When Both Layers Learn: Training Dynamics of Representing Linear Models via ReLU Networks

Berk Tinaz TINAZ@USC.EDU
Changzhi Xie CHANGZHI@USC.EDU
Mahdi Soltanolkotabi SOLTANOL@USC.EDU
University of Southern California, Los Angeles, CA.

Editors: Steve Hanneke and Tor Lattimore

Abstract

In this paper, we study the gradient descent dynamics for jointly training *both* layers of a one-hidden-layer ReLU network to fit a linear target function. Concretely, we consider a realizable setting where inputs are drawn i.i.d. from a Gaussian distribution and labels follow a planted linear model. This stylized framework captures salient features of end-to-end training in inverse problems and certain auto-encoder models. Despite its apparent simplicity, the dynamics remain poorly understood, in part because the loss landscape contains multiple non-strict saddle points, making it unclear why gradient descent from random initialization reliably escapes bad stationary regions. We provide a detailed characterization of the optimization landscape and prove that gradient descent from a moderately small random initialization-*simultaneously training both layers*-converges to a global minimizer at a linear rate with order-wise optimal sample complexity. Our analysis tracks the trajectory through three phases: an *alignment phase* in which hidden weights progressively align with the planted direction while the output weights maintain the correct sign pattern; a *growth phase* in which the norms of both layers increase while preserving alignment; and a *local refinement phase* in which the aligned neurons rapidly converge to the planted direction, yielding fast local convergence. To rigorously show that GD avoids non-strict saddles, we develop trajectory-level control arguments for the end-to-end dynamics. In addition, we establish novel uniform concentration results that hold along the entire trajectory, and are essential for obtaining order-wise optimal sample complexity. We corroborate our theory with extensive experiments across a range of configurations.

Keywords: ReLU networks, gradient descent, learning theory, linear functions

1. Introduction

1.1. Motivation

End-to-end training of neural networks (NNs) via Gradient Descent (GD) has recently achieved remarkable success on many tasks. Of particular interest, these models have been adopted to solve inverse problems by taking the measurements as input and mapping them directly to the desired signal with successful scientific applications in computer vision (Ledig et al., 2017; Wang et al., 2018), MRI reconstruction (Sriram et al., 2020; Fabian et al., 2022), sparse-view computed tomography (CT) (Jin et al., 2017b), and phase retrieval (Hand et al., 2018). These models not only fit the training data but also appear to capture useful features and nuanced priors that enable them to generalize to unseen test examples. Despite this empirical success, the reasons behind the success of NNs for end-to-end training and how they can extract useful features from data remain unclear.

Perhaps the most classical form of end-to-end training is that arising in autoencoder type problems, where the goal is to teach a neural network to learn a linear mapping (e.g., identity for autoencoders). Surprisingly, the dynamics of training such a model are not well understood for nonlinear models. For linear networks, a classical result by Baldi and Hornik (1989) provided a complete characterization, showing how gradient descent recovers the principal components of the data. In contrast, understanding the dynamics of non-linear encoders has remained an open and challenging problem, even for simple target functions. In this paper, we aim to take a step towards a systematic understanding of the training dynamics of such problems by addressing the following question:

> *How do the dynamics of training ReLU neural networks with gradient descent starting from random initialization facilitate learning simple priors and structures such as linear target functions?*

Understanding this question requires reasoning not only about the final solution reached by GD, but about the entire trajectory of the optimization process. Recent empirical work suggests that several phenomena observed during neural network training, including grokking (or delayed generalization) (Power et al., 2022), are closely tied to the temporal evolution of gradient descent. In such settings, models may fit the training data well before exhibiting improved generalization, indicating that learning can unfold through distinct stages over the course of optimization. This perspective motivates a careful, trajectory-level analysis even in simple problem settings.

Despite significant recent progress in understanding neural networks (especially shallow networks) (Chizat et al., 2019; Soltanolkotabi et al., 2018; Jacot et al., 2018; Du et al., 2018; Ongie et al., 2019)[1], many aspects of the dynamics of GD and how it facilitates learning remain mysterious even in seemingly simple settings. A particularly simple one involves learning linear target functions via GD, that is, teaching a one-hidden-layer network to mimic the output of a simple linear model. Surprisingly, understanding the dynamics of GD in this simple setting has remained elusive. Although there are many results on learning specific target functions such as ReLUs (Xu and Du, 2023; Soltanolkotabi, 2017) and polynomials (Damian et al., 2022), these results typically exclude linear function classes. In fact, many of the existing papers use a pre-processing step or alter the early optimization trajectory to avoid complications arising from the dynamics of learning linear functions or genuinely training both layers (Damian et al., 2022). This is in part due to the fact that the optimization landscape of learning linear target functions contains multiple non-strict saddle points (i.e. where the gradient vanishes and the Hessian is PSD but has a 0 eigenvalue) requiring a subtle trajectory analysis to ensure GD avoid these bad points (See Section 2 for further details). We note that despite the simple formulation, quite a few interesting scenarios, including autoencoder training dynamics, are captured in this framework.

Our main contributions are as follows:

- We present one of the first works that analyzes training dynamics of learning *both* layers in a one-hidden-layer ReLU network in a practical regime. That is, we do not use pre-processing or alter the early optimization trajectory to avoid complications that arise from non-linear training dynamics of optimizing both layers.

1. Due to space constraints, we refer the reader to Appendix F for in-depth discussion on related work.

- We develop a theory for running GD on the NN with moderately small initialization, demonstrating exact convergence to the ground truth at a linear rate and with an optimal sample complexity that scales linearly in the number of parameters. That is, we show that the inner weights of the NN recover the target directions *exactly*, while the outer layer maintains the correct sign pattern.

- As detailed further in Section 2 the training landscape studied in this paper contains multiple non-strict saddles. To prove that the trajectory of GD from moderately small random initialization avoids these bad stationary points, we develop new techniques to control the GD trajectory which we combine with intricate uniform concentration bounds. In particular, our refined analysis tracks the trajectory through three phases (*alignment*, *growth*, and *local refinement* phases). We believe our refined trajectory analysis may have broader implications for the analysis of non-convex optimization problems involving non-strict saddles.

- Since gradient descent repeatedly reuses the same finite dataset across all phases, the iterates become statistically dependent on the samples. We address this by proving new uniform concentration bounds for the gradient along the entire optimization trajectory, holding simultaneously for all iterates encountered by GD. A key component of our uniform concentration result is that the accuracy of the concentration increases as we get closer and closer to the global optima. These refined bounds are a key technical ingredient for achieving order-optimal sample complexity.

- We further corroborate our results with various experimental investigations.

1.2. Problem Formulation

We first state the general family of problems of interest in this paper.

Data Model – We assume there are n pairs of training data consisting of input features $\boldsymbol{x}_i \in \mathbb{R}^d$ and corresponding targets $y_i \in \mathbb{R}$. As mentioned before, we consider the class of linear models where the relationship between $\boldsymbol{x}_i$ and y_i is given by the equation: $y_i = \boldsymbol{a}^T \boldsymbol{x}_i$ where $\boldsymbol{a} \in \mathbb{R}^d$ is the labeling vector. Conceptually, $\boldsymbol{a}$ is the target *direction* that our predictor should *learn*. For our theoretical analysis we assume the data points $\boldsymbol{x}_i$ are drawn i.i.d. according to a standard normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$.

Network Model – We consider one-hidden-layer neural networks of the form $f(\boldsymbol{v}, \boldsymbol{W}, \boldsymbol{x}) := \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x})$ as our predictor. Here k denotes the number of hidden-units, $\boldsymbol{v} \in \mathbb{R}^k$ is the outer layer of the neural network, $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ is the inner layer of the neural network, and $\phi(\boldsymbol{z})$ is the activation function. We refer to individual rows of $\boldsymbol{v}/\boldsymbol{W}$ as $v_i/\boldsymbol{w}_i$ respectively, In this paper, we specifically consider neural networks with ReLU activation functions i.e. $\phi(\boldsymbol{z}) = \text{ReLU}(\boldsymbol{z}) = \max(0, \boldsymbol{z})$, where $\max$ is applied to the input vector $\boldsymbol{z}$ element-wise. Furthermore, we focus on the exact parametrized setting, i.e. $k = 2$, as a step towards understanding the behavior of the over-specified/parameterized setting with $k > 2$ neurons.

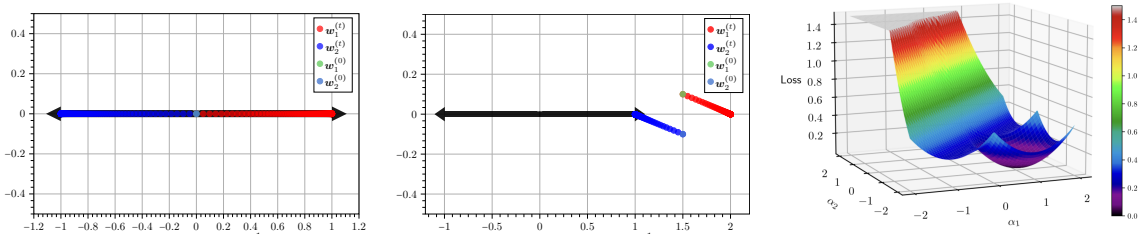
Training Loss – We minimize the squared loss between the target and the prediction

$$\widehat{\mathcal{L}}(\boldsymbol{v}, \boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i \right)^2 \tag{1}$$

using gradient descent. For part of our theoretical analysis of GD, we also consider the population loss (i.e. infinite data asymptotics as $n \to \infty$) with $\boldsymbol{x}$ drawn randomly from an isotropic Gaussian distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I_d})$. Concretely, the population loss is given by

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}} \left[\left(\sum_{i=1}^{k} v_i \phi\left(\boldsymbol{w}_i^T \boldsymbol{x}\right) - \boldsymbol{a}^T \boldsymbol{x} \right)^2 \right]. \tag{2}$$

2. Landscape Analysis: Why is learning linear functions with ReLUs challenging?



(a) Trajectory of convergence to the global optimum.

(b) Trajectory of convergence to one of the non-strict saddle points.

(c) 3D visualization of the loss landscape.

Figure 1: **GD trajectories and population loss landscape.** We run gradient descent on the population loss for a one-hidden-layer ReLU network with two hidden units and fixed output weights $\boldsymbol{v} = \left[1, -1\right]^T$. Panels (a) and (b) use two different initializations of $\boldsymbol{w}_1^{(0)}$ and $\boldsymbol{w}_2^{(0)}$. To visualize the dynamics in 2D, we plot each neuron in the plane spanned by $\boldsymbol{a}$ (black arrows indicate $\pm\boldsymbol{a}$) and a randomly chosen direction orthogonal to $\boldsymbol{a}$ (y-axis). In (a), a small initialization near the origin converges to the global optimum. In (b), initializing near $1.5\boldsymbol{a}$ leads GD to $(\boldsymbol{w}_1, \boldsymbol{w}_2) = (\boldsymbol{a}, 2\boldsymbol{a})$, a non-strict saddle of (2). Finally, (c) plots the population landscape in the reduced slice $\boldsymbol{w}_1 = \alpha_1 \boldsymbol{a}$, $\boldsymbol{w}_2 = \alpha_2 \boldsymbol{a}$.

Despite the simplicity of the target function, the gradient descent dynamics in this setting are surprisingly subtle. The difficulty is that the loss landscape is riddled with non-strict saddle points. Indeed, infinitely many of them—creating large flat directions where naive intuition about descent can fail. The next theorem makes this phenomenon precise for the population loss.

Theorem 1 (Landscape Characterization) *For $v_1, v_2 > 0$, the stationary points of the population loss (2) are either*

1. *global optima:* $v_1 \boldsymbol{w}_1 = \boldsymbol{a}, \quad v_2 \boldsymbol{w}_2 = -\boldsymbol{a}$,
2. *or non-strict saddles:* $v_1 \boldsymbol{w}_1 = (c+1)\,\boldsymbol{a}$, $v_2 \boldsymbol{w}_2 = c\boldsymbol{a}$, *where $c > 0$ or $c < -1$.*

We provide the proof of Theorem 1 in Appendix D.1.

Theorem 1 above shows that, beyond the global minima, the population loss contains a continuum of stationary points forming non-strict saddle manifolds parameterized by c. In particular, for every $c > 0$ and every $c < -1$, the equations $v_1\boldsymbol{w}_1 = (c+1)\boldsymbol{a}$ and $v_2\boldsymbol{w}_2 = c\boldsymbol{a}$ define a stationary point with flat directions in the loss. Thus the landscape is highly degenerate: instead of isolated critical points, there are infinitely many saddle regions that gradient descent can enter and move along without encountering negative curvature. This proliferation of flat saddles is the primary geometric obstruction to analyzing the global behavior of gradient descent.

In Figure 1, we illustrate how the initialization determines whether GD converges to a global optimum or drifts toward a non-strict saddle. Figure 1(a) shows a trajectory that converging to the global optimum, while Figure 1(b) shows a trajectory that stalls near a saddle. To further visualize the landscape, Figure 1(c) fixes $v_1 = v_2 = 1$ and plots the loss in the reduced two-dimensional slice $\boldsymbol{w}_1 = \alpha_1\boldsymbol{a}$, $\boldsymbol{w}_2 = \alpha_2\boldsymbol{a}$. In this slice, the gradient vanishes along the $\alpha_1 - \alpha_2 = 1$ *valley*, even though the loss remains strictly positive.

3. Main Result: Convergence of the Gradient Descent Trajectory

We now present our main result, which characterizes the training dynamics when *both* layers of a ReLU network are trained in the practical empirical regime.

Theorem 2 (Convergence of GD Trajectory) *Suppose we have n feature vectors $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ that are sampled i.i.d. according to a Gaussian distribution $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. We assume the corresponding outputs are generated according to a linear target function of the form $y_i = \boldsymbol{a}^T\boldsymbol{x}_i$, where $\boldsymbol{a} \in \mathbb{R}^d$ is an arbitrary weight vector. To learn this linear function, we fit a one-hidden-layer ReLU network with two hidden nodes*

$$\boldsymbol{x} \mapsto \boldsymbol{v}^T ReLU(\boldsymbol{W}\boldsymbol{x}) = v_1 ReLU(\boldsymbol{w}_1^T\boldsymbol{x}) - v_2 ReLU(\boldsymbol{w}_2^T\boldsymbol{x}).$$

by minimizing the empirical loss

$$\widehat{\mathcal{L}}(\boldsymbol{v}, \boldsymbol{W}) = \frac{1}{2n}\sum_{i=1}^{n}\left(\boldsymbol{v}^T ReLU\left(\boldsymbol{W}\boldsymbol{x}_i\right) - \boldsymbol{a}^T\boldsymbol{x}_i\right)^2.$$

over $\boldsymbol{v} = \left[v_1, -v_2\right]^T$ and $\boldsymbol{W} = \left[\boldsymbol{w}_1, \boldsymbol{w}_2\right]^T \in \mathbb{R}^{2\times d}$ using gradient descent with step size $\mu := \frac{\bar{\mu}}{\|\boldsymbol{a}\|}$ with $\bar{\mu} \le \frac{\mu_0}{\ln\left(\frac{\sqrt{\|\boldsymbol{a}\|}}{\sigma}\right)}$:

$$\boldsymbol{W}^{(\tau+1)} = \boldsymbol{W}^{(\tau)} - \mu\nabla_{\boldsymbol{W}}\widehat{\mathcal{L}}(\boldsymbol{v}^{(\tau)}, \boldsymbol{W}^{(\tau)}), \quad \boldsymbol{v}^{(\tau+1)} = \boldsymbol{v}^{(\tau)} - \mu\nabla_{\boldsymbol{v}}\widehat{\mathcal{L}}(\boldsymbol{v}^{(\tau)}, \boldsymbol{W}^{(\tau)})$$

Assume the initialization

$$\boldsymbol{w}_1^{(0)}, \boldsymbol{w}_2^{(0)} \sim \mathcal{N}\Big(0, \tfrac{\sigma^2}{d}\boldsymbol{I}_d\Big), \qquad v_1^{(0)}, v_2^{(0)} \sim \frac{\sigma}{\sqrt{d}}\xi, \quad \xi^2 \sim \chi_d^2,$$

with $\sigma \le \sigma_0\sqrt{\|\boldsymbol{a}\|}$, χ_d^2 a chi-squared distribution with d degrees of freedom, and define $\boldsymbol{W}^* = [\boldsymbol{a}, -\boldsymbol{a}]^T$. As long as the number of training samples satisfies $n \ge Cd$, then with probability at

least $1 - Ce^{-cd}$ *there exists* $T \geq c' \frac{1}{\bar{\mu}} \ln\left(\frac{\sqrt{\|\boldsymbol{a}\|}}{\sigma}\right)$ *such that for all iterations* $\tau > T$,

$$\left\| diag\left(\begin{bmatrix} v_1^{(\tau)} \\ v_2^{(\tau)} \end{bmatrix}\right) \boldsymbol{W}^{(\tau)} - \boldsymbol{W}^* \right\|_F^2 \leq \tilde{c}\,(1 - c\bar{\mu})^{(\tau - T)} \left\| diag\left(\begin{bmatrix} v_1^{(T)} \\ v_2^{(T)} \end{bmatrix}\right) \boldsymbol{W}^{(T)} - \boldsymbol{W}^* \right\|_F^2.$$

Here, $\mu_0, \sigma_0, c, \tilde{c}, c', C$ *are fixed numerical constants independent of any problem dimensions.*

This theorem shows that gradient descent can provably train a fully end-to-end one-hidden-layer ReLU network to learn a linear target from finitely many samples, despite the highly degenerate and saddle-rich optimization landscape. In particular, the result gives a global convergence guarantee for simultaneous optimization of the hidden and output weights; going beyond analyses that rely on effectively fixed features or only local perturbations around initialization. Starting from a small random initialization with the standard $\sigma/\sqrt{d}$ scaling—consistent with common "default" initializations used in practice—the two student neurons $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ rapidly align with the ground-truth direction $\boldsymbol{a}$, after which the *effective parameters* $\text{diag}\left(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right) \boldsymbol{W}$ converge geometrically to the planted solution. This linear-rate convergence kicks in after only a short burn-in period of $T \geq c' \frac{1}{\bar{\mu}} \ln\left(\frac{\sqrt{\|\boldsymbol{a}\|}}{\sigma}\right)$ iterations and requires only $n \geq Cd$ samples, which is information-theoretically optimal.

We note that the *same* finite dataset is reused across all iterations of gradient descent. Controlling the resulting dependence between the iterates and the samples requires uniform concentration arguments that hold *along the entire trajectory*, i.e., controlling population–empirical deviations simultaneously for all iterates visited by GD rather than at a fixed parameter value. Finally, the χ^2-based initialization for the output weights is used for technical convenience. One can alternatively initialize v_1 and v_2 as Gaussians with variance $\sigma^2/2$; the same qualitative convergence behavior persists, but the success probability degrades to a fixed constant, rather than $1 - Ce^{-cd}$. This constant failure with Gaussian initialization is unavoidable for $k = 2$. For large k, scaling the initialization with $\frac{\sigma^2}{k}$ yields failure probability decaying as e^{-k}. We therefore use a slight non-Gaussian modification in the initialization to demonstrate that except for this $k = 2$ artifact our result holds with much higher probability.

4. Experiments

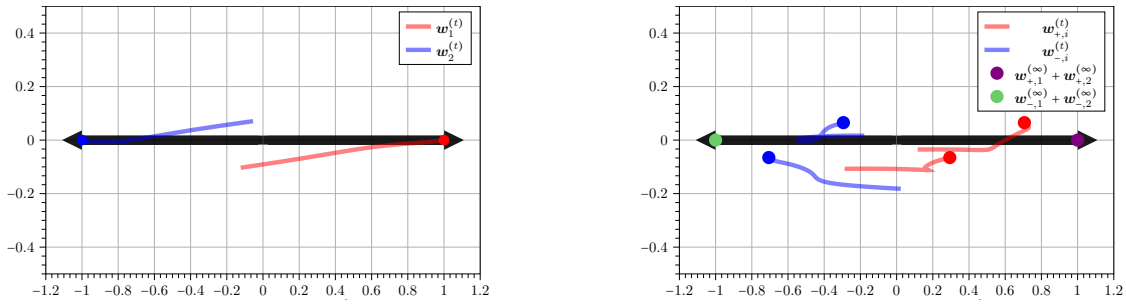
We run experiments on various output dimension (denoted with $r = 1$ vs. $r > 1$), and initialization scale (small vs large). In this section we show experimental results for single output $r = 1$ case and refer the reader to Appendix E.1 for multi-output $r > 1$ results. We use PyTorch for experiments and unless mentioned otherwise, network weights are initialized with Xavier Normal initialization (for a matrix $\boldsymbol{W} \in \mathbb{R}^{r \times d}$,$\boldsymbol{W}_{ij} \sim \mathcal{N}\left(0, \frac{2}{r+d}\right)$).

