u} + \frac{1}{1-\alpha}\mathbf{u}^\top H'_{i,i}\mathbf{u} = \lambda\left(\frac{1}{\alpha} + \frac{1}{1-\alpha}\right) < 0.$$

Next, letting $\tilde{\mathbf{u}} \in \mathbb{R}^{(n+1)d}$ be the all-zero vector, except for entries $(i-1)d+1$ to $(i+1)d$ which equal $(\mathbf{u}, -\mathbf{u})$, then

$$\tilde{\mathbf{u}}^\top H(\mathbf{w}_1^n(\alpha, i))\tilde{\mathbf{u}} = \lambda\left(\frac{1}{\alpha} + \frac{1}{1-\alpha}\right) < 0,$$

thus $H(\mathbf{w}_1^n(\alpha, i))$ is not a PSD matrix.

For the case where $\alpha \in \{0,1\}$, since the objective is not differentiable in this case, we will show that the point cannot be a local minimum by showing that in any neighborhood containing it also contains a point with a strictly smaller objective.

Assuming $\alpha = 0$, we have from the above derivation that there exists $i \in [n]$ such that for all $\alpha' \in (0,1)$, $\mathbf{w}_1^n(\alpha', i)$ is not a local minimum. In particular, given some $\delta > 0$, choose $\alpha' > 0$ small enough so that $\|\mathbf{w}_1^n(\alpha', i) - \mathbf{w}_1^n(0, i)\| \leq \delta/2$. Since $\mathbf{w}_1^n(\alpha', i)$ is not a local minimum, there exists $\tilde{\mathbf{w}}_1^{n+1}$ such that $\|\mathbf{w}_1^n(\alpha', i) - \tilde{\mathbf{w}}_1^{n+1}\| \leq \delta/2$ and $F(\tilde{\mathbf{w}}_1^{n+1}) < F(\mathbf{w}_1^n(\alpha', i))$. Since

$$\left[\langle \alpha'\mathbf{w}, \mathbf{x}\rangle\right]_+ + \left[\langle (1-\alpha')\mathbf{w}, \mathbf{x}\rangle\right]_+ = [\langle \mathbf{w}, \mathbf{x}\rangle]_+$$

for all $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$ and $\alpha' \in [0,1]$, this entails

$$F(\tilde{\mathbf{w}}_1^{n+1}) < F(\mathbf{w}_1^n(\alpha', i)) = F(\mathbf{w}_1^n(0, i))$$

and $\|\tilde{\mathbf{w}}_1^{n+1} - \mathbf{w}_1^n(0, i)\| \leq \delta$, hence $\mathbf{w}_1^n(0, i)$ is not a local minimum.

Finally, the case for $\alpha = 1$ follows from the $\alpha = 0$ case by permuting the neurons. ■

We are now ready to prove Thm. 14.

Proof [Proof of Thm. 14] Throughout the proof we will assume that F is differentiable at $\mathbf{w}_1^n$, which by Lemma 1 also implies that it is twice continuously differentiable there, and as would be evident later in the proof we will see that this is necessarily the case.

We will show that there exist $i \in [n]$ and some unit vector $\mathbf{u}$ such that

$$\lambda \coloneqq \mathbf{u}^\top H'_{i,i}\mathbf{u} = \sum_{j\in[n]\setminus\{i\}} \frac{\sin(\theta_{\mathbf{w}_i,\mathbf{w}_j})\|\mathbf{w}_j\|}{2\pi\|\mathbf{w}_i\|} - \sum_{j=1}^{k} \frac{\sin(\theta_{\mathbf{w}_i,\mathbf{v}_j})}{2\pi\|\mathbf{w}_i\|} < 0, \tag{68}$$

hence $H'_{i,i}$ is not a PSD matrix. We begin with letting $O^\top DO$ be the eigendecomposition of the symmetric matrix $\bar{\mathbf{w}}\bar{\mathbf{w}}^\top - \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}^\top$, where O is an orthonormal matrix and D is diagonal. We have that $\mathrm{diag}(D) = (1, -1, 0, \ldots, 0)$, as readily seen by taking the orthogonal eigenvectors $\bar{\mathbf{w}}, \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}$ which correspond to the eigenvalues $1, -1$ respectively, where the remaining $d-2$ vectors orthogonal to $\bar{\mathbf{w}}, \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}$ comprise the rest of the spectrum with all zero correponding eigenvalues. Taking expectation over a random vector $\hat{\mathbf{u}} = (u_1, \ldots, u_d)$ uniformly on the unit hypersphere we have