In order to change the initialization scale, we multiply the default initialization with a positive scalar σ. For *small* initialization experiments, we use $\sigma = 10^{-8}$, otherwise it is set to $\sigma = 1$. We set $d = 100$ and $\mu = 0.1$. All experiments are run on a server with an Intel Xeon Gold 5220R CPU. We would like to stress that even though the visualizations in this paper are based on a single trial, we ran these experiments for different random seeds and the behavior of the visualizations did not change.

In experiments w.l.o.g. we choose $\boldsymbol{a} = \boldsymbol{e}_1$ where $\boldsymbol{e}_1$ is the first standard basis in $\mathbb{R}^d$. This does not effect the results due to the rotational symmetry of isotropic Gaussian distribution of which $\boldsymbol{x}$ are drawn from. Note that this implies $\|\boldsymbol{a}\| = 1$ in our experiments. Finally, in this section we focus our experiments on the population loss. Similar results continue to hold in the empirical case with moderate sample sizes i.e. when $n \geq crd$ with c a sufficiently large constant.

When the model is exactly parameterized with two hidden nodes ($k = 2$), we empirically see that the model cannot converge to the global optima consistently. When it does, $\boldsymbol{v}^{(\infty)}$ indeed becomes ± 1 and $\boldsymbol{w}_1^{(\infty)}$ and $\boldsymbol{w}_2^{(\infty)}$ recover $\pm\boldsymbol{a}$ exactly. For the remaining time, the GD iterates converge to one of the many stationary points of this problem similar to the depiction in Figure 1 (part b). We further observe that iterates get stuck only when $\boldsymbol{v}_1^{(0)}$ and $\boldsymbol{v}_2^{(0)}$ both have the same signs which happens with probability $\frac{1}{2}$. This is also the reason for why we fix the correct sign pattern at initialization in Theorem 2.

When $k > 2$, the probability of *all* $\boldsymbol{v}_i$'s having the same sign decreases rapidly. Therefore, iterates typically converge to the global optima. However, in this case global minima is not unique anymore. To demonstrate this, consider the case where there are four hidden units ($k = 4$) instead of two. The trajectory of the inner weights across GD iterations is depicted in Figure 2.



(a) Trajectory of neurons when $k = 2$.

(b) Trajectory of neurons when $k > 2$.

Figure 2: **Trajectory of neurons for different values of** k. We run gradient descent updates on the population loss. A randomly selected orthogonal direction to $\boldsymbol{a}$ is shown for the y-axis in order to visualize the neurons in 2D. Black arrows indicate $\pm\boldsymbol{a}$ direction. We use colors red and blue to indicate whether $\boldsymbol{v}_i$ corresponding to $\boldsymbol{w}_i$ is positive or negative respectively. Points at the end of each trajectory denotes the final weight GD converges to.

We observe that while no individual $\boldsymbol{w}_i$ align itself with $\pm\boldsymbol{a}$ direction, grouping hidden units based on their corresponding signs in $\boldsymbol{v}$ and summing them recovers $\pm\boldsymbol{a}$ *exactly* (purple and green points in Figure 2). Although not depicted here, we have tried various values for $k > 2$ and the observation that grouping weights recover $\pm\boldsymbol{a}$ was consistent. This suggests that combining node aggregation technique from (Li et al., 2024) with our proof strategy may extend our results for the $k > 2$ setting. We leave this to future work.

5. Overview and Key Ideas of the Proof

In this section, we outline the main ideas underlying our analysis. As mentioned previously, a major challenge is that the optimization landscape is riddled with non-strict saddle points that gradient descent can get stuck in. Thus, our analysis requires a very refined control of the trajectory to

guarantee that the iterates escape these saddle regions. We will show that the trajectory of full-batch gradient descent partitions into three distinct phases discussed below. Figure 3 illustrates the three phases and their interaction.

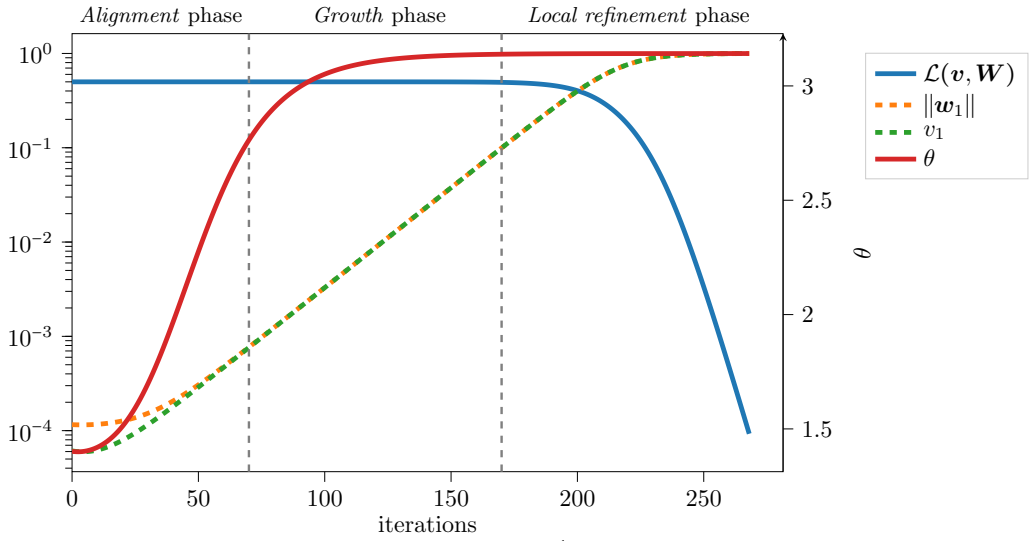


Figure 3: **Phases of GD Trajectory**. We run gradient descent updates on the population loss with small initialization $\sigma = 10^{-4}$. We track the population loss $\mathcal{L}$ (blue), norms $v_1, \|\boldsymbol{w}_1\|$ (green and yellow), and the θ – i.e. angle between $\boldsymbol{w}_1$, $\boldsymbol{w}_2$ – (red). For visualization purposes θ uses the right vertical axis. v_2 and $\|\boldsymbol{w}_2\|$ behave similarly but omitted for clarity.

(1) Alignment phase (Section 5.4). Starting from a small random initialization, we show that the hidden weights progressively align with the planted direction while the output weights maintain the correct sign pattern.

(2) Growth phase (Section 5.5). Once sufficient alignment has been established, we prove that the norms of both the hidden and output layers grow in a coordinated fashion while preserving this alignment. This phase drives the effective parameters toward the correct scale and pushes the iterates away from flat saddle regions of the loss landscape. A key technical challenge here is to show that gradient descent does not drift into spurious stationary points despite the non-strict nature of these saddles.

(3) Local refinement phase (Section 5.6). After the alignment and growth phases we enter a well-behaved region of the planted solution, where the dynamics become locally well-conditioned. In this phase, We show that the aligned neurons then converge rapidly to the ground-truth direction, and the effective parameters enjoy a linear rate of convergence to the global minimizer.

Throughout all three phases, the same finite dataset is reused across iterations. To control the resulting dependence between the iterates and the samples, we establish new trajectory-level uniform concentration bounds that hold simultaneously for all points visited by gradient descent. These results are crucial for obtaining order-wise optimal sample complexity. We give an overview of these

uniform concentration results in Section 5.7. Before we detail the specific phases of the trajectory, we also need to establish two sets of key identities. The first set demonstrates a specific property of balancedness between the inner and outer weights (Section 5.2). The second set concerns the stability of our training dynamics, ensuring that the evolution is monotonic in the sense that once the iterates enter a new phase, they do not revert to a previous one (Section 5.3). We begin with some quick notation used throughout our proofs.

5.1. Notation

In this section we gather some simple notation used in our proofs. As a reminder we use $\widehat{\mathcal{L}}$ and $\mathcal{L}$ to denote the empirical and population losses, respectively. We use

$$\Delta\mathcal{G}_1 := \frac{2}{v_1}\left(\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L}\right) \quad \text{and} \quad \Delta\mathcal{G}_2 := \frac{2}{v_2}\left(\nabla_{\boldsymbol{w}_2}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_2}\mathcal{L}\right)$$

to denote the scaled difference between the empirical and population gradients. Finally, we use θ to denote the angle between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. We also define θ_1 to be the angle between $\boldsymbol{w}_1$ and $\boldsymbol{a}$, θ_2 to be the angle between $\boldsymbol{w}_2$ and $-\boldsymbol{a}$. We note that all lemmas stated in this proof overview our under the assumptions of the main theorem, we avoid repeating these assumptions repeatedly for readability.

5.2. Controlling the imbalance term

A crucial identity used throughout our proofs is that from moderately small initialization the norms of the inner and outer weights remain close to each other. Concretely, we define the imbalance term as $b_1^{(\tau)} := \left\|\boldsymbol{w}_1^{(\tau)}\right\|^2 - \left(v_1^{(\tau)}\right)^2$ and $b_2^{(\tau)} := \left\|\boldsymbol{w}_2^{(\tau)}\right\|^2 - \left(v_2^{(\tau)}\right)^2$. A constant bound for the absolute value of these terms is required to prove that the norms remain bounded throughout the training process (see Lemma 5).

While the imbalance is invariant in gradient flow (Ji and Telgarsky, 2019), the discretization in gradient descent introduces a small drift given by:

$$b_1^{(\tau+1)} = b_1^{(\tau)} + \mu^2\left(\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 - \left(\nabla_{v_1}\widehat{\mathcal{L}}\right)^2\right).$$

A simple constant bound on this drift is insufficient for our analysis, as the errors could accumulate to infinity over an infinite number of iterations.

To address this, we prove a stronger result: the drift in each step is bounded by the distance between the effective weights $\operatorname{diag}\left(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right)\boldsymbol{W}$ and the planted solution. Since $\operatorname{diag}\left(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right)\boldsymbol{W}$ converges to the planted solution exponentially fast in Phase 3, the total accumulated drift remains finite even as $\tau \to \infty$. Concretely, we prove the following lemma.

Lemma 3 (Imbalance bound) *Assume that $v_1^{(\tau)}, v_2^{(\tau)} > 0$. For any $i \in \{1, 2\}$, we have*

$$\left|b_i^{(\tau+1)} - b_i^{(\tau)}\right| \le c_6\mu^2\left(\left(v_i^{(\tau)}\right)^2 + \left\|\boldsymbol{w}_i^{(\tau)}\right\|^2\right)\left(\left\|v_1^{(\tau)}\boldsymbol{w}_1^{(\tau)} - \boldsymbol{a}\right\|^2 + \left\|v_2^{(\tau)}\boldsymbol{w}_2^{(\tau)} + \boldsymbol{a}\right\|^2\right).$$

Here, we set the constant as $c_6 = 6$.

This lemma is proven in Section C.1.

5.3. Stability of Training Dynamics

In this section, we establish two key stability properties that serve as the foundation for our proof. These results ensuring that the evolution is monotonic in the sense that once the iterates enter a new phase, they do not revert to a previous phase. The first lemma ensures that once the angle becomes small (at the end of the first phase) it continues to remain sufficiently small.

Lemma 4 (Angle stays small) *Assume that* $\mu \leq \frac{c_0}{\|\boldsymbol{a}\|}$. *For any iteration* τ *such that* $0 < v_1^{(\tau)}, v_2^{(\tau)} \leq c_2\sqrt{\|\boldsymbol{a}\|}$, $\theta_1^{(\tau)}, \theta_2^{(\tau)} \leq c_4$, $\left|b_1^{(\tau)}\right|, \left|b_2^{(\tau)}\right| \leq \gamma \|\boldsymbol{a}\|$ *and* $\left\|\Delta\mathcal{G}_1^{(\tau)}\right\|, \left\|\Delta\mathcal{G}_2^{(\tau)}\right\| \leq c_5 \|\boldsymbol{a}\|$, *we have*

$$\theta_1^{(\tau+1)}, \theta_2^{(\tau+1)} \leq c_4.$$

Here, we set the constants as $c_0 \leq \frac{1}{2}, c_2 = 2, c_5 = \frac{1}{50}, c_4 = \frac{\pi}{20}, \gamma = \frac{1}{4}$.

This lemma is proven in Section C.2. The second result ensures the norms remain bounded.

Lemma 5 (Norms remain bounded) *Assume that* $\mu \leq \frac{c_0}{\|\boldsymbol{a}\|}$. *For any* $0 \leq \beta \leq \frac{1}{4}$ *and iteration* τ *such that* $\beta\sqrt{\|\boldsymbol{a}\|} < v_1^{(\tau)}, v_2^{(\tau)} \leq c_2\sqrt{\|\boldsymbol{a}\|}$, $\theta_1^{(\tau)}, \theta_2^{(\tau)} \leq c_4$, $\left|b_1^{(\tau)}\right|, \left|b_2^{(\tau)}\right| \leq \gamma \|\boldsymbol{a}\|$ *and* $\left\|\Delta\mathcal{G}_1^{(\tau)}\right\|, \left\|\Delta\mathcal{G}_2^{(\tau)}\right\| \leq c_5 \|\boldsymbol{a}\|$, *we have*

$$\beta\sqrt{\|\boldsymbol{a}\|} < v_1^{(\tau+1)}, v_2^{(\tau+1)} \leq c_2\sqrt{\|\boldsymbol{a}\|}.$$

Here, we set the constants as $c_0 \leq \frac{4}{25}, c_2 = 2, c_5 = \frac{1}{50}, c_4 = \frac{\pi}{20}, \gamma = \frac{1}{4}$.

This lemma is proven in Section C.3.

5.4. Overview of *Alignment* Phase

The primary objective of Phase 1 is to demonstrate that $\boldsymbol{w}_1$ becomes approximately aligned with $\boldsymbol{a}$ (and $\boldsymbol{w}_2$ with $-\boldsymbol{a}$) within a constant number of steps, which is crucial for Phases 2 and 3. By symmetry, we focus on $\boldsymbol{w}_1$.

Our key observation is that the gradient update is dominated by the signal direction. Specifically, the update can be decomposed as:

$$\boldsymbol{w}_1^{(\tau+1)} = \boldsymbol{w}_1^{(\tau)} + \mu\left(\frac{v_1}{2}\boldsymbol{a} + \boldsymbol{\zeta}^{(\tau)}\right).$$

where the remainder term $\boldsymbol{\zeta}^{(\tau)}$ consists of terms involving the weights and the empirical noise. Since we use small initialization, these weight-dependent terms are much smaller than the signal term. This implies that the projection of $\boldsymbol{w}_1$ onto the signal direction $\boldsymbol{a}$ grows much faster than its projection onto the orthogonal subspace. Specifically, we have the following lemma proven in Section C.4:

Lemma 6 (Angle alignment) *Assume that* $\mu \leq \frac{c_0}{\|\boldsymbol{a}\|}$, $\sigma \leq \sigma_0\sqrt{\|\boldsymbol{a}\|}$. *After* $T_1 = \lceil \frac{c_9}{\mu\|\boldsymbol{a}\|} \rceil$ *iterations, it holds that*

$$\theta_1^{(T_1)}, \theta_2^{(T_1)} \leq c_4$$

with probability at least $1 - Ce^{-cd}$. *Moreover, this alignment is achieved while maintaining that:*

$$c_3\sigma \leq v_1^{(T_1)}, v_2^{(T_1)} \leq c_2\sqrt{\|\boldsymbol{a}\|}, \quad \left|b_1^{(T_1)}\right|, \left|b_2^{(T_1)}\right| \leq c_{10}\|\boldsymbol{a}\|.$$

Here, we set the constants as $c_0 \leq 1, c_4 = \frac{\pi}{20}, c_9 = \frac{64}{\tan c_4}, c_2 = 2, c_3 = \frac{1}{4}$. *With* $\alpha = \frac{65}{\tan c_4}$, *we have* $c_{10} = \frac{1}{4e^{2\alpha}}, \sigma_0 = \frac{1}{8e^{2\alpha}}$.

5.5. Overview of *Growth* Phase

In Phase 3, we show that the effective weights $\operatorname{diag}\left(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right)\boldsymbol{W}$ converge to the planted solution at an exponential rate. A key ingredient is a Polyak–Lojasiewicz (PL) inequality for the population loss (Lemma 8), which lower-bounds the squared gradient norm in terms of the suboptimality gap. Importantly, the PL constant depends on the magnitudes of v_1 and v_2.

At the end of Phase 1, $|v_1|$ and $|v_2|$ remain at their initialization scale, so the PL inequality only yields a weak contraction and therefore a slow convergence rate. The main goal of Phase 2 is to grow v_1 and v_2 to a sufficiently large scale, thereby strengthening the PL constant and enabling fast linear convergence in Phase 3.

Lemma 7 (Norm growth) *Assume that* $\mu \leq \frac{c_0}{\|\boldsymbol{a}\|\ln\left(\frac{\sqrt{\|\boldsymbol{a}\|}}{\sigma}\right)}$. *After* $T_2 = \lceil \frac{c_8}{\mu\|\boldsymbol{a}\|}\ln\left(\frac{\sqrt{\|\boldsymbol{a}\|}}{\sigma}\right)\rceil$ *iterations, we have*

$$v_1^{(T_1+T_2)}, v_2^{(T_1+T_2)} \geq c_7\sqrt{\|\boldsymbol{a}\|}, \quad \left|b_1^{(T_1+T_2)}\right|, \left|b_2^{(T_1+T_2)}\right| \leq c_{11}\|\boldsymbol{a}\|.$$

Here, we set the constants as $c_0 \leq \frac{1}{10^8}, c_7 = \frac{1}{4}, c_8 = 32, c_{11} = \frac{1}{50}$.

This lemma is proven in Section C.5.

5.6. Overview of the local *Refinement* Phase

In the final *local refinement* phase, we show that the effective weights $\operatorname{diag}\left(\begin{bmatrix} v_1^{(\tau)} \\ v_2^{(\tau)} \end{bmatrix}\right)\boldsymbol{W}^{(\tau)}$ converge to the planted solution $\boldsymbol{W}^* = [\boldsymbol{a}, -\boldsymbol{a}]^T$. Rather than tracking parameters directly, we first prove that along empirical gradient descent the *population* loss decreases rapidly, and then convert this decay into the stated parameter convergence. The full argument (Proof of Theorem 2 in Section D.2) is technical: it couples a population-level gradient-descent analysis with trajectory-uniform concentration bounds (next section). For clarity, we sketch only the population argument here. This population reduction is essential because the empirical loss is not smooth and as discussed below, even the population loss is not uniformly smooth. The proof proceeds in two parts.

Part 1 (PL inequality) In this step we will show the following PL inequality proven in Section B.2.

Lemma 8 (PL Inequality for the population loss) *For* $v_1, v_2 > 0$,

$$\|\nabla_{\boldsymbol{w}_1}\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|^2 + \|\nabla_{\boldsymbol{w}_2}\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})\|^2 \geq \alpha \min\left(v_1^2, v_2^2\right)\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})$$

holds with $\alpha = 0.05$ *as long as* $\theta > \frac{\pi}{2}$.

To prove this PL inequality we first show that it can be deduced by establishing the PL inequality when $v_1 = v_2 = 1$ via a clever reduction argument. To prove the latter we define

$$h\left(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{a}\right) = \left\|\nabla_{\boldsymbol{w}_1} \mathcal{L}\left(\boldsymbol{W}\right)\right\|^2 + \left\|\nabla_{\boldsymbol{w}_2} \mathcal{L}\left(\boldsymbol{W}\right)\right\|^2 - \alpha \mathcal{L}\left(\boldsymbol{W}\right). \tag{3}$$

Note that since we set $v_1 = v_2 = 1$ the loss is now only a function of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. Also note that to prove the PL inequality it sufficies to show that h is always positive. To do this, in our proof we show that $\tilde{h}\left(\boldsymbol{w}_1, \boldsymbol{w}_2\right) = \min_{\boldsymbol{a}} h\left(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{a}\right)$ is always positive. The way we establish this is by showing that $\frac{1}{\|\boldsymbol{w}_2\|^2}\tilde{h}\left(\boldsymbol{w}_1, \boldsymbol{w}_2\right)$ is only a function of θ (the angle between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$) and $\frac{\|\boldsymbol{w}_1\|}{\|\boldsymbol{w}_2\|}$. Since this is now only a function of two variables θ and $\frac{\|\boldsymbol{w}_1\|}{\|\boldsymbol{w}_2\|}$ it is easy to establish non-negativity as long as $\theta > \frac{\pi}{2}$. The latter holds at the end of the growth phase and continues to remain large utilizing the stability of the dynamics established in Section 5.3 (concretely, Lemma 4 shows the angles with planted directions remain small which implies the angle between the weight vectors remain large).

Part 2 (Gradient smoothness) – While the population loss is not smooth in the entire domain, we show that in the region of the local refinement phase it is indeed smooth. Leaving the *growth* phase, we have lower/upper bounds on $v_1, v_2, \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\|$. Additionally, due to the stability analysis in Section 5.3 we continue to have lower/upper bounds on these quantities. In the next lemma we show that assuming such lower/upper bounds the population loss is indeed smooth. This lemma is proven in Section B.3.

Lemma 9 (Smoothness of the population loss) *$\left\|\nabla^2 \mathcal{L}\left(\boldsymbol{v}, \boldsymbol{W}\right)\right\|_F \leq L\|\boldsymbol{a}\|$ holds for all $\boldsymbol{v} \in \mathbb{R}^2, \boldsymbol{W} \in \mathbb{R}^{2\times d}$ such that $c_1\sqrt{\|\boldsymbol{a}\|} \leq v_1, v_2, \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \leq c_2\sqrt{\|\boldsymbol{a}\|}$ holds. Here c_1, c_2, L are fixed constants.*

Showing geometric decrease of the *population* loss under a PL inequality and smoothness is a classical optimization result. Our setting is more delicate because we run *empirical* gradient descent rather than its population counterpart. See the proof of Theorem 2 in Section D.2 for how we combine the trajectory-uniform concentration bounds (next section) with the PL and smoothness properties of the population loss stated above to obtain a geometric decrease in the population loss.

5.7. Uniform Concentration

In this section, we provide an overview of the novel uniform concentration result that we have established which is key to our near optimal sample complexity. In particular, the concentration holds along the entire trajectory of GD and is used across the three phases. We provide the setup next. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$ in $\mathbb{R}^d$. For $\boldsymbol{w}, \boldsymbol{w}^* \in \mathbb{S}^{d-1}$ define

$$M(\boldsymbol{w}, \boldsymbol{w}^*) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}\rangle \geq 0\}} \mathbb{1}_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}^*\rangle \geq 0\}} \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{4}$$

We prove a high-probability bound on

$$\sup_{\boldsymbol{w}, \boldsymbol{w}^* \in \mathbb{S}^{d-1}} \left\|M(\boldsymbol{w}, \boldsymbol{w}^*) - \mathbb{E}\left[M(\boldsymbol{w}, \boldsymbol{w}^*)\right]\right\|.$$

Lemma 10 (Uniform Concentration) *Fix $\delta \in (0, 1/2)$. There exist universal constants $C, c > 0$ such that the following holds. If*

$$n \;\geq\; C\, d\, \frac{\log^2(1/\delta)}{\delta^2}, \tag{5}$$

then with probability at least $1 - 3e^{-cd}$,

$$\sup_{\boldsymbol{w}, \boldsymbol{w}^* \in \mathbb{S}^{d-1}} \|M(\boldsymbol{w}, \boldsymbol{w}^*) - \mathbb{E}\left[M(\boldsymbol{w}, \boldsymbol{w}^*)\right]\| \;\leq\; \delta.$$

This lemma is proven in Section C.6. Notably, this lemma allows us to establish separate high-probability bounds for the deviations of the gradient components with respect to $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. Concretely, it allows us to show the following lemma proven in Section C.7.

Lemma 11 (Component-wise Gradient Deviation Bounds) *Fix $\delta \in (0, 1/2)$. Define the error vectors $\mathbf{h}_1 = v_1\boldsymbol{w}_1 - \boldsymbol{a}$ and $\mathbf{h}_2 = v_2\boldsymbol{w}_2 + \boldsymbol{a}$. Under the sample complexity $n \;\geq\; C\, d\, \frac{\log^2(1/\delta)}{\delta^2}$, with probability at least $1 - 3e^{-cd}$, the following bounds hold simultaneously:*

$$\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L}\right\| \leq v_1\delta(\|\mathbf{h}_1\| + \|\mathbf{h}_2\|), \tag{6}$$

$$\left\|\nabla_{\boldsymbol{w}_2}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_2}\mathcal{L}\right\| \leq v_2\delta(\|\mathbf{h}_1\| + \|\mathbf{h}_2\|). \tag{7}$$

This lemma highlights a particularly favorable self-regularizing property of the dynamics: the empirical–population gradient deviation scales *linearly* with the current error $\|\boldsymbol{h}_1\| + \|\boldsymbol{h}_2\|$. As the iterates approach the global optimum and the errors shrink, the concentration bounds automatically tighten, yielding increasingly accurate gradient estimates along the trajectory. In other words, concentration improves precisely when it is most needed in the local refinement regime, enabling stable geometric convergence in this region.

Acknowledgements

This work was partially supported by AWS credits through an Amazon Faculty Research Award, a NAIRR Pilot Award, and generous funding by Coefficient Giving. M. Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, NSF CAREER Award #1846369, DARPA FastNICS program, NSF CIF Awards #1813877 and #2008443, and NIH Award DP2LM014564-01.

References

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *36th International Conference on Machine Learning, ICML 2019*, pages 477–502. International Machine Learning Society (IMLS), 2019.

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

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90014-2. URL https://www.sciencedirect.com/science/article/pii/0893608089900142.

Etienne Boursier and Nicolas Flammarion. Simplicity bias and optimization threshold in two-layer relu networks, 2025. URL https://arxiv.org/abs/2410.02348.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris, Ser. I*, 334(6):495–500, 2002.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs, 2017. URL https://arxiv.org/abs/1702.07966.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inf. Theory*, 61(4):1985–2007, 2015. ISSN 0018-9448.

Ji Chen, Dekai Liu, and Xiaodong Li. Nonconvex rectangular matrix completion via gradient descent without ℓ_2, ∞ regularization. *IEEE Trans. Inf. Theory*, 66(9):5806–5841, 2020. ISSN 0018-9448.

Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Commun. Pure Appl. Math.*, 70(5):822–883, 2017. ISSN 0010-3640; 1097-0312/e.

Dmitry Chistikov, Matthias Englert, and Ranko Lazic. Learning a neuron by a shallow relu network: Dynamics and implicit bias for correlated inputs, 2023. URL https://arxiv.org/abs/2306.06479.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

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

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in neural information processing systems*, 29, 2016.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

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

Zalan Fabian, Berk Tinaz, and Mahdi Soltanolkotabi. Humus-net: Hybrid unrolled multi-scale network architecture for accelerated mri reconstruction. *Advances in Neural Information Processing Systems*, 35:25306–25319, 2022.

Rong Ge, Furong Huang, Chi Jin, and Yang. Yuan. Escaping from saddle points: online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in Neural Information Processing Systems*, 29:2973–2981, 2016.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *arXiv preprint arXiv:2006.13409*, 2020.

Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks, 2019. URL https://arxiv.org/abs/1810.02032.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. page 1724–1732, 2017a.

Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE transactions on image processing*, 26(9): 4509–4522, 2017b.

Kevin Kögler, Alexander Shevchenko, Hamed Hassani, and Marco Mondelli. Compression of structured data with autoencoders: Provable benefit of nonlinearities and depth, 2024. URL https://arxiv.org/abs/2402.05013.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

Binghui Li, Zhixuan Pan, Kaifeng Lyu, and Jian Li. Feature averaging: An implicit bias of gradient descent leading to non-robustness in neural networks. *arXiv preprint arXiv:2410.10322*, 2024.

Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Appl. Comput. Harmon. Anal.*, 47(3):893–934, 2019. ISSN 1063-5203.

Shuyang Ling and Thomas Strohmer. Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing. *Inf. Inference*, 8(1):1–49, 2019. ISSN 2049-8764; 2049-8772/e.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found. Comput. Math.*, 20(3):451–632, 2020. ISSN 1615-3375; 1615-3383/e.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108(1 (A)):177–205, 2006. ISSN 0025-5610; 1436-4646/e.

Jorge Nocedal and Stephen J. Wright. Trust-region methods. *Numerical Optimization*, pages 66–100, 2006.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. pages 1674–1703, 2017.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in sgd learning of shallow neural networks, 2025. URL https://arxiv.org/abs/2504.19983.

Alexander Shevchenko, Kevin Kögler, Hamed Hassani, and Marco Mondelli. Fundamental limits of two-layer autoencoders, and achieving them with gradient methods, 2022.

Mahdi Soltanolkotabi. Learning relus via gradient descent. *Advances in neural information processing systems*, 30, 2017.

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, 2018.

Mahdi Soltanolkotabi, Dominik Stöger, and Changzhi Xie. Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5140–5142. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/soltanolkotabi23a.html.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated mri reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 64–73. Springer, 2020.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23831–23843. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c82836ed448c41094025b4a872c5341e-Paper.pdf.

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

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Found. Comput. Math.*, 18(5):1131–1198, 2018. ISSN 1615-3375; 1615-3383/e.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

Aad W. Van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, NY, 1996. ISBN 978-0-387-94640-5.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Learning a single neuron with bias using gradient descent, 2022. URL https://arxiv.org/abs/2106.01101.

Martin J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. ISBN 978-1-108-49802-9. doi: 10.1017/9781108627771. URL https://doi.org/10.1017/9781108627771. A non-asymptotic viewpoint.

Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. *Advances in neural information processing systems*, 31, 2018.

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect, 2022. URL https://arxiv.org/abs/2110.03677.

Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult, 2023. URL https://arxiv.org/abs/2310.17087.

Chenwei Wu, Jiajun Luo, and Jason D. Lee. No spurious local minima in a two hidden unit reLU network, 2018. URL https://openreview.net/forum?id=B14uJzW0b.

Weihang Xu and Simon S. Du. Over-parameterization exponentially slows down gradient descent for learning a single neuron, 2023.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods, 2022. URL https://arxiv.org/abs/2001.05205.

Richard Y. Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *J. Mach. Learn. Res.*, 20 (114):1–34, 2019. URL http://jmlr.org/papers/v20/19-020.html.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent, 2018. URL https://arxiv.org/abs/1806.07808.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1980–2022. PMLR, 07–10 Jul 2017. URL http://proceedings.mlr.press/v65/zhang17b.html.

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. Recovery guarantees for one-hidden-layer neural networks, 2017. URL https://arxiv.org/abs/1706.03175.

Zhenyu Zhu, Fanghui Liu, and Volkan Cevher. How gradient descent balances features: A dynamical analysis for two-layer neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=25j2ZEgwTj.

Appendix A. Useful Calculations

In this section we provide the derivation of several useful identities.

A.1. Population Loss

Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ be two arbitrary vectors. Define

$$\begin{aligned} f(\boldsymbol{a}, \boldsymbol{b}) &= \mathbb{E}_{\boldsymbol{x}}\left[\left[\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x}\right]_+\left[\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{x}\right]_+\right] \\ &\overset{(a)}{=} \frac{1}{2\pi}\|\boldsymbol{a}\|\|\boldsymbol{b}\|\left(\sin\left(\theta_{\boldsymbol{a},\boldsymbol{b}}\right)+\left(\pi-\theta_{\boldsymbol{a},\boldsymbol{b}}\right)\cos\left(\theta_{\boldsymbol{a},\boldsymbol{b}}\right)\right) \end{aligned} \tag{8}$$

where $\theta_{\boldsymbol{a},\boldsymbol{b}} = \cos^{-1}\left(\frac{\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}\right)$, expectation is over $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and inequality (a) follows from the Table 1 in (Daniely et al., 2016).

Using these we calculate the closed form for the population loss (2) as:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\left\|\sum_{i=1}^{k} v_i\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}\right)-\boldsymbol{a}^T\boldsymbol{x}\right\|^2\right] \\ &= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_j \mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}\right)\phi\left(\boldsymbol{w}_{\boldsymbol{j}}^{\boldsymbol{T}}\boldsymbol{x}\right)\right] - \sum_{i=1}^{k} v_i \boldsymbol{a}^T\mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}\right)\boldsymbol{x}\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{a}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{a}\right] \\ &= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_j \mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}\right)\phi\left(\boldsymbol{w}_{\boldsymbol{j}}^{\boldsymbol{T}}\boldsymbol{x}\right)\right] - \sum_{i=1}^{k} v_i \boldsymbol{a}^T\mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}\right)\boldsymbol{x}\right] + \frac{\|\boldsymbol{a}\|^2}{2} \\ &\overset{(a)}{=} \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_j f\left(\boldsymbol{w}_i, \boldsymbol{w}_j\right) - \sum_{i=1}^{k} v_i \boldsymbol{a}^T\mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}\right)\boldsymbol{x}\right] + \frac{\|\boldsymbol{a}\|^2}{2} \\ &\overset{(b)}{=} \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_j f\left(\boldsymbol{w}_i, \boldsymbol{w}_j\right) - \sum_{i=1}^{k} v_i \boldsymbol{a}^T\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}}\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}\right)\right] + \frac{\|\boldsymbol{a}\|^2}{2} \\ &= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_j f\left(\boldsymbol{w}_i, \boldsymbol{w}_j\right) - \sum_{i=1}^{k} v_i \boldsymbol{a}^T\boldsymbol{w}_i\mathbb{E}_{\boldsymbol{x}}\left[\phi'\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}\right)\right] + \frac{\|\boldsymbol{a}\|^2}{2} \\ &\overset{(c)}{=} \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_j f\left(\boldsymbol{w}_i, \boldsymbol{w}_j\right) - \frac{1}{2}\sum_{i=1}^{k} v_i \boldsymbol{a}^T\boldsymbol{w}_i + \frac{\|\boldsymbol{a}\|^2}{2} \\ &= \frac{1}{4\pi}\sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_j \|\boldsymbol{w}_i\|\|\boldsymbol{w}_j\|\left(\sin\theta_{ij}+\left(\pi-\theta_{ij}\right)\cos\theta_{ij}\right) - \frac{1}{2}\sum_{i=1}^{k} v_i \boldsymbol{a}^T\boldsymbol{w}_i + \frac{\|\boldsymbol{a}\|^2}{2} \end{aligned} \tag{9}$$

where equation (a) follows from the definition of $f(\boldsymbol{a}, \boldsymbol{b})$, (b) follows from the Stein's Lemma, and finally (c) follows from the fact that derivative of ReLU activation is the step function and $\boldsymbol{w}_i^T\boldsymbol{x} > 0$ with probability $\frac{1}{2}$.

We also write this in a more compact matrix form as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{4\pi}\boldsymbol{u}^T\left(\sin(\boldsymbol{\Theta}) + \left(\pi\mathbb{1}\mathbb{1}^T - \boldsymbol{\Theta}\right)\odot\cos(\boldsymbol{\Theta})\right)\boldsymbol{u} - \frac{1}{2}\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{v} + \frac{1}{2}\|\boldsymbol{a}\|^2$$

where $\boldsymbol{\omega}_i = \|\boldsymbol{w}_i\|$, $\boldsymbol{u} = diag(\boldsymbol{\omega})\,\boldsymbol{v}$, and $\boldsymbol{\theta}_{ij}$ is the angle between $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$.

A.2. Population Gradient

Gradient w.r.t. $\boldsymbol{W}$: Let us define,

$$\begin{aligned} g(\boldsymbol{a},\boldsymbol{b}) &= \frac{\partial}{\partial \boldsymbol{a}} f(\boldsymbol{a},\boldsymbol{b}) \\ &= \frac{1}{2\pi}\left(\|\boldsymbol{b}\|\sin(\theta_{\boldsymbol{a},\boldsymbol{b}})\,\bar{\boldsymbol{a}} + (\pi - \theta_{\boldsymbol{a},\boldsymbol{b}})\,\boldsymbol{b}\right) \quad \left(\bar{\boldsymbol{a}} = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}, \quad \bar{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right) \\ &= \frac{\|\boldsymbol{b}\|}{2\pi}\left(\sin(\theta_{\boldsymbol{a},\boldsymbol{b}})\,\bar{\boldsymbol{a}} + (\pi - \theta_{\boldsymbol{a},\boldsymbol{b}})\,\bar{\boldsymbol{b}}\right). \end{aligned} \tag{10}$$

Taking the derivative of (9) with respect to $\boldsymbol{w}_i$, we get

$$\begin{aligned} \nabla_{\boldsymbol{w_i}}\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2}v_i^2\boldsymbol{w_i} + \sum_{\substack{j=1\\ i\neq j}}^{k} v_i v_j g(\boldsymbol{w_i},\boldsymbol{w_j}) - \frac{1}{2}v_i\boldsymbol{a} \\ &= \sum_{j=1}^{k} v_i v_j g(\boldsymbol{w_i},\boldsymbol{w_j}) - \frac{1}{2}v_i\boldsymbol{a} \\ &= \frac{1}{2\pi}\sum_{j=1}^{k} v_i v_j \|\boldsymbol{w_j}\|\left(\sin\theta_{ij}\bar{\boldsymbol{w}}_{\boldsymbol{i}} + (\pi - \theta_{ij})\,\bar{\boldsymbol{w}}_{\boldsymbol{j}}\right) - \frac{1}{2}v_i\boldsymbol{a} \end{aligned}$$

In matrix form:

$$\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2\pi}diag(\boldsymbol{v})\left(\left(\pi\mathbb{1}\mathbb{1}^T - \boldsymbol{\Theta}\right)diag(\boldsymbol{u}) + diag(\sin(\boldsymbol{\Theta})\,\boldsymbol{u})\right)\bar{\boldsymbol{W}} - \frac{1}{2}\boldsymbol{v}\boldsymbol{a}^T \tag{11}$$

where $\boldsymbol{\omega}_i = \|\boldsymbol{w}_i\|$, and $\boldsymbol{u} = diag(\boldsymbol{\omega})\,\boldsymbol{v}$.