$$\begin{aligned}
\mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top h_1(\mathbf{w},\mathbf{v})\hat{\mathbf{u}}\right] &= \mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top \frac{\sin(\theta_{\mathbf{w},\mathbf{v}})\|\mathbf{v}\|}{2\pi\|\mathbf{w}\|}\left(I - \bar{\mathbf{w}}\bar{\mathbf{w}}^\top + \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}^\top\right)\hat{\mathbf{u}}\right] \\
&= \frac{\sin(\theta_{\mathbf{w},\mathbf{v}})\|\mathbf{v}\|}{2\pi\|\mathbf{w}\|} - \mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top\left(\bar{\mathbf{w}}\bar{\mathbf{w}}^\top - \bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}\bar{\mathbf{n}}_{\mathbf{v},\mathbf{w}}^\top\right)\hat{\mathbf{u}}\right] \\
&= \frac{\sin(\theta_{\mathbf{w},\mathbf{v}})\|\mathbf{v}\|}{2\pi\|\mathbf{w}\|} - \mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top O^\top DO\hat{\mathbf{u}}\right] \\
&= \frac{\sin(\theta_{\mathbf{w},\mathbf{v}})\|\mathbf{v}\|}{2\pi\|\mathbf{w}\|} - \mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top D\hat{\mathbf{u}}\right] \qquad (69) \\
&= \frac{\sin(\theta_{\mathbf{w},\mathbf{v}})\|\mathbf{v}\|}{2\pi\|\mathbf{w}\|} - \mathbb{E}_{\hat{\mathbf{u}}}\left[u_1^2 - u_2^2\right] = \\
&= \frac{\sin(\theta_{\mathbf{w},\mathbf{v}})\|\mathbf{v}\|}{2\pi\|\mathbf{w}\|}, \qquad (70)
\end{aligned}$$

where equality (69) is due to a uniform distribution on the unit hypersphere being invariant to orthonormal transformations, and equality (70) is due to all coordinates of $\hat{\mathbf{u}}$ being i.i.d. From Eq. (70), the definition of $H'_{i,i}$, the linearity of expectation and the fact that $\|\mathbf{v}_i\| = 1$ for all $i \in [k]$, we have

$$\mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top H'_{i,i}\hat{\mathbf{u}}\right] = \sum_{j\in[n]\setminus\{i\}} \mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top h(\mathbf{w}_i,\mathbf{w}_j)\hat{\mathbf{u}}\right] - \sum_{j=1}^{k} \mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top h(\mathbf{w}_i,\mathbf{v}_j)\hat{\mathbf{u}}\right] = \lambda \tag{71}$$

for all $i \in [k]$. We will show this implies the existence of a particular vector $\mathbf{u}$ satisfying the above equality. Choose an arbitrary unit vector $\mathbf{u}'$. If $\mathbf{u}'$ satisfies the above equality we are done. Otherwise, assume w.l.o.g. that $\mathbf{u}'^\top H'_{i,i}\mathbf{u}' < \lambda$, in which case there must exist another unit vector $\mathbf{u}''$ such that $\mathbf{u}''^\top H'_{i,i}\mathbf{u}'' > \lambda$ (since otherwise $\mathbb{E}_{\hat{\mathbf{u}}}\left[\hat{\mathbf{u}}^\top H'_{i,i}\hat{\mathbf{u}}\right] < \lambda$, contradicting Eq. (71)). Let $\gamma(t) =$

$\frac{t\mathbf{u}'+(1-t)\mathbf{u}''}{\|t\mathbf{u}'+(1-t)\mathbf{u}''\|}$, then by the continuity of $\gamma(t)^\top H'_{i,i}\gamma(t)$ in t and the intermediate value theorem we have some t' satisfying $\gamma(t')^\top H'_{i,i}\gamma(t') = \lambda$, and by taking $\mathbf{u} = \gamma(t')$ we have $\mathbf{u}^\top H'_{i,i}\mathbf{u} = \lambda$.