Gradient w.r.t. $\boldsymbol{v}$: Taking the derivative of (9) with respect to v_i, we get

$$\begin{aligned} \nabla_{v_i}\mathcal{L}(\boldsymbol{\theta}) &= v_i f(\boldsymbol{w_i},\boldsymbol{w_i}) + \sum_{\substack{j=1\\ i\neq j}}^{k} v_j f(\boldsymbol{w_i},\boldsymbol{w_j}) - \frac{1}{2}\boldsymbol{a}^T\boldsymbol{w_i} \\ &= \sum_{j=1}^{k} v_j f(\boldsymbol{w_i},\boldsymbol{w_j}) - \frac{1}{2}\boldsymbol{a}^T\boldsymbol{w_i} \\ &= \frac{1}{2\pi}\sum_{j=1}^{k} v_j \|\boldsymbol{w_i}\|\|\boldsymbol{w_j}\|\left(\sin\theta_{ij} + (\pi - \theta_{ij})\cos\theta_{ij}\right) - \frac{1}{2}\boldsymbol{a}^T\boldsymbol{w_i} \end{aligned}$$

In matrix form:

$$\nabla_{\boldsymbol{v}} Loss = \frac{1}{2\pi} diag\left(\boldsymbol{\omega}\right)\left(\sin\boldsymbol{\Theta} + \cos\boldsymbol{\Theta}\odot\left(\pi\mathbb{1}\mathbb{1}^T - \boldsymbol{\Theta}\right)\right) diag\left(\boldsymbol{\omega}\right)\boldsymbol{v} - \frac{1}{2}\boldsymbol{W}\boldsymbol{a} \tag{12}$$

where $\boldsymbol{\omega}_i = \|\boldsymbol{w}_i\|$. Finally, we note that the gradient w.r.t $\boldsymbol{w}_i$ and v_i are related with the following simple identity:

$$\boldsymbol{w}_i^T \nabla_{\boldsymbol{w_i}} \mathcal{L}\left(\boldsymbol{\theta}\right) = v_i \nabla_{v_i} \mathcal{L}\left(\boldsymbol{\theta}\right). \tag{13}$$

A.3. Population Hessian

The Hessian consists of four blocks (3 unique) due to interaction of $\boldsymbol{v}$ and $\boldsymbol{W}$ terms. We provide these individual blocks below and calculations in the following subsections. Define $\bar{\boldsymbol{w}}_l = \frac{\boldsymbol{w}_l}{\|\boldsymbol{w}_l\|}$, $\boldsymbol{P}_{\boldsymbol{w}_l^\perp} = \left(I - \bar{\boldsymbol{w}}_l \bar{\boldsymbol{w}}_l{}^T\right)$, and $\boldsymbol{w}_{\ell,m^\perp} = \boldsymbol{P}_{\boldsymbol{w}_m^\perp}\boldsymbol{w}_\ell$. Then we have,

$$\nabla^2_{v_\ell, v_m}\mathcal{L}\left(\boldsymbol{\theta}\right) = \frac{\|\boldsymbol{w}_\ell\|\|\boldsymbol{w}_m\|}{2\pi}\left(\left(\pi - \theta_{\ell,m}\right)\cos\theta_{\ell,m} + \sin\theta_{\ell,m}\right), \tag{14}$$

$$\nabla^2_{v_\ell, \boldsymbol{w}_m}\mathcal{L}\left(\boldsymbol{\theta}\right) = \begin{cases} \frac{v_\ell \boldsymbol{w}_\ell^T - \boldsymbol{a}^T}{2} + \sum_{i=1}^k \frac{v_i\|\boldsymbol{w}_i\|}{2\pi}\left(\left(\pi - \theta_{\ell,i}\right)\bar{\boldsymbol{w}}_i^T + \sin\theta_{\ell,i}\bar{\boldsymbol{w}}_\ell^T\right) & \ell = m \\ \frac{v_m\|\boldsymbol{w}_\ell\|}{2\pi}\left(\left(\pi - \theta_{\ell,m}\right)\bar{\boldsymbol{w}}_\ell^T + \sin\theta_{\ell,m}\bar{\boldsymbol{w}}_m^T\right) & \ell \neq m \end{cases},$$

$$\nabla^2_{\boldsymbol{w}_\ell, \boldsymbol{w}_m}\mathcal{L}\left(\boldsymbol{\theta}\right) = \begin{cases} \frac{\boldsymbol{v}_\ell^2}{2}\boldsymbol{I} + \frac{\boldsymbol{v}_\ell}{2\pi\|\boldsymbol{w}_\ell\|}\sum_{i=1}^k \boldsymbol{v}_i\|\boldsymbol{w}_i\|\sin\left(\theta_{\ell,i}\right)\left(\boldsymbol{P}_{\boldsymbol{w}_l^\perp} + \frac{\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\bar{\boldsymbol{w}}_i^T\boldsymbol{P}_{\boldsymbol{w}_l^\perp}}{\left\|\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\right\|^2}\right) & \ell = m \\ \frac{\boldsymbol{v}_\ell\boldsymbol{v}_m}{2\pi}\left(\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}_{m,\ell^\perp}^T + \bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_{\ell,m^\perp}^T + \left(\pi - \theta_{\ell,m}\right)\boldsymbol{I}\right) & \ell \neq m \end{cases}$$

A.3.1. CALCULATING THE $\boldsymbol{v}, \boldsymbol{v}$ BLOCK

We have,

$$\begin{aligned}\nabla^2_{v_\ell, v_m}\mathcal{L}\left(\boldsymbol{\theta}\right) &= \mathbb{E}_{\boldsymbol{x}}\left[\left(f\left(\boldsymbol{\theta};\boldsymbol{x}\right) - \boldsymbol{a}^T\boldsymbol{x}\right)\nabla^2_{v_\ell, v_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right) + \nabla_{v_\ell} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\nabla_{v_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)^T\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\nabla_{v_\ell} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\nabla_{v_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)^T\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_\ell^T\boldsymbol{x}\right)\phi\left(\boldsymbol{w}_m^T\boldsymbol{x}\right)\right] \\ &= \frac{\|\boldsymbol{w}_\ell\|\|\boldsymbol{w}_m\|}{2\pi}\left(\left(\pi - \theta_{\ell,m}\right)\cos\theta_{\ell,m} + \sin\theta_{\ell,m}\right)\end{aligned}$$

where the last step follows from the Table 1 in (Daniely et al., 2016).

A.3.2. CALCULATING THE $\boldsymbol{v}$, VECT $(\boldsymbol{W})$ BLOCK

We have,

$$\nabla^2_{v_\ell, \boldsymbol{w}_m}\mathcal{L}\left(\boldsymbol{\theta}\right) = \mathbb{E}_{\boldsymbol{x}}\left[\left(f\left(\boldsymbol{\theta};\boldsymbol{x}\right) - \boldsymbol{a}^T\boldsymbol{x}\right)\nabla^2_{v_\ell, \boldsymbol{w}_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right) + \nabla_{v_\ell} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\nabla_{\boldsymbol{w}_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)^T\right].$$

for calculation of individual terms refer down below.

Calculating the $\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{v_\ell} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\nabla_{\boldsymbol{w}_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)^T\right]$ term: We have,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{v_\ell} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\nabla_{\boldsymbol{w}_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)^T\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_\ell^T\boldsymbol{x}\right) v_m\phi'\left(\boldsymbol{w}_m^T\boldsymbol{x}\right)\boldsymbol{x}^T\right] \\
&= v_m\left\|\boldsymbol{w}_\ell\right\|\mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\boldsymbol{x}^T\right] \\
&\overset{(a)}{=} v_m\left\|\boldsymbol{w}_\ell\right\|\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}}\left(\phi\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_m^T\boldsymbol{x}\right)\right)\right] \\
&= v_m\left\|\boldsymbol{w}_\ell\right\|\mathbb{E}_{\boldsymbol{x}}\left[\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\bar{\boldsymbol{w}}_\ell^T + \phi\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\bar{\boldsymbol{w}}_m^T\right] \\
&\overset{(b)}{=} v_m\left\|\boldsymbol{w}_\ell\right\|\left(\left(\frac{\pi-\theta_{\ell,m}}{2\pi}\right)\bar{\boldsymbol{w}}_\ell^T + \mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\bar{\boldsymbol{w}}_m^T\right]\right)
\end{aligned}$$

where equation (a) follows from Stein's Lemma, and (b) follows from the dual activation of a step function. The handle the remaining expectation, we first define $g = \bar{\boldsymbol{w}}_m^T\boldsymbol{x} \sim \mathcal{N}\left(0,1\right)$. Then,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{P}_{\boldsymbol{w}_m^\perp}\boldsymbol{x} + \bar{\boldsymbol{w}}_\ell^T\bar{\boldsymbol{w}}_m g\right)\delta\left(g\right)\right] \\
&= \frac{1}{\sqrt{2\pi}}\mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{P}_{\boldsymbol{w}_m^\perp}\boldsymbol{x}\right)\right] \quad \text{(Delta integration)} \\
&= \frac{\left\|\boldsymbol{P}_{\boldsymbol{w}_m^\perp}\bar{\boldsymbol{w}}_\ell\right\|}{\sqrt{2\pi}}\mathbb{E}_u\left[\phi\left(u\right)\right] \quad \left(u \sim N\left(0,1\right)\right) \\
&= \frac{\sin\theta_{\ell,m}}{2\pi} \quad \left(\text{Expectation of rectified Gaussian } f_{\boldsymbol{x}}(0) = \frac{1}{\sqrt{2\pi}}\right).
\end{aligned}$$

Combining everything:

$$\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{v_\ell} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\nabla_{\boldsymbol{w}_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)^T\right] = \frac{v_m\left\|\boldsymbol{w}_\ell\right\|}{2\pi}\left(\left(\pi-\theta_{\ell,m}\right)\bar{\boldsymbol{w}}_\ell^T + \sin\theta_{\ell,m}\bar{\boldsymbol{w}}_m^T\right).$$

Calculating the $\mathbb{E}_{\boldsymbol{x}}\left[\left(f\left(\boldsymbol{\theta};\boldsymbol{x}\right) - \boldsymbol{a}^T\boldsymbol{x}\right)\nabla^2_{v_\ell,\boldsymbol{w}_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\right]$ term: Note that

$$\nabla^2_{v_\ell,\boldsymbol{w}_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right) = \mathbb{1}\left\{\ell = m\right\}\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\boldsymbol{x}^T.$$

Hence we focus only on $\ell = m$ case.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\left(f\left(\boldsymbol{\theta};\boldsymbol{x}\right) - \boldsymbol{a}^T\boldsymbol{x}\right)\nabla^2_{v_\ell,\boldsymbol{w}_m} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\right] &= \mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\boldsymbol{x}^T\right] \\
&\overset{(a)}{=} \mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}} r\left(\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right) + r\left(\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\bar{\boldsymbol{w}}_\ell^T\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i=1}^k v_i\phi'\left(\boldsymbol{w}_i^T\boldsymbol{x}\right)\boldsymbol{w}_i - \boldsymbol{a}\right)^T\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right) + r\left(\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\bar{\boldsymbol{w}}_\ell^T\right] \\
&\overset{(b)}{=} \sum_{i=1}^k v_i\boldsymbol{w}_i^T\left(\frac{\pi-\theta_{\ell,i}}{2\pi}\right) - \frac{\boldsymbol{a}}{2} + \mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\right]\bar{\boldsymbol{w}}_\ell^T,
\end{aligned}$$

where (a) follows from the Stein's identity, and (b) follows from the dual activation of step function. To handle the remaining expectation term, define $g = \bar{\boldsymbol{w}}_\ell^T \boldsymbol{x} \sim \mathcal{N}(0,1)$. Then,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[r(\boldsymbol{x})\,\delta\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\right] &= \mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{P}_{\boldsymbol{w}_\ell^\perp}\boldsymbol{x} + \bar{\boldsymbol{w}}_\ell g\right)\delta(g)\right] \\
&= \frac{1}{\sqrt{2\pi}}\mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\boldsymbol{x}\right)\right] \quad \text{(Delta integration)} \\
&= \frac{1}{\sqrt{2\pi}}\sum_{i=1}^{k}\boldsymbol{v}_i\mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^T\boldsymbol{P}_{\boldsymbol{w}_\ell^\perp}\boldsymbol{x}\right)\right] \\
&= \frac{1}{\sqrt{2\pi}}\sum_{i=1}^{k}\boldsymbol{v}_i\left\|\boldsymbol{P}_{\boldsymbol{w}_\ell^\perp}\boldsymbol{w}_i\right\|\mathbb{E}_u\left[\phi(u)\right] \quad (u\sim N(0,1)) \\
&= \frac{1}{2\pi}\sum_{i=1}^{k}\boldsymbol{v}_i\left\|\boldsymbol{w}_i\right\|\sin\left(\theta_{\ell,i}\right) \quad \left(\text{Expectation of rectified Gaussian } f_{\boldsymbol{x}}(0) = \frac{1}{\sqrt{2\pi}}\right)
\end{aligned}$$

Combining everything,

$$\mathbb{E}_{\boldsymbol{x}}\left[\left(f(\boldsymbol{\theta};\boldsymbol{x}) - \boldsymbol{a}^T\boldsymbol{x}\right)\nabla^2_{v_\ell,\boldsymbol{w}_m} f(\boldsymbol{\theta};\boldsymbol{x})\right] = \sum_{i=1}^{k}\frac{v_i\left\|\boldsymbol{w}_i\right\|}{2\pi}\left(\left(\pi-\theta_{\ell,i}\right)\bar{\boldsymbol{w}}_i^T + \sin\theta_{\ell,i}\bar{\boldsymbol{w}}_\ell^T\right) - \frac{\boldsymbol{a}^T}{2},$$

when $\ell = m$. Otherwise, this term is $\mathbf{0}$.

A.3.3. CALCULATING THE VECT$(\boldsymbol{W})$, VECT$(\boldsymbol{W})$ BLOCK

We have,

$$\nabla^2_{\boldsymbol{w}_\ell,\boldsymbol{w}_m}\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\left(f(\boldsymbol{\theta};\boldsymbol{x}) - \boldsymbol{a}^T\boldsymbol{x}\right)\nabla^2_{\boldsymbol{w}_\ell,\boldsymbol{w}_m} f(\boldsymbol{\theta};\boldsymbol{x}) + \nabla_{\boldsymbol{w}_\ell} f(\boldsymbol{\theta};\boldsymbol{x})\,\nabla_{\boldsymbol{w}_m} f(\boldsymbol{\theta};\boldsymbol{x})^T\right].$$

for calculation of individual terms refer down below.
Calculating the $\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{w}_\ell} f(\boldsymbol{\theta};\boldsymbol{x})\,\nabla_{\boldsymbol{w}_m} f(\boldsymbol{\theta};\boldsymbol{x})^T\right]$ term: We have,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{w}_\ell} f(\boldsymbol{\theta};\boldsymbol{x})\,\nabla_{\boldsymbol{w}_m} f(\boldsymbol{\theta};\boldsymbol{x})^T\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{v}_\ell\phi'\left(\boldsymbol{w}_\ell^T\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^T\phi'\left(\boldsymbol{w}_m^T\boldsymbol{x}\right)\boldsymbol{v}_m\right] \\
&= \boldsymbol{v}_\ell\boldsymbol{v}_m\mathbb{E}_{\boldsymbol{x}}\left[\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^T\right]
\end{aligned}$$

To tackle the expectation term, we use second order Stein's Lemma, $\mathbb{E}_{\boldsymbol{x}}\left[g(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\nabla^2_{\boldsymbol{x}} g(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}}\left[g(\boldsymbol{x})\right]\boldsymbol{I}$.

$$\mathbb{E}_{\boldsymbol{x}}\left[\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^T\right] = \mathbb{E}_{\boldsymbol{x}}\left[\nabla^2_{\boldsymbol{x}}\left(\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\right)\right] + \mathbb{E}_{\boldsymbol{x}}\left[\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\right]\boldsymbol{I}$$

<u>First term is:</u>

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\nabla^2_{\boldsymbol{x}}\left(\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\right)\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\delta'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}_\ell^T\right] \\
&\quad + \mathbb{E}_{\boldsymbol{x}}\left[\delta\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\left(\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}_m^T + \bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_\ell^T\right)\right] \\
&\quad + \mathbb{E}_{\boldsymbol{x}}\left[\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\delta'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_m^T\right]
\end{aligned}$$

These terms can be grouped in two.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[\delta'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\right] &= \mathbb{E}_{\boldsymbol{x},g}\left[\delta'\left(g\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{\ell}}^\perp}\boldsymbol{x}+\bar{\boldsymbol{w}}_m^T\bar{\boldsymbol{w}}_\ell g\right)\right]\\ &= -\frac{\cos\left(\theta_{\ell,m}\right)}{\sqrt{2\pi}}\mathbb{E}_{\boldsymbol{x}}\left[\delta\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{\ell}}^\perp}\boldsymbol{x}\right)\right] \quad \left(\delta'\left(x\right)f\left(x\right)=-f'\left(0\right)\delta\left(x\right)\right)\\ &= -\frac{\cos\left(\theta_{\ell,m}\right)}{2\pi\left\|\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{\ell}}^\perp}\bar{\boldsymbol{w}}_m\right\|} = -\frac{\cos\left(\theta_{\ell,m}\right)}{2\pi\sin\left(\theta_{\ell,m}\right)}\end{aligned}$$

and the other one is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[\delta\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\right] &= \mathbb{E}_{\boldsymbol{x},g}\left[\delta\left(g\right)\delta\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{\ell}}^\perp}\boldsymbol{x}+\bar{\boldsymbol{w}}_m^T\bar{\boldsymbol{w}}_\ell g\right)\right]\\ &= \frac{1}{2\pi\left\|\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{\ell}}^\perp}\bar{\boldsymbol{w}}_m\right\|} = \frac{1}{2\pi\sin\left(\theta_{\ell,m}\right)}\end{aligned}$$

Therefore we get:

$$\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}}^2\left(\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\right)\right] = \frac{\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}_m^T+\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_\ell^T}{2\pi\sin\left(\theta_{\ell,m}\right)} - \frac{\cos\left(\theta_{\ell,m}\right)\left(\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}_\ell^T+\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_m^T\right)}{2\pi\sin\left(\theta_{\ell,m}\right)}$$

Second term is:

$$\mathbb{E}_{\boldsymbol{x}}\left[\phi'\left(\bar{\boldsymbol{w}}_\ell^T\boldsymbol{x}\right)\phi'\left(\bar{\boldsymbol{w}}_m^T\boldsymbol{x}\right)\right]\boldsymbol{I} = \left(\frac{\pi-\theta_{\ell,m}}{2\pi}\right)\boldsymbol{I} \quad \text{(Dual activation of step function)}$$

Combining everything:

$$\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{w}_\ell}f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\nabla_{\boldsymbol{w}_m}f\left(\boldsymbol{\theta};\boldsymbol{x}\right)^T\right] = \boldsymbol{v}_\ell\boldsymbol{v}_m\left(\frac{\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}_m^T+\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_\ell^T-\cos\left(\theta_{\ell,m}\right)\left(\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}_\ell^T+\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_m^T\right)}{2\pi\sin\left(\theta_{\ell,m}\right)}+\left(\frac{\pi-\theta_{\ell,m}}{2\pi}\right)\boldsymbol{I}\right)$$

or alternatively (by substituting $\cos\left(\theta_{i,j}\right)=\bar{w}_i^T\bar{w}_j$):

$$\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{w}_\ell}f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\nabla_{\boldsymbol{w}_m}f\left(\boldsymbol{\theta};\boldsymbol{x}\right)^T\right] = \frac{\boldsymbol{v}_\ell\boldsymbol{v}_m}{2\pi}\left(\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}_{m,\ell^\perp}^T+\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}_{\ell,m^\perp}^T+\left(\pi-\theta_{\ell,m}\right)\boldsymbol{I}\right)$$

Calculating the $\mathbb{E}_{\boldsymbol{x}}\left[\left(f\left(\boldsymbol{\theta};\boldsymbol{x}\right)-\boldsymbol{a}^T\boldsymbol{x}\right)\nabla_{\boldsymbol{w}_\ell,\boldsymbol{w}_m}^2f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\right]$ term: Note that

$$\nabla_{\boldsymbol{w}_\ell,\boldsymbol{w}_m}^2f\left(\boldsymbol{\theta};\boldsymbol{x}\right) = diag\left(\boldsymbol{v}\odot\phi''\left(\boldsymbol{W}\boldsymbol{x}\right)\right)_{\ell,m}\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}.$$

This expectation is $\mathbf{0}$ when $\ell\neq m$. Define $g=\bar{\boldsymbol{w}}_{\boldsymbol{l}}{}^T\boldsymbol{x}\sim\mathcal{N}\left(0,1\right)$.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[\left(f\left(\boldsymbol{\theta};\boldsymbol{x}\right)-\boldsymbol{a}^T\boldsymbol{x}\right)\nabla_{\boldsymbol{w}_\ell,\boldsymbol{w}_\ell}^2f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\right] &= \mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{x}\right)\boldsymbol{v}_\ell\delta\left(\boldsymbol{w}_\ell^T\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\right]\\ &= \frac{\boldsymbol{v}_\ell}{\left\|\boldsymbol{w}_\ell\right\|}\mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{x}\right)\delta\left(\bar{\boldsymbol{w}}_\ell{}^T\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\right]\\ &= \frac{\boldsymbol{v}_\ell}{\left\|\boldsymbol{w}_\ell\right\|}\mathbb{E}_{\boldsymbol{x},g}\left[r\left(\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{l}}^\perp}\boldsymbol{x}+\bar{\boldsymbol{w}}_{\boldsymbol{l}}g\right)\delta\left(g\right)\left(\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{l}}^\perp}\boldsymbol{x}+\bar{\boldsymbol{w}}_{\boldsymbol{l}}g\right)\left(\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{l}}^\perp}\boldsymbol{x}+\bar{\boldsymbol{w}}_{\boldsymbol{l}}g\right)^T\right]\\ &= \frac{\boldsymbol{v}_\ell}{\sqrt{2\pi}\left\|\boldsymbol{w}_\ell\right\|}\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{l}}^\perp}\mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{l}}^\perp}\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\right]\boldsymbol{P}_{\boldsymbol{w}_{\boldsymbol{l}}^\perp}\end{aligned}$$

To tackle the expectation term, we use second order Stein's Lemma, $\mathbb{E}_{\boldsymbol{x}}\left[g(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\right]=\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}}^{2}g\left(\boldsymbol{x}\right)\right]+\mathbb{E}_{\boldsymbol{x}}\left[g\left(\boldsymbol{x}\right)\right]\boldsymbol{I}$.

$$\mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\right]=\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}}^{2}r\left(\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\right]+\mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\right]\boldsymbol{I}$$

First term is:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}}^{2}r\left(\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}}^{2}\left(\sum_{i=1}^{k}\boldsymbol{v}_{i}\phi\left(\boldsymbol{w}_{i}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\right)\right] \quad (\boldsymbol{a}^{T}\boldsymbol{x}\text{ vanishes.})\\
&= \sum_{i=1}^{k}\boldsymbol{v}_{i}\mathbb{E}_{\boldsymbol{x}}\left[\nabla_{\boldsymbol{x}}^{2}\phi\left(\boldsymbol{w}_{i}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\right]\\
&= \sum_{i=1}^{k}\boldsymbol{v}_{i}\mathbb{E}_{\boldsymbol{x}}\left[\delta\left(\boldsymbol{w}_{i}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{w}_{i}\boldsymbol{w}_{i}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\right]\\
&= \sum_{i=1}^{k}\frac{\boldsymbol{v}_{i}}{\left\|\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{w}_{\boldsymbol{i}}\right\|}\mathbb{E}_{u}\left[\delta\left(u\right)\right]\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{w}_{i}\boldsymbol{w}_{i}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\\
&= \frac{1}{\sqrt{2\pi}}\sum_{i=1}^{k}\frac{\boldsymbol{v}_{i}}{\left\|\boldsymbol{w}_{i}\right\|\sin\left(\theta_{\ell,i}\right)}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{w}_{i}\boldsymbol{w}_{i}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}} \quad \text{(Delta integration)}\\
&= \frac{1}{\sqrt{2\pi}}\sum_{i=1}^{k}\frac{\boldsymbol{v}_{i}\left\|\boldsymbol{w}_{i}\right\|}{\sin\left(\theta_{\ell,i}\right)}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\bar{\boldsymbol{w}}_{i}\bar{\boldsymbol{w}}_{i}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}
\end{aligned}$$

Second term is:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\right] &= \sum_{i=1}^{k}\boldsymbol{v}_{i}\mathbb{E}_{\boldsymbol{x}}\left[\phi\left(\boldsymbol{w}_{\boldsymbol{i}}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\right]\\
&= \sum_{i=1}^{k}\boldsymbol{v}_{i}\left\|\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{w}_{\boldsymbol{i}}\right\|\mathbb{E}_{u}\left[\phi\left(u\right)\right] \quad (u\sim N\left(0,1\right))\\
&= \frac{1}{\sqrt{2\pi}}\sum_{i=1}^{k}\boldsymbol{v}_{i}\left\|\boldsymbol{w}_{i}\right\|\sin\left(\theta_{\ell,i}\right) \quad \left(\text{Expectation of rectified Gaussian } f_{\boldsymbol{x}}(0)=\frac{1}{\sqrt{2\pi}}\right)
\end{aligned}$$

Combining both terms we get

$$\mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\right]=\frac{1}{\sqrt{2\pi}}\sum_{i=1}^{k}\boldsymbol{v}_{i}\left\|\boldsymbol{w}_{i}\right\|\left(\sin\left(\theta_{\ell,i}\right)\boldsymbol{I}+\frac{\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}\bar{\boldsymbol{w}}_{i}\bar{\boldsymbol{w}}_{i}^{T}\boldsymbol{P}_{\boldsymbol{w}_{l}^{\perp}}}{\sin\left(\theta_{\ell,i}\right)}\right).$$

Finally we plug this back to get:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[r\left(\boldsymbol{x}\right)\nabla^2_{\boldsymbol{w}_\ell,\boldsymbol{w}_\ell} f\left(\boldsymbol{\theta};\boldsymbol{x}\right)\right] &= \frac{\boldsymbol{v}_\ell}{2\pi\left\|\boldsymbol{w}_\ell\right\|}\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\left(\sum_{i=1}^k \boldsymbol{v}_i\left\|\boldsymbol{w}_i\right\|\left(\sin\left(\theta_{\ell,i}\right)\boldsymbol{I}+\frac{\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\bar{\boldsymbol{w}}_i^T\boldsymbol{P}_{\boldsymbol{w}_l^\perp}}{\sin\left(\theta_{\ell,i}\right)}\right)\right)\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\\
&= \frac{\boldsymbol{v}_\ell}{2\pi\left\|\boldsymbol{w}_\ell\right\|}\sum_{i=1}^k \boldsymbol{v}_i\left\|\boldsymbol{w}_i\right\|\left(\sin\left(\theta_{\ell,i}\right)\boldsymbol{P}_{\boldsymbol{w}_l^\perp}+\frac{\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\bar{\boldsymbol{w}}_i^T\boldsymbol{P}_{\boldsymbol{w}_l^\perp}}{\sin\left(\theta_{\ell,i}\right)}\right)\\
&= \frac{\boldsymbol{v}_\ell}{2\pi\left\|\boldsymbol{w}_\ell\right\|}\sum_{i=1}^k \boldsymbol{v}_i\left\|\boldsymbol{w}_i\right\|\sin\left(\theta_{\ell,i}\right)\left(\boldsymbol{P}_{\boldsymbol{w}_l^\perp}+\frac{\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\bar{\boldsymbol{w}}_i^T\boldsymbol{P}_{\boldsymbol{w}_l^\perp}}{\left\|\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\right\|^2}\right).
\end{aligned}$$

Appendix B. Proof of Key Lemmas in the Population Setting

For the simplicity of notation, let $\boldsymbol{w}_1=\boldsymbol{w}_1^{(\tau)},\boldsymbol{w}_2=\boldsymbol{w}_2^{(\tau)}$.

B.1. Proof of Gradient Smoothness Towards the Global Optima in the Population Case (Lemma 12)

To establish the imbalance bound (Lemma 3), we first introduce a key lemma that characterizes the gradient smoothness toward the global optima in the population case. This result relates the norm of the population gradient to the relative distance between the current parameters and the global optima:

Lemma 12 *Under the constraint $v_1=v_2=1$, the following inequality holds for all $\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{R}^d$:*

$$\left\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2+\left\|\nabla_{\boldsymbol{w}_2}\mathcal{L}\right\|^2\le\frac{5}{2}\left(\left\|\boldsymbol{w}_1-\boldsymbol{a}\right\|^2+\left\|\boldsymbol{w}_2+\boldsymbol{a}\right\|^2\right).$$

Proof

We begin by demonstrating that:

$$\left\|\sin\theta\left\|\boldsymbol{w}_2\right\|\bar{\boldsymbol{w}}_1+\left(\pi-\theta\right)\boldsymbol{w}_2\right\|\le\pi\left\|\boldsymbol{w}_1+\boldsymbol{w}_2\right\|. \tag{15}$$

Note that $0\le\theta\le\pi$ since it is the angle between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. We proceed by case analysis on the value of θ. When $0\le\theta<\frac{\pi}{2}$, we have

$$\begin{aligned}
\left\|\sin\theta\left\|\boldsymbol{w}_2\right\|\bar{\boldsymbol{w}}_1+\left(\pi-\theta\right)\boldsymbol{w}_2\right\| &\overset{(a)}{\le}\left\|\sin\theta\left\|\boldsymbol{w}_2\right\|\bar{\boldsymbol{w}}_1\right\|+\left\|\left(\pi-\theta\right)\boldsymbol{w}_2\right\|\\
&=\sin\theta\left\|\boldsymbol{w}_2\right\|+\left(\pi-\theta\right)\left\|\boldsymbol{w}_2\right\|\\
&\overset{(b)}{\le}\pi\left\|\boldsymbol{w}_2\right\|\\
&\overset{(c)}{\le}\pi\left\|\boldsymbol{w}_1+\boldsymbol{w}_2\right\|.
\end{aligned}$$

In Inequality (a) we use the triangle inequality. Inequality (b) follows from the fact that $\sin\theta\le\theta$ when $\theta\ge0$. Inequality (c) follows from the fact that $\theta\le\frac{\pi}{2}$.

When $\theta \geq \frac{\pi}{2}$, we observe that

$$\begin{aligned}
\|\sin\theta \|\boldsymbol{w}_2\| \bar{\boldsymbol{w}}_1 + (\pi-\theta)\boldsymbol{w}_2\| &\overset{(a)}{\leq} \|\sin\theta \|\boldsymbol{w}_2\| \bar{\boldsymbol{w}}_1\| + \|(\pi-\theta)\boldsymbol{w}_2\| \\
&= \sin\theta \|\boldsymbol{w}_2\| + (\pi-\theta)\|\boldsymbol{w}_2\| \\
&= \left(1 + \frac{\pi-\theta}{\sin\theta}\right)\sin\theta\|\boldsymbol{w}_2\| \\
&\overset{(b)}{\leq} \left(1 + \frac{\pi-\theta}{\sin\theta}\right)\|\boldsymbol{w}_1 + \boldsymbol{w}_2\| \\
&\overset{(c)}{\leq} \left(1 + \frac{\pi}{2}\right)\|\boldsymbol{w}_1 + \boldsymbol{w}_2\| \\
&\overset{(d)}{\leq} \pi\|\boldsymbol{w}_1 + \boldsymbol{w}_2\|.
\end{aligned}$$

In Inequality (a) we use the triangle inequality. Inequality (b) follows from the fact that $\boldsymbol{w}_1 + \boldsymbol{w}_2$ has a component with magnitude $\sin\theta\|\boldsymbol{w}_2\|$ perpendicular to $\boldsymbol{w}_1$. Inequality (c) follows since $\frac{\pi-\theta}{\sin\theta}$ attains its maximum at $\theta = \frac{\pi}{2}$ when restricted to the range $\theta \geq \frac{\pi}{2}$. Finally, (d) follows because $1 \leq \frac{\pi}{2}$. This finishes the proof of Ineq. 15. Note that due to symmetry we get the following as a corollary:

$$\|\sin\theta \|\boldsymbol{w}_1\| \bar{\boldsymbol{w}}_2 + (\pi-\theta)\boldsymbol{w}_1\| \leq \pi\|\boldsymbol{w}_1 + \boldsymbol{w}_2\|. \tag{16}$$

Under the constraint $v_1 = v_2 = 1$, the partial gradients with respect to $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are given separately by:

$$\begin{aligned}
\nabla_{\boldsymbol{w}_1}\mathcal{L} &= -\frac{\boldsymbol{a}}{2} + \frac{1}{2\pi}\left(\pi\boldsymbol{w}_1 - \sin\theta\|\boldsymbol{w}_2\|\bar{\boldsymbol{w}}_1 - (\pi-\theta)\boldsymbol{w}_2\right) \\
\nabla_{\boldsymbol{w}_2}\mathcal{L} &= \frac{\boldsymbol{a}}{2} - \frac{1}{2\pi}\left((\pi-\theta)\boldsymbol{w}_1 + \sin\theta\|\boldsymbol{w}_1\|\bar{\boldsymbol{w}}_2 - \pi\boldsymbol{w}_2\right).
\end{aligned}$$

Using Ineq. 15, we can write

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2 &= \left\|-\frac{\boldsymbol{a}}{2} + \frac{1}{2\pi}\left(\pi\boldsymbol{w}_1 - \sin\theta\|\boldsymbol{w}_2\|\bar{\boldsymbol{w}}_1 - (\pi-\theta)\boldsymbol{w}_2\right)\right\|^2 \\
&= \left\|\frac{\boldsymbol{w}_1 - \boldsymbol{a}}{2} - \frac{1}{2\pi}\left(\sin\theta\|\boldsymbol{w}_2\|\bar{\boldsymbol{w}}_1 + (\pi-\theta)\boldsymbol{w}_2\right)\right\|^2 \\
&\leq 2\left\|\frac{\boldsymbol{w}_1 - \boldsymbol{a}}{2}\right\|^2 + 2\left\|\frac{1}{2\pi}\left(\sin\theta\|\boldsymbol{w}_2\|\bar{\boldsymbol{w}}_1 + (\pi-\theta)\boldsymbol{w}_2\right)\right\|^2 \\
&\leq \frac{1}{2}\|\boldsymbol{w}_1 - \boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{w}_1 + \boldsymbol{w}_2\|^2 \\
&= \frac{1}{2}\|\boldsymbol{w}_1 - \boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{w}_1 - \boldsymbol{a} + \boldsymbol{a} + \boldsymbol{w}_2\|^2 \\
&\leq \frac{3}{2}\|\boldsymbol{w}_1 - \boldsymbol{a}\|^2 + \|\boldsymbol{w}_2 + \boldsymbol{a}\|^2.
\end{aligned}$$

Similarly, using Eq. 16 on the gradient for $\boldsymbol{w}_2$, we get

$$\|\nabla_{\boldsymbol{w}_2}\mathcal{L}\|^2 \leq \frac{3}{2}\|\boldsymbol{w}_2 + \boldsymbol{a}\|^2 + \|\boldsymbol{w}_1 - \boldsymbol{a}\|^2.$$

Combining these, we obtain

$$\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2+\|\nabla_{\boldsymbol{w}_2}\mathcal{L}\|^2\leq\frac{5}{2}\left(\|\boldsymbol{w}_1-\boldsymbol{a}\|^2+\|\boldsymbol{w}_2+\boldsymbol{a}\|^2\right).$$

This completes the proof of Lemma 12. ■

B.2. Proof of the PL Inequality in the Population Case (Lemma 8)

First, we show it is sufficient to analyze $v_1=v_2=1$. For $v_1,v_2>0$, we define $\tilde{\boldsymbol{w}}_i=v_i\boldsymbol{w}_i$. Then,

$$\mathcal{L}\left(\boldsymbol{\theta}\right)=\mathbb{E}_{\boldsymbol{x}}\left[\left(v_1\text{ReLU}\left(\boldsymbol{w}_1^T\boldsymbol{x}\right)-v_2\text{ReLU}\left(\boldsymbol{w}_2^T\boldsymbol{x}\right)-\boldsymbol{a}^T\boldsymbol{x}\right)^2\right]=\mathbb{E}_{\boldsymbol{x}}\left[\left(\text{ReLU}\left(\tilde{\boldsymbol{w}}_1^T\boldsymbol{x}\right)-\text{ReLU}\left(\tilde{\boldsymbol{w}}_2^T\boldsymbol{x}\right)-\boldsymbol{a}^T\boldsymbol{x}\right)^2\right].$$

Let us focus on squared gradient norms:

$$\begin{aligned}\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2+\|\nabla_{\boldsymbol{w}_2}\mathcal{L}\|^2&=\|v_1\nabla_{\tilde{\boldsymbol{w}}_1}\mathcal{L}\|^2+\|v_2\nabla_{\tilde{\boldsymbol{w}}_2}\mathcal{L}\|^2\\&\geq\min\left(v_1^2,v_2^2\right)\left(\|\nabla_{\tilde{\boldsymbol{w}}_1}\mathcal{L}\|^2+\|\nabla_{\tilde{\boldsymbol{w}}_2}\mathcal{L}\|^2\right)\end{aligned}$$

This suggests that proving $\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2+\|\nabla_{\boldsymbol{w}_2}\mathcal{L}\|^2\geq\alpha\mathcal{L}$ when $v_1=v_2=1$ implies that

$$\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2+\|\nabla_{\boldsymbol{w}_2}\mathcal{L}\|^2\geq\min\left(v_1^2,v_2^2\right)\alpha\mathcal{L}$$

for arbitrary $v_1,v_2>0$. Now, we assume $v_1=v_2=1$. We define

$$h\left(\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{a}\right)=\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2+\|\nabla_{\boldsymbol{w}_2}\mathcal{L}\|^2-\alpha\mathcal{L}.$$

Using the gradient calculations in (11), we can write it equivalently as

$$\begin{aligned}h\left(\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{a}\right)=&\frac{1}{4}\left(1-\alpha+\frac{\left(\pi-\theta\right)^2+2\left(\pi-\theta\right)\sin\theta\cos\theta+\left(\sin\theta\right)^2}{\pi^2}\right)\left(\|\boldsymbol{w}_1\|^2+\|\boldsymbol{w}_2\|^2\right)\\&-\left(1-\frac{\alpha}{2}\right)\frac{\left(\pi-\theta\right)\cos\theta+\sin\theta}{\pi}\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|\\&-\frac{1}{2}\left(\left(\frac{\pi-\theta}{\pi}+1-\alpha\right)\|\boldsymbol{w}_1\|-\frac{\sin\theta}{\pi}\|\boldsymbol{w}_2\|\right)\bar{\boldsymbol{w}}_1^T\boldsymbol{a}\\&+\frac{1}{2}\left(\left(\frac{\pi-\theta}{\pi}+1-\alpha\right)\|\boldsymbol{w}_2\|-\frac{\sin\theta}{\pi}\|\boldsymbol{w}_1\|\right)\bar{\boldsymbol{w}}_2^T\boldsymbol{a}\\&+\left(\frac{1-\alpha}{2}\right)\|\boldsymbol{a}\|^2\\=&\frac{1-\alpha}{2}\|\boldsymbol{a}\|^2+\boldsymbol{b}^T\boldsymbol{a}+c\end{aligned}$$

where $\boldsymbol{b}$ and c are defined by the following terms for brevity,

$$\alpha_1 = -\frac{1}{2}\left(\left(\frac{\pi-\theta}{\pi}+1-\alpha\right)\|\boldsymbol{w}_1\| - \frac{\sin\theta}{\pi}\|\boldsymbol{w}_2\|\right)$$
$$\alpha_2 = +\frac{1}{2}\left(\left(\frac{\pi-\theta}{\pi}+1-\alpha\right)\|\boldsymbol{w}_2\| - \frac{\sin\theta}{\pi}\|\boldsymbol{w}_1\|\right)$$
$$c = \frac{1}{4}\left(1-\alpha+\frac{(\pi-\theta)^2+2(\pi-\theta)\sin\theta\cos\theta+(\sin\theta)^2}{\pi^2}\right)\left(\|\boldsymbol{w}_1\|^2+\|\boldsymbol{w}_2\|^2\right)$$
$$-\left(1-\frac{\alpha}{2}\right)\frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|$$
$$\boldsymbol{b} = \alpha_1\bar{\boldsymbol{w}}_1+\alpha_2\bar{\boldsymbol{w}}_2$$

Noting that the expression above is quadratic in $\boldsymbol{a}$, we compute $\tilde{h}(\boldsymbol{w}_1,\boldsymbol{w}_2) = \min_{\boldsymbol{a}} h(\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{a})$. The choice of $\boldsymbol{a}$ that minimizes the expression is $\boldsymbol{a} = -\frac{\boldsymbol{b}}{1-\alpha}$. Plugging this in back we get,

$$\tilde{h}(\boldsymbol{w}_1,\boldsymbol{w}_2) = c - \frac{\|\boldsymbol{b}\|^2}{2(1-\alpha)} = c - \frac{\alpha_1^2+2\alpha_1\alpha_2\cos\theta+\alpha_2^2}{2(1-\alpha)}$$

Taking the norm out: Note that we are only interested in the positivity of $\tilde{h}$, therefore dividing it by $\|\boldsymbol{w}_2\|^2$ does not change the sign. Denote $\frac{\|\boldsymbol{w}_1\|}{\|\boldsymbol{w}_2\|} = r$. Then we still have

$$\frac{\tilde{h}(\boldsymbol{w}_1,\boldsymbol{w}_2)}{\|\boldsymbol{w}_2\|^2} = c - \frac{\alpha_1^2+2\alpha_1\alpha_2\cos\theta+\alpha_2^2}{2(1-\alpha)}$$

but the variables are modified as

$$\alpha_1 = -\frac{1}{2}\left(\left(\frac{\pi-\theta}{\pi}+1-\alpha\right)r - \frac{\sin\theta}{\pi}\right)$$
$$\alpha_2 = +\frac{1}{2}\left(\left(\frac{\pi-\theta}{\pi}+1-\alpha\right) - \frac{\sin\theta}{\pi}r\right)$$
$$c = \frac{1}{4}\left(1-\alpha+\frac{(\pi-\theta)^2+2(\pi-\theta)\sin\theta\cos\theta+(\sin\theta)^2}{\pi^2}\right)\left(r^2+1\right) - \left(1-\frac{\alpha}{2}\right)\frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}r$$