Next, we show that $\lambda < 0$ under the assumptions in the theorem statement. To this end, it suffices to show that for some $i \in [n]$

$$\sum_{j\neq i}^{n} \sin(\theta_{\mathbf{w}_i,\mathbf{w}_j})\|\mathbf{w}_j\| - \sum_{j=1}^{k} \sin(\theta_{\mathbf{w}_i,\mathbf{v}_j}) < 0. \tag{72}$$

Beginning with the positive term, we have

$$\sum_{j\neq i}^{n} \sin(\theta_{\mathbf{w}_i,\mathbf{w}_j})\|\mathbf{w}_j\| \leq \sum_{j\neq i}^{n} \sin(\theta_{\mathbf{w}_i,\mathbf{w}_j})\frac{k\|\mathbf{w}_j\|}{\sum_{m=1}^{n}\|\mathbf{w}_m\|} \leq \sum_{j\neq i}^{n} \frac{k\|\mathbf{w}_j\|}{\sum_{m=1}^{n}\|\mathbf{w}_m\|}.$$

Since $\sum_{j=1}^{n} \frac{k\|\mathbf{w}_j\|}{\sum_{m=1}^{n}\|\mathbf{w}_m\|} = k$, there exists some $i \in [n]$ such that

$$\frac{k\|\mathbf{w}_i\|}{\sum_{m=1}^{n}\|\mathbf{w}_m\|} \geq \frac{k}{n} \geq 1,$$

thus the above equals

$$\sum_{j=1}^{n} \frac{k\|\mathbf{w}_j\|}{\sum_{m=1}^{n}\|\mathbf{w}_m\|} - \frac{k\|\mathbf{w}_i\|}{\sum_{m=1}^{n}\|\mathbf{w}_m\|} \leq k-1. \tag{73}$$

Turning to the negative term in Eq. (72), recall that $\mathbf{w}_i = (w_{i,1}, \ldots, w_{i,d})$ and $d \geq k$. Assume w.l.o.g. that $\mathbf{v}_j$ is a standard unit vector for all $j \in [k]$ (otherwise apply a change of basis under which the following argument is invariant), and compute

$$\sum_{j=1}^{k} \sin(\theta_{\mathbf{w}_i,\mathbf{v}_j}) = \sum_{j=1}^{k} \sqrt{1 - \frac{\langle \mathbf{w}_i, \mathbf{v}_j\rangle^2}{\|\mathbf{w}_i\|^2}} = \sum_{j=1}^{k} \sqrt{1 - \frac{w_{i,j}^2}{\|\mathbf{w}_i\|^2}} \geq \sum_{j=1}^{k} \left(1 - \frac{w_{i,j}^2}{\|\mathbf{w}_i\|^2}\right) \geq k - \sum_{j=1}^{d} \frac{w_{i,j}^2}{\|\mathbf{w}_i\|^2} = k-1.$$

Observe that if Eq. (73) is not a strict inequality, then it must hold that $n = k$ and $\|\mathbf{w}_j\| = 1$ for all $j \in [n]$. In such case, since $\mathbf{w}_1^n$ is not global, we have from Thm. 2 that it is not a permutation of the standard basis, therefore there must exist $\mathbf{w}_i$ of unit norm which is non-zero in at least two coordinates. For this $\mathbf{w}_i$, the above must be a strict inequality since $\sqrt{x} > x$ for any $x \in (0, 1)$, which guarantees a strict inequality for at least two summands. Now, combining the above with Eq. (73) and plugging in Eq. (72) establishes Eq. (68). Next, we invoke Lemma 29 with what was shown in Eq. (68).

To conclude the proof of the theorem, it only remains to show that $\mathbf{w}_1^n$ cannot be non-differentiable (which also implies that $\mathbf{w}_1^n(\alpha, i)$ is not a local minimum for $\alpha \in \{0, 1\}$). Assume $\mathbf{w}_1^n$ is non-differentiable, then by Lemma 1, there exists some $i \in [n]$ such that $\mathbf{w}_i = \mathbf{0}$.

First assume that $\mathbf{w}_1^n$ is not the origin. If we remove all zero vector neurons to obtain a differentiable point $\mathbf{w}_1^{n'} \in \mathbb{R}^{n'd}$ for $n' < n$, then we reduce to the previous case and there exists j such that $\mathbf{w}_1^{n'}(0, j)$ is a saddle point, and clearly adding more zero vector neurons (and permuting the neurons accordingly) till we recover $\mathbf{w}_1^n$, we have that it cannot be a local minimum, contradicting the theorem assumption.

Finally, if $\mathbf{w}_1^n$ is the origin then from Lemma 26, $\mathbf{w}_1^n$ is not a local minimum.