We note that the expression is of the form $C_1\left(r^2+1\right)+C_2 r$. Without changing the sign, we can take out r outside. Then we notice that the minima is achieved at $r=1$. Therefore, it is sufficient for us to check the positivity of the expression at $r=1$. That is, we draw:

$$\left.\frac{\tilde{h}(\boldsymbol{w}_1,\boldsymbol{w}_2)}{\|\boldsymbol{w}_2\|^2}\right|_{r=1} = c - \frac{(1-\cos\theta)}{(1-\alpha)}\tilde{\alpha}^2$$

where

$$\tilde{\alpha} = \frac{1}{2}\left(\frac{\pi-\theta-\sin\theta}{\pi}+1-\alpha\right),$$
$$c = \frac{1}{2}\left(1-\alpha+\frac{(\pi-\theta)^2+2(\pi-\theta)\sin\theta\cos\theta+(\sin\theta)^2}{\pi^2}\right) - \left(1-\frac{\alpha}{2}\right)\frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}.$$

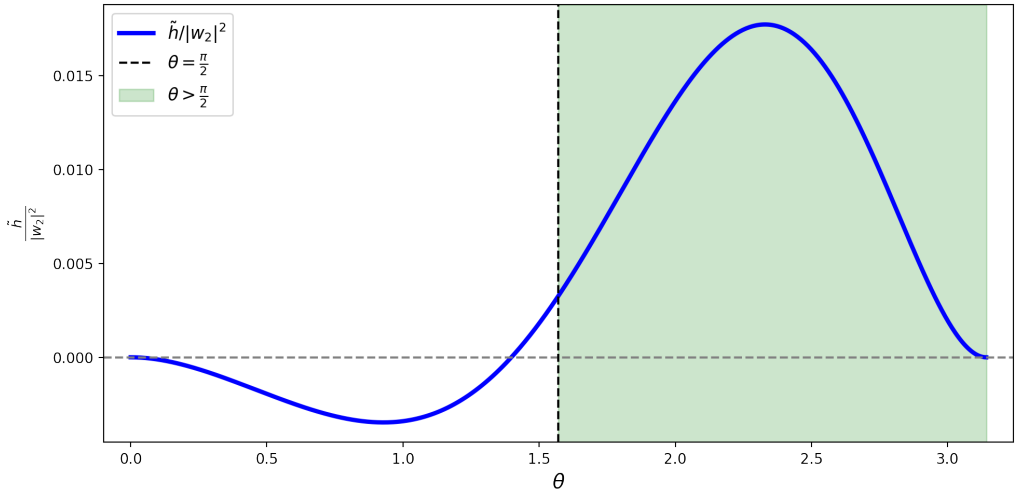


Figure 4: $\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2 + \|\nabla_{\boldsymbol{w}_2}\mathcal{L}\|^2 - \alpha\mathcal{L}$ **is non-negative**. We set $\alpha = 0.05$ and draw $\frac{1}{\|\boldsymbol{w}_2\|}\tilde{h}\left(\boldsymbol{w}_1, \boldsymbol{w}_2\right)$ for $\theta \in [0, \pi]$. We show that $\frac{1}{\|\boldsymbol{w}_2\|}\tilde{h}\left(\boldsymbol{w}_1, \boldsymbol{w}_2\right)$ is non-negative inside the shaded region ($\theta \geq \frac{\pi}{2}$).

To complete the proof, in Figure 4, we set $\alpha = 0.05$ and draw $\left.\frac{\tilde{h}(\boldsymbol{w}_1, \boldsymbol{w}_2)}{\|\boldsymbol{w}_2\|^2}\right|_{r=1}$ as a 1D plot for $\theta \in [0, \pi]$. The plot demonstrates that $\tilde{h}$ is non-negative for $\theta > \frac{\pi}{2}$. This finishes the proof.

B.3. Bound on the Smoothness of the Population Loss (Lemma 9)

We bound the population Hessian $\nabla^2\mathcal{L}\left(\boldsymbol{v}, \boldsymbol{W}\right)$ in the local refinement phase. That is, we assume $c_1\sqrt{\|\boldsymbol{a}\|} \leq v_1, v_2, \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \leq c_2\sqrt{\|\boldsymbol{a}\|}$. By the sub-additivity properties of the spectral norm, we have

$$\left\|\nabla^2\mathcal{L}\left(\boldsymbol{v}, \boldsymbol{W}\right)\right\|_2 \leq \left\|\nabla^2_{\boldsymbol{v},\boldsymbol{v}}\mathcal{L}\left(\boldsymbol{v}, \boldsymbol{W}\right)\right\|_2 + 2\left\|\nabla^2_{\boldsymbol{v},\mathrm{vect}(\boldsymbol{W})}\mathcal{L}\left(\boldsymbol{v}, \boldsymbol{W}\right)\right\|_2 + \left\|\nabla^2_{\mathrm{vect}(\boldsymbol{W}),\mathrm{vect}(\boldsymbol{W})}\mathcal{L}\left(\boldsymbol{v}, \boldsymbol{W}\right)\right\|_2.$$

We bound each term separately below.

$\nabla^2_{\boldsymbol{v},\boldsymbol{v}}\mathcal{L}\left(\boldsymbol{\theta}\right)$ term: We have

$$\left\|\nabla^2_{\boldsymbol{v},\boldsymbol{v}}\mathcal{L}\left(\boldsymbol{\theta}\right)\right\|_2 \leq \sum_{\ell=1}^{2}\sum_{m=1}^{2}\left|\nabla^2_{v_\ell, v_m}\mathcal{L}\left(\boldsymbol{\theta}\right)\right|.$$

where,

$$\begin{aligned}
\left|\nabla^2_{v_\ell, v_m}\mathcal{L}\left(\boldsymbol{\theta}\right)\right| =& \frac{\|\boldsymbol{w}_\ell\|\,\|\boldsymbol{w}_m\|}{2\pi}\left(\left(\pi - \theta_{\ell,m}\right)\cos\theta_{\ell,m} + \sin\theta_{\ell,m}\right) \\
\leq& \frac{c_2^2\,\|\boldsymbol{a}\|}{2\pi}\left(\left(\pi - \theta_{\ell,m}\right)\cos\theta_{\ell,m} + \sin\theta_{\ell,m}\right) \\
\leq& \frac{c_2^2\,\|\boldsymbol{a}\|}{2\pi}\left(\pi + 1\right) \\
\leq& c_2^2\,\|\boldsymbol{a}\|\,\frac{1+\pi}{2\pi}.
\end{aligned}$$

Then,

$$\begin{aligned}\left\|\nabla^2_{\boldsymbol{v},\boldsymbol{v}}\mathcal{L}\left(\boldsymbol{\theta}\right)\right\|_2 \leq & \sum_{\ell=1}^{2}\sum_{m=1}^{2}\left|\nabla^2_{v_\ell,v_m}\mathcal{L}\left(\boldsymbol{\theta}\right)\right| \\ \leq & 4c_2^2\left\|\boldsymbol{a}\right\|\frac{1+\pi}{2\pi} = \left(2+\frac{2}{\pi}\right)c_2^2\left\|\boldsymbol{a}\right\|.\end{aligned}$$

$\nabla^2_{\boldsymbol{v},\mathbf{vect}(\boldsymbol{W})}\mathcal{L}\left(\boldsymbol{\theta}\right)$ **term:** We have

$$\left\|\nabla^2_{\boldsymbol{v},\mathrm{vect}(\boldsymbol{W})}\mathcal{L}\left(\boldsymbol{\theta}\right)\right\|_2 \leq \sum_{\ell=1}^{2}\sum_{m=1}^{2}\left\|\nabla^2_{v_\ell,\boldsymbol{w}_m}\mathcal{L}\left(\boldsymbol{\theta}\right)\right\|.$$

For $\ell \neq m$, we have

$$\begin{aligned}\left\|\nabla^2_{v_\ell,\boldsymbol{w}_m}\mathcal{L}\left(\boldsymbol{\theta}\right)\right\| = & \left\|\frac{v_m\left\|\boldsymbol{w}_\ell\right\|}{2\pi}\left(\left(\pi-\theta_{\ell,m}\right)\bar{\boldsymbol{w}}_\ell^T+\sin\theta_{\ell,m}\bar{\boldsymbol{w}}_m^T\right)\right\| \\ = & \frac{v_m\left\|\boldsymbol{w}_\ell\right\|}{2\pi}\left\|\left(\left(\pi-\theta_{\ell,m}\right)\bar{\boldsymbol{w}}_\ell+\sin\theta_{\ell,m}\bar{\boldsymbol{w}}_m\right)\right\| \\ \leq & \frac{v_m\left\|\boldsymbol{w}_\ell\right\|}{2\pi}\left(\left\|\left(\pi-\theta_{\ell,m}\right)\bar{\boldsymbol{w}}_\ell\right\|+\left\|\sin\theta_{\ell,m}\bar{\boldsymbol{w}}_m\right\|\right) \\ \leq & \frac{v_m\left\|\boldsymbol{w}_\ell\right\|}{2\pi}\left(\pi+1\right) \\ \leq & c_2^2\left\|\boldsymbol{a}\right\|\frac{1+\pi}{2\pi}.\end{aligned}$$

For $\ell = m$, we have

$$\begin{aligned}\left\|\nabla^2_{v_\ell,\boldsymbol{w}_m}\mathcal{L}\left(\boldsymbol{\theta}\right)\right\| = & \left\|\frac{v_\ell\boldsymbol{w}_\ell^T-\boldsymbol{a}^T}{2}+\sum_{i=1}^{2}\frac{v_i\left\|\boldsymbol{w}_i\right\|}{2\pi}\left(\left(\pi-\theta_{\ell,i}\right)\bar{\boldsymbol{w}}_i^T+\sin\theta_{\ell,i}\bar{\boldsymbol{w}}_\ell^T\right)\right\| \\ \leq & \frac{\left\|v_\ell\boldsymbol{w}_\ell\right\|}{2}+\frac{\left\|\boldsymbol{a}\right\|}{2}+\sum_{i=1}^{2}\left\|\frac{v_i\left\|\boldsymbol{w}_i\right\|}{2\pi}\left(\left(\pi-\theta_{\ell,i}\right)\bar{\boldsymbol{w}}_i^T+\sin\theta_{\ell,i}\bar{\boldsymbol{w}}_\ell^T\right)\right\| \\ \leq & \frac{\left\|v_\ell\boldsymbol{w}_\ell\right\|}{2}+\frac{\left\|\boldsymbol{a}\right\|}{2}+\sum_{i=1}^{2}\frac{v_i\left\|\boldsymbol{w}_i\right\|}{2\pi}\left(\left\|\left(\pi-\theta_{\ell,i}\right)\bar{\boldsymbol{w}}_i\right\|+\left\|\sin\theta_{\ell,i}\bar{\boldsymbol{w}}_\ell\right\|\right) \\ \leq & \frac{c_2^2\left\|\boldsymbol{a}\right\|}{2}+\frac{\left\|\boldsymbol{a}\right\|}{2}+\sum_{i=1}^{2}\frac{c_2^2\left\|\boldsymbol{a}\right\|}{2\pi}\left(\pi+1\right) = \left(\frac{1}{2}+\left(\frac{3}{2}+\frac{1}{\pi}\right)c_2^2\right)\left\|\boldsymbol{a}\right\|.\end{aligned}$$

Combining both inequalities, we have

$$\begin{aligned}\left\|\nabla^2_{\boldsymbol{v},\mathrm{vect}(\boldsymbol{W})}\mathcal{L}\left(\boldsymbol{\theta}\right)\right\|_2 \leq & \sum_{\ell=1}^{2}\sum_{m=1}^{2}\left\|\nabla^2_{v_\ell,\boldsymbol{w}_m}\mathcal{L}\left(\boldsymbol{\theta}\right)\right\| \\ \leq & 2\left(c_2^2\left\|\boldsymbol{a}\right\|\frac{1+\pi}{2\pi}+\left(\frac{1}{2}+\left(\frac{3}{2}+\frac{1}{\pi}\right)c_2^2\right)\left\|\boldsymbol{a}\right\|\right) \\ = & \left(1+\left(4+\frac{3}{\pi}\right)c_2^2\right)\left\|\boldsymbol{a}\right\|.\end{aligned}$$

$\nabla^2_{\mathbf{vect}(\boldsymbol{W}),\mathbf{vect}(\boldsymbol{W})}\mathcal{L}(\boldsymbol{\theta})$ **term:** We have

$$\left\|\nabla^2_{\mathrm{vect}(\boldsymbol{W}),\mathrm{vect}(\boldsymbol{W})}\mathcal{L}(\boldsymbol{\theta})\right\|_2 \leq \sum_{\ell=1}^{2}\sum_{m=1}^{2}\left\|\nabla^2_{\boldsymbol{w}_\ell,\boldsymbol{w}_m}\mathcal{L}(\boldsymbol{\theta})\right\|.$$

For $\ell \neq m$, we have

$$\begin{aligned}
\left\|\nabla^2_{\boldsymbol{w}_\ell,\boldsymbol{w}_m}\mathcal{L}(\boldsymbol{\theta})\right\|_2 =& \left\|\frac{v_\ell v_m}{2\pi}\left(\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}^T_{m,\ell^\perp}+\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}^T_{\ell,m^\perp}+(\pi-\theta_{\ell,m})\boldsymbol{I}\right)\right\|_2\\
=&\frac{v_\ell v_m}{2\pi}\left\|\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}^T_{m,\ell^\perp}+\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}^T_{\ell,m^\perp}+(\pi-\theta_{\ell,m})\boldsymbol{I}\right\|_2\\
\leq&\frac{v_\ell v_m}{2\pi}\left(\left\|\bar{\boldsymbol{w}}_\ell\bar{\boldsymbol{w}}^T_{m,\ell^\perp}\right\|_2+\left\|\bar{\boldsymbol{w}}_m\bar{\boldsymbol{w}}^T_{\ell,m^\perp}\right\|_2+\left\|(\pi-\theta_{\ell,m})\boldsymbol{I}\right\|_2\right)\\
\leq&\frac{v_\ell v_m}{2\pi}(1+1+\pi)\\
=&v_\ell v_m\frac{2+\pi}{2\pi}\\
\leq&c_2^2\left\|\boldsymbol{a}\right\|\frac{2+\pi}{2\pi}.
\end{aligned}$$

For $\ell = m$, we have

$$\begin{aligned}
\left\|\nabla^2_{\boldsymbol{w}_\ell,\boldsymbol{w}_m}\mathcal{L}(\boldsymbol{\theta})\right\|_2 =& \left\|\frac{v_\ell^2}{2}\boldsymbol{I}+\frac{v_\ell}{2\pi\left\|\boldsymbol{w}_\ell\right\|}\sum_{i=1}^{2}v_i\left\|\boldsymbol{w}_i\right\|\sin(\theta_{\ell,i})\left(\boldsymbol{P}_{\boldsymbol{w}_l^\perp}+\frac{\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\bar{\boldsymbol{w}}_i^T\boldsymbol{P}_{\boldsymbol{w}_l^\perp}}{\left\|\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\right\|^2}\right)\right\|_2\\
\leq&\left\|\frac{v_\ell^2}{2}\boldsymbol{I}\right\|_2+\frac{v_\ell}{2\pi\left\|\boldsymbol{w}_\ell\right\|}\sum_{i=1}^{2}\left\|v_i\left\|\boldsymbol{w}_i\right\|\sin(\theta_{\ell,i})\left(\boldsymbol{P}_{\boldsymbol{w}_l^\perp}+\frac{\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\bar{\boldsymbol{w}}_i^T\boldsymbol{P}_{\boldsymbol{w}_l^\perp}}{\left\|\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\right\|^2}\right)\right\|_2\\
\leq&\left\|\frac{v_\ell^2}{2}\boldsymbol{I}\right\|_2+\frac{v_\ell}{2\pi\left\|\boldsymbol{w}_\ell\right\|}\sum_{i=1}^{2}v_i\left\|\boldsymbol{w}_i\right\|\sin(\theta_{\ell,i})\left(\left\|\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\right\|+\left\|\frac{\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\bar{\boldsymbol{w}}_i^T\boldsymbol{P}_{\boldsymbol{w}_l^\perp}}{\left\|\boldsymbol{P}_{\boldsymbol{w}_l^\perp}\bar{\boldsymbol{w}}_i\right\|^2}\right\|_2\right)\\
\leq&\frac{v_\ell^2}{2}+\frac{v_\ell}{2\pi\left\|\boldsymbol{w}_\ell\right\|}\sum_{i=1}^{2}v_i\left\|\boldsymbol{w}_i\right\|(1+1)\\
\leq&\frac{c_2^2\left\|\boldsymbol{a}\right\|}{2}+\frac{c_2}{\pi c_1}\left(2c_2^2\left\|\boldsymbol{a}\right\|\right)=c_2^2\left\|\boldsymbol{a}\right\|\left(\frac{1}{2}+\frac{2c_2}{\pi c_1}\right).
\end{aligned}$$

Combining both inequalities, we have

$$\begin{aligned}
\left\|\nabla^2_{\mathrm{vect}(\boldsymbol{W}),\mathrm{vect}(\boldsymbol{W})}\mathcal{L}(\boldsymbol{\theta})\right\|_2 \leq&\sum_{\ell=1}^{2}\sum_{m=1}^{2}\left\|\nabla^2_{\boldsymbol{w}_\ell,\boldsymbol{w}_m}\mathcal{L}(\boldsymbol{\theta})\right\|\\
\leq&2\left(c_2^2\left\|\boldsymbol{a}\right\|\frac{2+\pi}{2\pi}+c_2^2\left\|\boldsymbol{a}\right\|\left(\frac{1}{2}+\frac{2c_2}{\pi c_1}\right)\right)\\
=&\left(2+\frac{2}{\pi}+\frac{4c_2}{\pi c_1}\right)c_2^2\left\|\boldsymbol{a}\right\|.
\end{aligned}$$

Combining the terms: Putting everything together,

$$\begin{aligned}\left\|\nabla^2\mathcal{L}\left(\boldsymbol{v},\boldsymbol{W}\right)\right\|_2 \leq& \left\|\nabla^2_{\boldsymbol{v},\boldsymbol{v}}\mathcal{L}\left(\boldsymbol{v},\boldsymbol{W}\right)\right\|_2 + 2\left\|\nabla^2_{\boldsymbol{v},\mathrm{vect}(\boldsymbol{W})}\mathcal{L}\left(\boldsymbol{v},\boldsymbol{W}\right)\right\|_2 + \left\|\nabla^2_{\mathrm{vect}(\boldsymbol{W}),\mathrm{vect}(\boldsymbol{W})}\mathcal{L}\left(\boldsymbol{v},\boldsymbol{W}\right)\right\|_2 \\ \leq& \left(2+\frac{2}{\pi}\right)c_2^2\left\|\boldsymbol{a}\right\| + 2\left(1+\left(4+\frac{3}{\pi}\right)c_2^2\right)\left\|\boldsymbol{a}\right\| + \left(2+\frac{2}{\pi}+\frac{4c_2}{\pi c_1}\right)c_2^2\left\|\boldsymbol{a}\right\| \\ =& \left(2+\left(12+\frac{10}{\pi}+\frac{4c_2}{\pi c_1}\right)c_2^2\right)\left\|\boldsymbol{a}\right\| \\ :=& L\left\|\boldsymbol{a}\right\|.\end{aligned}$$

This completes the proof of Lemma 9.