■

Proof [Proof of Proposition 15] Given a point $\mathbf{w}_1^n$, let $\mathbf{w}_1^n(t) = (t\mathbf{w}_1, \dots, t\mathbf{w}_n)$. From Eq. (66), we have that the objective is quadratic in t and therefore in particular, for a point to be minimal in our original nd-dimensional space, it must be minimal over t. Optimizing over t we have that the optimum t^* is given by

$$t^* = \frac{\sum_{i\in[n],j\in[k]} f\left(\mathbf{w}_i, \mathbf{v}_j\right)}{\sum_{i,j=1}^{n} f\left(\mathbf{w}_i, \mathbf{w}_j\right)},$$

therefore any local minimum must be of the form $\mathbf{w}_1^n(t^*)$, in which case its sum of Euclidean norms is given using Eq. (65) by

$$\sum_{i=1}^{n} \|t^*\mathbf{w}_i\| = t^* \sum_{i=1}^{n} \|\mathbf{w}_i\| = \frac{\sum_{i\in[n],j\in[k]} \|\mathbf{w}_i\|\|\mathbf{v}_j\| \left(\sin\left(\theta_{\mathbf{w}_i,\mathbf{v}_j}\right) + \left(\pi - \theta_{\mathbf{w}_i,\mathbf{v}_j}\right)\cos\left(\theta_{\mathbf{w}_i,\mathbf{v}_j}\right)\right)}{\sum_{i,j=1}^{n} \|\mathbf{w}_i\|\|\mathbf{w}_j\| \left(\sin\left(\theta_{\mathbf{w}_i,\mathbf{w}_j}\right) + \left(\pi - \theta_{\mathbf{w}_i,\mathbf{w}_j}\right)\cos\left(\theta_{\mathbf{w}_i,\mathbf{w}_j}\right)\right)} \sum_{i=1}^{n} \|\mathbf{w}_i\|.$$

Elementary calculus reveals that in $[0, \pi]$, the function $x \mapsto \sin(x)+(\pi-x)\cos(x)$ is monotonically decreasing, thus its image is bounded in the same interval, and the above displayed equation is upper bounded by

$$\frac{\pi \sum_{i\in[n],j\in[k]} \|\mathbf{w}_i\|\|\mathbf{v}_j\|}{\sum_{i=1}^{n} \pi\|\mathbf{w}_i\|^2 + \sum_{i\neq j}^{n} \|\mathbf{w}_i\|\|\mathbf{w}_j\| \left(\sin\left(\theta_{\mathbf{w}_i,\mathbf{w}_j}\right) + \left(\pi - \theta_{\mathbf{w}_i,\mathbf{w}_j}\right)\cos\left(\theta_{\mathbf{w}_i,\mathbf{w}_j}\right)\right)} \sum_{i=1}^{n} \|\mathbf{w}_i\|,$$

which in turn is at most

$$\frac{\pi \sum_{i\in[n],j\in[k]} \|\mathbf{w}_i\|\|\mathbf{v}_j\|}{\pi \sum_{i=1}^{n} \|\mathbf{w}_i\|^2} \sum_{i=1}^{n} \|\mathbf{w}_i\| = \frac{k \sum_{i\in[n]} \|\mathbf{w}_i\|}{\sum_{i=1}^{n} \|\mathbf{w}_i\|^2} \sum_{i=1}^{n} \|\mathbf{w}_i\| = k \frac{\left(\sum_{i=1}^{n} \|\mathbf{w}_i\|\right)^2}{\sum_{i=1}^{n} \|\mathbf{w}_i\|^2}. \tag{74}$$

Letting $\mathbf{u} = (\|\mathbf{w}_1\|, \dots, \|\mathbf{w}_n\|) \in \mathbb{R}^n$ and $\mathbf{u}' = (1, \dots, 1) \in \mathbb{R}^n$, we have from CS

$$\left(\sum_{i=1}^{n} \|\mathbf{w}_i\|\right)^2 = \langle \mathbf{u}, \mathbf{u}' \rangle^2 \leq \|\mathbf{u}\|^2 \|\mathbf{u}'\|^2 = n \sum_{i=1}^{n} \|\mathbf{w}_i\|^2,$$

thus by plugging the above we have that Eq. (74) is upper bounded by kn. ■

Proof [Proof of Thm. 16] Simply invoke Lemma 29 with each $i \in [n]$ satisfying the assumption in the theorem statement to obtain the result. ■