B.4. Population Loss Lower Bound (Lemma 13)

Lemma 13 (Population Loss Lower Bound) *For $v_1, v_2 > 0$ and $\theta > \frac{\pi}{2}$. We have*

$$\left\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\right\|^2 + \left\|v_2\boldsymbol{w}_2 + \boldsymbol{a}\right\|^2 \leq 20\mathcal{L}\left(\boldsymbol{v},\boldsymbol{W}\right).$$

Proof Define $\tilde{\boldsymbol{w}}_i = v_i\boldsymbol{w}_i$. For $v_1, v_2 > 0$, both $\left\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\right\|^2 + \left\|v_2\boldsymbol{w}_2 + \boldsymbol{a}\right\|^2$ and $\mathcal{L}\left(\boldsymbol{v},\boldsymbol{W}\right)$ are only functions of $\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2$. Next, we define

$$h\left(\tilde{\boldsymbol{w}}_1,\tilde{\boldsymbol{w}}_2,\boldsymbol{a}\right) = \mathcal{L}\left(\tilde{\boldsymbol{w}}_1,\tilde{\boldsymbol{w}}_2\right) - \tilde{\alpha}\left(\left\|\tilde{\boldsymbol{w}}_1 - \boldsymbol{a}\right\|^2 + \left\|\tilde{\boldsymbol{w}}_2 + \boldsymbol{a}\right\|^2\right).$$

We can write it equivalently as

$$\begin{aligned}h\left(\tilde{\boldsymbol{w}}_1,\tilde{\boldsymbol{w}}_2,\boldsymbol{a}\right) =& \left(\frac{1}{4}-\tilde{\alpha}\right)\left(\left\|\tilde{\boldsymbol{w}}_1\right\|^2 + \left\|\tilde{\boldsymbol{w}}_2\right\|^2\right) - \frac{(\pi-\theta)\cos\theta + \sin\theta}{2\pi}\left\|\tilde{\boldsymbol{w}}_1\right\|\left\|\tilde{\boldsymbol{w}}_2\right\| \\ &- \left(\frac{1}{2}-2\tilde{\alpha}\right)\boldsymbol{a}^T\left(\tilde{\boldsymbol{w}}_1 - \tilde{\boldsymbol{w}}_2\right) + \left(\frac{1}{2}-2\tilde{\alpha}\right)\left\|\boldsymbol{a}\right\|^2\end{aligned}$$

Noting that the expression above is quadratic in $\boldsymbol{a}$, we compute $\tilde{h}\left(\tilde{\boldsymbol{w}}_1,\tilde{\boldsymbol{w}}_2\right) = \min_{\boldsymbol{a}} h\left(\tilde{\boldsymbol{w}}_1,\tilde{\boldsymbol{w}}_2,\boldsymbol{a}\right)$. The choice of $\boldsymbol{a}$ that minimizes the expression is $\boldsymbol{a} = \frac{\tilde{\boldsymbol{w}}_1 - \tilde{\boldsymbol{w}}_2}{2}$. Plugging this in back we get,

$$\begin{aligned}\tilde{h}\left(\boldsymbol{w}_1,\boldsymbol{w}_2\right) =& \left(\frac{1}{4}-\tilde{\alpha}\right)\left(\left\|\tilde{\boldsymbol{w}}_1\right\|^2 + \left\|\tilde{\boldsymbol{w}}_2\right\|^2\right) - \frac{(\pi-\theta)\cos\theta + \sin\theta}{2\pi}\left\|\tilde{\boldsymbol{w}}_1\right\|\left\|\tilde{\boldsymbol{w}}_2\right\| \\ &- \left(\frac{1}{8}-\frac{\tilde{\alpha}}{2}\right)\left\|\tilde{\boldsymbol{w}}_1 - \tilde{\boldsymbol{w}}_2\right\|^2 \\ =& \left(\frac{1}{8}-\frac{\tilde{\alpha}}{2}\right)\left(\left\|\tilde{\boldsymbol{w}}_1\right\|^2 + \left\|\tilde{\boldsymbol{w}}_2\right\|^2\right) - \frac{\left(\pi\left(\frac{1}{2}+2\tilde{\alpha}\right)-\theta\right)\cos\theta + \sin\theta}{2\pi}\left\|\tilde{\boldsymbol{w}}_1\right\|\left\|\tilde{\boldsymbol{w}}_2\right\|\end{aligned}$$

Taking the norm out: Note that we are only interested in the positivity of $\tilde{h}$, therefore dividing it by $\left\|\tilde{\boldsymbol{w}}_2\right\|^2$ does not change the sign. Denote $\frac{\left\|\tilde{\boldsymbol{w}}_1\right\|}{\left\|\tilde{\boldsymbol{w}}_2\right\|} = r$. Then,

$$\frac{\tilde{h}\left(\tilde{\boldsymbol{w}}_1,\tilde{\boldsymbol{w}}_2\right)}{\left\|\tilde{\boldsymbol{w}}_2\right\|^2} = \left(\frac{1}{8}-\frac{\tilde{\alpha}}{2}\right)\left(r^2+1\right) - \frac{\left(\pi\left(\frac{1}{2}+2\tilde{\alpha}\right)-\theta\right)\cos\theta + \sin\theta}{2\pi}r$$

We note that the expression is of the form $C_1\left(r^2+1\right)+C_2 r$. Note that the minima of this expression is achieved at $r=1$. Therefore, it is sufficient for us to check the positivity of the expression at $r=1$. To this aim we draw

$$\left.\frac{\tilde{h}\left(\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2\right)}{\left\|\tilde{\boldsymbol{w}}_2\right\|^2}\right|_{r=1}=\left(\frac{1}{4}-\tilde{\alpha}\right)-\frac{\left(\pi\left(\frac{1}{2}+2 \tilde{\alpha}\right)-\theta\right) \cos \theta+\sin \theta}{2 \pi}$$

To complete the proof, in Figure 5, we set $\tilde{\alpha}=0.05$ and draw $\left.\frac{\tilde{h}\left(\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2\right)}{\left\|\tilde{\boldsymbol{w}}_2\right\|^2}\right|_{r=1}$ as a 1D plot for

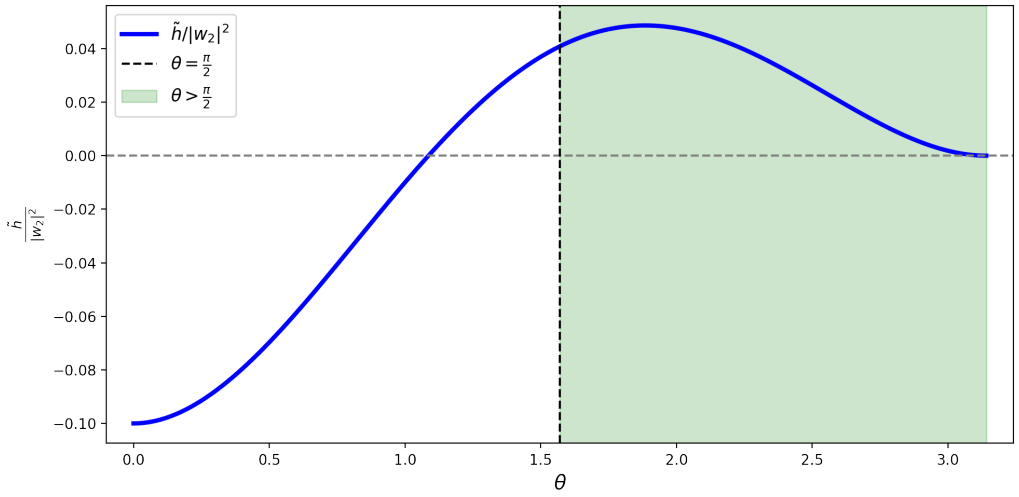


Figure 5: $\mathcal{L}(\boldsymbol{v}, \boldsymbol{W})-\tilde{\alpha}\left(\left\|v_1 \boldsymbol{w}_1-\boldsymbol{a}\right\|^2+\left\|v_2 \boldsymbol{w}_2+\boldsymbol{a}\right\|^2\right)$ **is non-negative**. We set $\tilde{\alpha}=0.05$ and draw $\frac{1}{\left\|\tilde{\boldsymbol{w}}_2\right\|^2} \tilde{h}\left(\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2\right)$ for $\theta \in[0, \pi]$. We show that $\frac{1}{\left\|\tilde{\boldsymbol{w}}_2\right\|^2} \tilde{h}\left(\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2\right)$ is non-negative inside the shaded region ($\theta \geq \frac{\pi}{2}$).

$\theta \in[0, \pi]$. The plot demonstrates that $\tilde{h}$ is non-negative for $\theta>\frac{\pi}{2}$. This finishes the proof of Lemma 13. ■

Appendix C. Proof of Key Lemmas in the Empirical Setting

C.1. Bound for Imbalance Term (Lemma 3)

By symmetry, it suffices to prove the bound for $b_1^{(\tau+1)} - b_1^{(\tau)}$. We first evaluate the per-step change in the imbalance term $b_1^{(\tau)}$. By the update rule of gradient descent, we have

$$\begin{aligned}
b_1^{(\tau+1)} =& \left\|\boldsymbol{w}_1^{(\tau+1)}\right\|^2 - \left(v_1^{(\tau+1)}\right)^2 \\
=& \left\|\boldsymbol{w}_1 - \mu\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 - \left(v_1 - \mu\nabla_{v_1}\widehat{\mathcal{L}}\right)^2 \\
=& \|\boldsymbol{w}_1\|^2 - 2\mu\boldsymbol{w}_1^T\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} + \mu^2\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 - \left(v_1^2 - 2\mu v_1\nabla_{v_1}\widehat{\mathcal{L}} + \mu^2\left(\nabla_{v_1}\widehat{\mathcal{L}}\right)^2\right) \\
\overset{(a)}{=}& \|\boldsymbol{w}_1\|^2 - v_1^2 + \mu^2\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 - \mu^2\left(\nabla_{v_1}\widehat{\mathcal{L}}\right)^2 \\
=& b_1^{(\tau)} + \mu^2\left(\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 - \left(\nabla_{v_1}\widehat{\mathcal{L}}\right)^2\right),
\end{aligned}$$

where (a) follows from Eq. 13. It follows that

$$\begin{aligned}
\left|b_1^{(\tau+1)} - b_1^{(\tau)}\right| =& \mu^2\left|\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 - \left(\nabla_{v_1}\widehat{\mathcal{L}}\right)^2\right| \\
\leq& \mu^2\left(\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 + \left(\nabla_{v_1}\widehat{\mathcal{L}}\right)^2\right). \qquad (17)
\end{aligned}$$

To bound the drift, we decompose the empirical gradients into their population counterparts and the associated estimation errors:

$$\begin{aligned}
&\left(\nabla_{v_1}\widehat{\mathcal{L}}\right)^2 + \left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 \\
=& \left(\nabla_{v_1}\mathcal{L} + \nabla_{v_1}\widehat{\mathcal{L}} - \nabla_{v_1}\mathcal{L}\right)^2 + \left\|\nabla_{\boldsymbol{w}_1}\mathcal{L} + \nabla_{w_1}\widehat{\mathcal{L}} - \nabla_{w_1}\mathcal{L}\right\|^2 \\
\leq& 2\left(\nabla_{v_1}\mathcal{L}\right)^2 + 2\left(\nabla_{v_1}\widehat{\mathcal{L}} - \nabla_{v_1}\mathcal{L}\right)^2 + 2\left\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2 + 2\left\|\nabla_{w_1}\widehat{\mathcal{L}} - \nabla_{w_1}\mathcal{L}\right\|^2 \\
=& 2\left(\left(\nabla_{v_1}\mathcal{L}\right)^2 + \left\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2\right) + 2\left(\nabla_{v_1}\widehat{\mathcal{L}} - \nabla_{v_1}\mathcal{L}\right)^2 + 2\left\|\nabla_{w_1}\widehat{\mathcal{L}} - \nabla_{w_1}\mathcal{L}\right\|^2 \\
=& 2\left(\left(\nabla_{v_1}\mathcal{L}\right)^2 + \left\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2\right) + 2\left(\frac{\boldsymbol{w}_1^T}{v_1}\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \frac{\boldsymbol{w}_1^T}{v_1}\nabla_{\boldsymbol{w}_1}\mathcal{L}\right)^2 + 2\left\|\nabla_{w_1}\widehat{\mathcal{L}} - \nabla_{w_1}\mathcal{L}\right\|^2 \\
\leq& 2\left(\left(\nabla_{v_1}\mathcal{L}\right)^2 + \left\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2\right) + 2\frac{\|\boldsymbol{w}_1\|^2\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2}{v_1^2} + 2\left\|\nabla_{w_1}\widehat{\mathcal{L}} - \nabla_{w_1}\mathcal{L}\right\|^2 \\
=& 2\left(\left(\nabla_{v_1}\mathcal{L}\right)^2 + \left\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2\right) + 2\left(1 + \frac{\|\boldsymbol{w}_1\|^2}{v_1^2}\right)\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2. \qquad (18)
\end{aligned}$$

Regarding the population component, note that by considering a reparameterized set of weights $\tilde{v}_1 = \tilde{v}_2 = 1, \tilde{\boldsymbol{w}}_1 = v_1\boldsymbol{w}_1, \tilde{\boldsymbol{w}}_2 = v_2\boldsymbol{w}_2$, we can leverage the smoothness properties of the population loss (Lemma 12):

$$\|\nabla_{\tilde{\boldsymbol{w}}_1}\mathcal{L}\|^2 + \|\nabla_{\tilde{\boldsymbol{w}}_2}\mathcal{L}\|^2 \leq \frac{5}{2}\left(\|\tilde{\boldsymbol{w}}_1 - \boldsymbol{a}\|^2 + \|\tilde{\boldsymbol{w}}_2 + \boldsymbol{a}\|^2\right).$$

Given $v_1, v_2 > 0$, it holds that $\nabla_{\tilde{\boldsymbol{w}}_1}\mathcal{L} = \frac{1}{v_1}\nabla_{\boldsymbol{w}_1}\mathcal{L}$. It then follows that:

$$\begin{aligned}
(\nabla_{v_1}\mathcal{L})^2 + \|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2 =& \left(\frac{\boldsymbol{w}_1^T}{v_1}\nabla_{\boldsymbol{w}_1}\mathcal{L}\right)^2 + \|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2 \\
\leq& \frac{\|\boldsymbol{w}_1\|^2 \|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2}{v_1^2} + \|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2 \\
=& \left(\|\boldsymbol{w}_1\|^2 + v_1^2\right)\|\nabla_{\tilde{\boldsymbol{w}}_1}\mathcal{L}\|^2 \\
\leq& \left(\|\boldsymbol{w}_1\|^2 + v_1^2\right)\cdot\frac{5}{2}\left(\|\tilde{\boldsymbol{w}}_1 - \boldsymbol{a}\|^2 + \|\tilde{\boldsymbol{w}}_2 + \boldsymbol{a}\|^2\right) \\
=& \frac{5}{2}\left(\|\boldsymbol{w}_1\|^2 + v_1^2\right)\left(\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\|^2 + \|v_2\boldsymbol{w}_2 + \boldsymbol{a}\|^2\right). \qquad (19)
\end{aligned}$$

As for the second term, Lemma 11 provides the following concentration bound:

$$\left\|\nabla_{w_1}\widehat{\mathcal{L}} - \nabla_{w_1}\mathcal{L}\right\| \leq v_1\delta\left(\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\| + \|v_2\boldsymbol{w}_2 + \boldsymbol{a}\|\right),$$

where $\delta \leq \frac{1}{2}$ is a constant. Squaring both sides and applying Jensen's inequality leads to:

$$\begin{aligned}
\left\|\nabla_{w_1}\widehat{\mathcal{L}} - \nabla_{w_1}\mathcal{L}\right\|^2 \leq& v_1^2\delta^2\left(\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\| + \|v_2\boldsymbol{w}_2 + \boldsymbol{a}\|\right)^2 \\
\leq& 2v_1^2\delta^2\left(\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\|^2 + \|v_2\boldsymbol{w}_2 + \boldsymbol{a}\|^2\right). \qquad (20)
\end{aligned}$$

Substituting the bounds from (19) and (20) into (18), we obtain that

$$\begin{aligned}
&\left(\nabla_{v_1}\widehat{\mathcal{L}}\right)^2 + \left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 \\
\leq& 2\left((\nabla_{v_1}\mathcal{L})^2 + \|\nabla_{\boldsymbol{w}_1}\mathcal{L}\|^2\right) + 2\left(1 + \frac{\|\boldsymbol{w}_1\|^2}{v_1^2}\right)\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L}\right\|^2 \\
\leq& 2\cdot\frac{5}{2}\left(\|\boldsymbol{w}_1\|^2 + v_1^2\right)\left(\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\|^2 + \|v_2\boldsymbol{w}_2 + \boldsymbol{a}\|^2\right) \\
&+ 2\left(1 + \frac{\|\boldsymbol{w}_1\|^2}{v_1^2}\right)\cdot 2v_1^2\delta^2\left(\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\|^2 + \|v_2\boldsymbol{w}_2 + \boldsymbol{a}\|^2\right) \\
=& \left(5 + 4\delta^2\right)\left(\|\boldsymbol{w}_1\|^2 + v_1^2\right)\left(\|v_1\boldsymbol{w}_1 - \boldsymbol{a}\|^2 + \|v_2\boldsymbol{w}_2 + \boldsymbol{a}\|^2\right).
\end{aligned}$$

Putting the above inequality into (17), we complete the proof of the lemma with the constant $c_6 = 6$.

C.2. Proof of the Stability of Angles (Lemma 4)

We prove the lemma with the following constants: $c_0 \leq \frac{1}{2}, c_2 = 2, c_5 = \frac{1}{50}, c_4 = \frac{\pi}{20}, \gamma = \frac{1}{4}$.

By symmetry, it suffices to prove the bound for $\theta_1^{(\tau+1)}$. By the update rule of gradient descent, we have

$$\begin{aligned}
&\boldsymbol{w}_1^{(\tau+1)} \\
=&\boldsymbol{w}_1 - \mu \nabla_{\boldsymbol{w}_1} \widehat{\mathcal{L}} \\
=&\boldsymbol{w}_1 - \mu \cdot \frac{v_1}{2} \left(v_1 \boldsymbol{w}_1 - \left(1 - \frac{\theta}{\pi}\right) v_2 \boldsymbol{w}_2 - \frac{\sin\theta}{\pi} v_2 \left\|\boldsymbol{w}_2\right\| \overline{\boldsymbol{w}}_1 - \boldsymbol{a} + \Delta\mathcal{G}_1 \right) \\
=&\left(\left(1 - \frac{\mu v_1^2}{2}\right) \left\|\boldsymbol{w}_1\right\| + \frac{\mu v_1}{2} \cdot \frac{\sin\theta}{\pi} v_2 \left\|\boldsymbol{w}_2\right\| \right) \overline{\boldsymbol{w}}_1 + \frac{\mu v_1}{2} \left(\left(1 - \frac{\theta}{\pi}\right) v_2 \boldsymbol{w}_2 + \boldsymbol{a} - \Delta\mathcal{G}_1 \right).
\end{aligned}$$

Given that $v_1, v_2 > 0$ and $1 - \frac{\mu v_1^2}{2} \geq 1 - \frac{c_0 c_2^2}{2} \geq 0$, the update vector $\boldsymbol{w}_1 - \mu \nabla_{\boldsymbol{w}_1} \widehat{\mathcal{L}}$ is a nonnegative linear combination of $\overline{\boldsymbol{w}}_1$ and the vector

$$\boldsymbol{q} := \left(1 - \frac{\theta}{\pi}\right) v_2 \boldsymbol{w}_2 + \boldsymbol{a} - \Delta\mathcal{G}_1.$$

Recall that the angle of a positive linear combination with a reference vector $\boldsymbol{a}$ is bounded by the maximum angle of its components. Since we assume $\angle(\overline{\boldsymbol{w}}_1, \boldsymbol{a}) \leq c_4$, it suffices to show that $\angle(\boldsymbol{q}, \boldsymbol{a}) \leq c_4$ to conclude the proof.

Geometrically, to ensure this angle constraint on $\boldsymbol{q}$, we need to bound the perturbation magnitude $\|\Delta\mathcal{G}_1\|$ by the Euclidean distance to the cone boundary. Specifically, under the assumption that $0 < v_2 \leq c_2 \sqrt{\|\boldsymbol{a}\|}$ and $\left|b_2^{(\tau)}\right| \leq \gamma \|\boldsymbol{a}\|$, it follows that $\|v_2 \boldsymbol{w}_2\| \leq c_2 \sqrt{c_2^2 + \gamma}\, \|\boldsymbol{a}\|$. This implies that the condition:

$$\|\Delta\mathcal{G}_1\| \leq c_5 \|\boldsymbol{a}\| \leq \left(\frac{\|\boldsymbol{a}\|}{2\cos c_4} - \frac{2 c_4 c_2 \sqrt{c_2^2 + \gamma}}{\pi} \|\boldsymbol{a}\| \right) \sin(2c_4)$$

is sufficient to guarantee $\angle(\boldsymbol{q}, \boldsymbol{a}) \leq c_4$. Numerical verification confirms that the chosen constants $c_2 = 2$, $c_4 = \frac{\pi}{20}$, $c_5 = \frac{1}{50}$, and $\gamma = \frac{1}{4}$ satisfy the required inequality. Consequently, we have $\angle(\boldsymbol{w}_1^{(\tau+1)}, \boldsymbol{a}) \leq c_4$, which completes the proof.

C.3. Proof of the Stability of Norms (Lemma 5)

We prove the lemma with the following constants: $c_2 = 2, c_0 \leq \frac{4}{25}, c_4 \leq \frac{\pi}{10}, c_5 \leq \frac{1}{3}, \gamma \leq \frac{1}{2}$.

By symmetry, it suffices to prove the bound for $v_1^{(\tau+1)}$. We first show that $v_1^{(\tau+1)} \leq 2\sqrt{\|\boldsymbol{a}\|}$. Applying the gradient descent update rule, the partial derivative with respect to v_1 is bounded as

follows:

$$\begin{aligned}
\nabla_{v_1}\widehat{\mathcal{L}} &= \frac{1}{2}\left(v_1\left\|\boldsymbol{w}_1\right\|^2 - \frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}v_2\left\|\boldsymbol{w}_1\right\|\left\|\boldsymbol{w}_2\right\| - \boldsymbol{w}_1^T\boldsymbol{a} + \boldsymbol{w}_1^T\Delta\mathcal{G}_1\right)\\
&\geq \frac{1}{2}\left(v_1\left\|\boldsymbol{w}_1\right\|^2 - \frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}v_2\left\|\boldsymbol{w}_1\right\|\left\|\boldsymbol{w}_2\right\| - \left\|\boldsymbol{w}_1\right\|\left\|\boldsymbol{a}\right\| - \left\|\boldsymbol{w}_1\right\|\left\|\Delta\mathcal{G}_1\right\|\right)\\
&= \frac{\left\|\boldsymbol{w}_1\right\|}{2}\left(v_1\left\|\boldsymbol{w}_1\right\| - \frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}v_2\left\|\boldsymbol{w}_2\right\| - \left\|\boldsymbol{a}\right\| - \left\|\Delta\mathcal{G}_1\right\|\right)\\
&\geq \frac{\left\|\boldsymbol{w}_1\right\|}{2}\left(v_1\left\|\boldsymbol{w}_1\right\| - \frac{1}{3}\left\|\boldsymbol{a}\right\| - \left\|\boldsymbol{a}\right\| - \frac{1}{3}\left\|\boldsymbol{a}\right\|\right)\\
&= \frac{\left\|\boldsymbol{w}_1\right\|}{2}\left(v_1\left\|\boldsymbol{w}_1\right\| - \frac{5}{3}\left\|\boldsymbol{a}\right\|\right).
\end{aligned}$$

In the penultimate line, we use the assumption $\left\|\Delta\mathcal{G}_1\right\| \leq c_5\left\|\boldsymbol{a}\right\| \leq \frac{1}{3}\left\|\boldsymbol{a}\right\|$ and the fact that

$$\frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}v_2\left\|\boldsymbol{w}_2\right\| \leq \frac{(2c_4)\cos(\pi-2c_4)+\sin(\pi-2c_4))}{\pi}c_2\sqrt{c_2^2+\gamma}\left\|\boldsymbol{a}\right\| \leq \frac{1}{3}\left\|\boldsymbol{a}\right\|.$$

To establish the bound, we consider the following two cases based on the magnitude of v_1:

- **Case 1:** $v_1 \geq \frac{5}{3}\sqrt{\left\|\boldsymbol{a}\right\|}$.
  Since $\left|\left\|\boldsymbol{w}_1\right\|^2 - v_1^2\right| = \left|b_1\right| \leq \gamma\left\|\boldsymbol{a}\right\| \leq \left\|\boldsymbol{a}\right\|$, we have $\left\|\boldsymbol{w}_1\right\| \geq \sqrt{\left\|\boldsymbol{a}\right\|}$. Thus, we have

  $$\begin{aligned}
  \nabla_{v_1}\widehat{\mathcal{L}} &\geq \frac{\left\|\boldsymbol{w}_1\right\|}{2}\left(v_1\left\|\boldsymbol{w}_1\right\| - \frac{5}{3}\left\|\boldsymbol{a}\right\|\right)\\
  &\geq \frac{\left\|\boldsymbol{w}_1\right\|}{2}\left(\frac{5}{3}\sqrt{\left\|\boldsymbol{a}\right\|}\cdot\sqrt{\left\|\boldsymbol{a}\right\|} - \frac{5}{3}\left\|\boldsymbol{a}\right\|\right)\\
  &\geq 0,
  \end{aligned}$$

  which means $v_1^{(\tau+1)} = v_1 - \mu\nabla_{v_1}\widehat{\mathcal{L}} \leq v_1 \leq 2\sqrt{\left\|\boldsymbol{a}\right\|}$.

- **Case 2:** $v_1 < \frac{5}{3}\sqrt{\left\|\boldsymbol{a}\right\|}$.
  We have

  $$\begin{aligned}
  \nabla_{v_1}\widehat{\mathcal{L}} &\geq \frac{\left\|\boldsymbol{w}_1\right\|}{2}\left(v_1\left\|\boldsymbol{w}_1\right\| - \frac{5}{3}\left\|\boldsymbol{a}\right\|\right)\\
  &\geq -\frac{5}{6}\left\|\boldsymbol{w}_1\right\|\left\|\boldsymbol{a}\right\|,
  \end{aligned}$$

  which means

  $$\begin{aligned}
  v_1^{(\tau+1)} =& v_1 - \mu\nabla_{v_1}\widehat{\mathcal{L}}\\
  \leq& \frac{5}{3}\sqrt{\left\|\boldsymbol{a}\right\|} + \frac{5}{6}\mu\left\|\boldsymbol{w}_1\right\|\left\|\boldsymbol{a}\right\|\\
  \leq& 2\sqrt{\left\|\boldsymbol{a}\right\|}.
  \end{aligned}$$

  Here we use the fact that $\mu\left\|\boldsymbol{w}_1\right\|\sqrt{\left\|\boldsymbol{a}\right\|} \leq c_0\sqrt{c_2^2+\gamma} \leq \frac{4}{25}\cdot\sqrt{5} \leq \frac{2}{5}$.

Next, we establish the lower bound $v_1^{(\tau+1)} > \beta\sqrt{\|\boldsymbol{a}\|}$. By the update rule of gradient descent, we have

$$\begin{aligned}
\nabla_{v_1}\widehat{\mathcal{L}} =& \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}v_2\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\| - \boldsymbol{w}_1^T\boldsymbol{a} + \boldsymbol{w}_1^T\Delta\mathcal{G}_1\right)\\
\leq& \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \boldsymbol{w}_1^T\boldsymbol{a} + \|\boldsymbol{w}_1\|\|\Delta\mathcal{G}_1\|\right)\\
=& \frac{\|\boldsymbol{w}_1\|}{2}\left(v_1\|\boldsymbol{w}_1\| - \cos\theta_1\|\boldsymbol{a}\| + \|\Delta\mathcal{G}_1\|\right)\\
\leq& \frac{\|\boldsymbol{w}_1\|}{2}\left(v_1\|\boldsymbol{w}_1\| - \frac{5}{6}\|\boldsymbol{a}\| + \frac{1}{3}\|\boldsymbol{a}\|\right)\\
=& \frac{\|\boldsymbol{w}_1\|}{2}\left(v_1\|\boldsymbol{w}_1\| - \frac{1}{2}\|\boldsymbol{a}\|\right).
\end{aligned}$$

In the penultimate line, we use the assumption $\|\Delta\mathcal{G}_1\| \leq c_5\|\boldsymbol{a}\| \leq \frac{1}{3}\|\boldsymbol{a}\|$ and $\theta_1 \leq c_4 \leq \frac{\pi}{10} \leq \arccos\left(\frac{5}{6}\right)$.

To establish the bound, we consider the following two cases based on the magnitude of $v_1\|\boldsymbol{w}_1\|$:

- **Case 1:** $v_1\|\boldsymbol{w}_1\| \leq \frac{1}{2}\|\boldsymbol{a}\|$.
  we have
  $$\begin{aligned}
  \nabla_{v_1}\widehat{\mathcal{L}} \leq& \frac{\|\boldsymbol{w}_1\|}{2}\left(v_1\|\boldsymbol{w}_1\| - \frac{1}{2}\|\boldsymbol{a}\|\right)\\
  \leq& 0,
  \end{aligned}$$
  which means $v_1^{(\tau+1)} = v_1 - \mu\nabla_{v_1}\widehat{\mathcal{L}} \geq \beta\sqrt{\|\boldsymbol{a}\|}$.

- **Case 2:** $v_1\|\boldsymbol{w}_1\| > \frac{1}{2}\|\boldsymbol{a}\|$.
  Since $\left|\|\boldsymbol{w}_1\|^2 - v_1^2\right| = |b_1| \leq \gamma\|\boldsymbol{a}\| \leq \frac{1}{2}\|\boldsymbol{a}\|$, we have $v_1 > \frac{1}{2}\sqrt{\|\boldsymbol{a}\|}$ and $\|\boldsymbol{w}_1\| < \sqrt{\|\boldsymbol{a}\|}$. Thus, we have
  $$\begin{aligned}
  \nabla_{v_1}\widehat{\mathcal{L}} \leq& \frac{\|\boldsymbol{w}_1\|}{2}\left(v_1\|\boldsymbol{w}_1\| - \frac{1}{2}\|\boldsymbol{a}\|\right)\\
  \leq& \frac{v_1\|\boldsymbol{w}_1\|^2}{2}\\
  <& \frac{v_1\|\boldsymbol{a}\|}{2}.
  \end{aligned}$$
  It follows that $v_1^{(\tau+1)} = v_1 - \mu\nabla_{v_1}\widehat{\mathcal{L}} \geq v_1 - \frac{\mu v_1\|\boldsymbol{a}\|}{2} \geq \frac{1}{2}v_1 > \frac{1}{4}\sqrt{\|\boldsymbol{a}\|} \geq \beta\sqrt{\|\boldsymbol{a}\|}$. Here we use the assumption $\mu \leq \frac{c_0}{\|\boldsymbol{a}\|} \leq \frac{1}{\|\boldsymbol{a}\|}$.

This concludes the proof of Lemma 5.

C.4. Proof of Phase 1 (Lemma 6)

We prove the lemma with the following constants: $c_0 \leq 1, c_4 \leq \frac{\pi}{4}, c_9 = \frac{64}{\tan c_4}, c_2 = 2, c_3 = \frac{1}{4}, c_{10} = \frac{1}{4e^{2\alpha}}, \sigma_0 = \frac{1}{8e^{2\alpha}}$, where $\alpha = \frac{65}{\tan c_4}$.

For notational simplicity, we assume without loss of generality that $\|\boldsymbol{a}\| = 1$. Recall that our initialization scheme is given by

$$\boldsymbol{w}_1^{(0)}, \boldsymbol{w}_2^{(0)} \sim \mathcal{N}\Big(0, \tfrac{\sigma^2}{d}\boldsymbol{I}_d\Big), \qquad v_1^{(0)}, v_2^{(0)} \sim \frac{\sigma}{\sqrt{d}}\xi, \quad \xi^2 \sim \chi_d^2,$$

where $\sigma \le \sigma_0\sqrt{\|\boldsymbol{a}\|}$ and χ_d^2 denotes the chi-squared distribution with d degrees of freedom.

By definition, both the squared scalar $\left(v_i^{(0)}\right)^2$ and the squared vector norm $\left\|\boldsymbol{w}_i^{(0)}\right\|^2$ follow the same Chi-squared distribution. By standard concentration inequalities, both variables concentrate sharply around the value σ^2 with probability at least $1 - O(e^{-cd})$. Furthermore, the projection $\boldsymbol{a}^T\boldsymbol{w}_i^{(0)}$ follows a Gaussian distribution $\mathcal{N}(0, \frac{\sigma^2}{d}\|\boldsymbol{a}\|^2)$, which concentrates around 0 with magnitude $O(1/\sqrt{d})$.

Specifically, applying the Laurent-Massart concentration bounds for the Chi-squared distribution and standard Gaussian tail bounds, we have that with probability at least $1-8\exp\left(-\min\left(\frac{1}{16}, \frac{\hat{c}_4^2}{4}\right)d\right)$, the following inequalities hold simultaneously:

$$\begin{aligned}
&v_1^{(0)} \ge \frac{1}{2}\left\|\boldsymbol{w}_1^{(0)}\right\|, v_2^{(0)} \ge \frac{1}{2}\left\|\boldsymbol{w}_2^{(0)}\right\| \\
&\frac{1}{2}\sigma \le v_1^{(0)}, v_2^{(0)} \le 2\sigma \\
&\boldsymbol{a}^T\boldsymbol{w}_1^{(0)} \ge -\hat{c}_4 v_1^{(0)}, \boldsymbol{a}^T\boldsymbol{w}_2^{(0)} \ge -\hat{c}_4 v_2^{(0)}.
\end{aligned}$$

where $\hat{c}_4$ is a fixed positive constant.

Assume that $c_0 \le 1, c_4 \le \frac{\pi}{4}$ and $T_1 = \left\lceil \frac{64}{\mu \tan c_4} \right\rceil$. We aim to establish the following properties for all iterations $\tau \le T$ via induction:

$$v_1^{(\tau)} \ge \frac{1}{2}v_1^{(0)}, v_2^{(\tau)} \ge \frac{1}{2}v_2^{(0)} \tag{21}$$

$$\boldsymbol{a}^T\boldsymbol{w}_1^{(\tau)} \ge -\hat{c}_4 v_1^{(0)}, \boldsymbol{a}^T\boldsymbol{w}_2^{(\tau)} \ge -\hat{c}_4 v_2^{(0)} \tag{22}$$

$$v_1^{(\tau)}, \left\|\boldsymbol{w}_1^{(\tau)}\right\| \le 2(1+\mu)^\tau v_1^{(0)} \le \hat{c}_3 v_1^{(0)} \le \frac{1}{\hat{c}_5} \tag{23}$$

$$v_2^{(\tau)}, \left\|\boldsymbol{w}_2^{(\tau)}\right\| \le 2(1+\mu)^\tau v_2^{(0)} \le \hat{c}_3 v_2^{(0)} \le \frac{1}{\hat{c}_5}$$

with constant $\hat{c}_4 = \frac{1}{2\alpha}, \hat{c}_3 = \hat{c}_5 = 2e^\alpha, \sigma_0 = \frac{1}{8e^{2\alpha}}$ where $\alpha = \frac{65}{\tan c_4}$.

We prove (21), (22), (23) by induction. At initialization it is true with probability at least $1 - \tilde{C}e^{-\tilde{c}d}$ as explained above. Assuming these hypotheses hold for some $\tau < T$, we proceed to show they remain valid for iteration $\tau + 1$. By symmetry, we focus on v_1 and $\boldsymbol{w}_1$. For the sake of notation simplicity, we suppress the superscript (τ) where the context is clear.

We start with bound for $\boldsymbol{a}^T\boldsymbol{w}_1$. Specifically, the update rule for the alignment term yields

$$\begin{aligned}
&\boldsymbol{a}^T\boldsymbol{w}_1^{(\tau+1)} \\
=&\boldsymbol{a}^T\boldsymbol{w}_1 - \mu\boldsymbol{a}^T\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} \\
=&\boldsymbol{a}^T\boldsymbol{w}_1 + \frac{\mu}{2}v_1 - \frac{\mu v_1}{2}\left(v_1\boldsymbol{a}^T\boldsymbol{w}_1 - \left(1 - \frac{\theta}{\pi}\right)v_2\boldsymbol{a}^T\boldsymbol{w}_2 - \frac{\sin\theta}{\pi}v_2\|\boldsymbol{w}_2\|\,\boldsymbol{a}^T\overline{\boldsymbol{w}}_1 + \boldsymbol{a}^T\Delta\mathcal{G}_1\right).
\end{aligned}$$

By Lemma 11, with probability at least $1 - 3e^{-cd}$, we have $\|\mathcal{G}_1\| \leq \frac{1}{\hat{c}_5^2} \|\boldsymbol{a}\|$. It follows that

$$\begin{aligned}
&\left| v_1 \boldsymbol{a}^T \boldsymbol{w}_1 - \left(1 - \frac{\theta}{\pi}\right) v_2 \boldsymbol{a}^T \boldsymbol{w}_2 - \frac{\sin\theta}{\pi} v_2 \|\boldsymbol{w}_2\| \boldsymbol{a}^T \overline{\boldsymbol{w}}_1 + \boldsymbol{a}^T \Delta\mathcal{G}_1 \right| \\
\leq & v_1 \|\boldsymbol{w}_1\| + v_2 \|\boldsymbol{w}_2\| + v_2 \|\boldsymbol{w}_2\| + \left\|\boldsymbol{a}^T \Delta\mathcal{G}_1\right\| \\
\leq & \frac{4}{\hat{c}_5^2} \leq \frac{1}{2}.
\end{aligned}$$

Combining the above bounds, the term inside the parenthesis is bounded by $\frac{1}{2}$ in absolute value. Consequently, the update satisfies

$$\boldsymbol{a}^T \boldsymbol{w}_1 + \frac{\mu}{4} v_1 \leq \boldsymbol{a}^T \boldsymbol{w}_1^{(\tau+1)} \leq \boldsymbol{a}^T \boldsymbol{w}_1 + \frac{3\mu}{4} v_1. \tag{24}$$

It follows that $\boldsymbol{a}^T \boldsymbol{w}_1^{(\tau+1)} \geq \boldsymbol{a}^T \boldsymbol{w}_1$, By the inductive hypothesis $\boldsymbol{a}^T \boldsymbol{w}_1 \geq -\hat{c}_4 v_1^{(0)}$, we conclude that (22) holds for iteration $\tau + 1$.

Next, we bound the orthogonal component $\left\|(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1^{(\tau+1)}\right\|$. Observe that:

$$\begin{aligned}
&(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1^{(\tau+1)} \\
=&(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1 - \mu(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\nabla_{\boldsymbol{w}_1} Loss \\
=&(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1 - \mu\frac{v_1}{2}(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\left(v_1 \boldsymbol{w}_1 - \left(1 - \frac{\theta}{\pi}\right) v_2 \boldsymbol{w}_2 - \frac{\sin\theta}{\pi} v_2 \|\boldsymbol{w}_2\| \overline{\boldsymbol{w}}_1 - \boldsymbol{a} + \Delta\mathcal{G}_1\right).
\end{aligned}$$

Since $\left(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T\right)\boldsymbol{a} = 0$, the update simplifies to:

$$\begin{aligned}
&(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1^{(\tau+1)} \\
=&(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1 - \mu\frac{v_1}{2}(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\left(v_1 \boldsymbol{w}_1 - \left(1 - \frac{\theta}{\pi}\right) v_2 \boldsymbol{w}_2 - \frac{\sin\theta}{\pi} v_2 \|\boldsymbol{w}_2\| \overline{\boldsymbol{w}}_1 + \Delta\mathcal{G}_1\right).
\end{aligned}$$

Applying the triangle inequality and substituting the bounds for $\|\boldsymbol{w}_1\|$, $\|\boldsymbol{w}_2\|$, and $\|\Delta\mathcal{G}_1\|$, we have

$$\begin{aligned}
&\left\|(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1^{(\tau+1)}\right\| \\
\leq & \left\|(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1\right\| + \mu\left\|\frac{v_1}{2}(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\left(v_1 \boldsymbol{w}_1 - \left(1 - \frac{\theta}{\pi}\right) v_2 \boldsymbol{w}_2 - \frac{\sin\theta}{\pi} v_2 \|\boldsymbol{w}_2\| \overline{\boldsymbol{w}}_1 + \Delta\mathcal{G}_1\right)\right\| \\
\leq & \left\|(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1\right\| + \frac{\mu v_1}{2}(v_1 \|\boldsymbol{w}_1\| + 2v_2 \|\boldsymbol{w}_2\| + \|\Delta\mathcal{G}_1\|) \\
\leq & \left\|(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1\right\| + \frac{2\mu v_1}{\hat{c}_5^2}.
\end{aligned} \tag{25}$$

Combining the upper bound for $\boldsymbol{a}^T\boldsymbol{w}_1^{(\tau+1)}$ (from inequality (24)) and the bound in (25), we derive the upper bound for $\left\|\boldsymbol{w}_1^{(\tau+1)}\right\|$ as follows:

$$\begin{aligned}\left\|\boldsymbol{w}_1^{(\tau+1)}\right\|^2 &= \left(\boldsymbol{a}^T\boldsymbol{w}_1^{(\tau+1)}\right)^2 + \left\|(\boldsymbol{I}-\boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1^{(\tau+1)}\right\|^2 \\ &\le \left(\boldsymbol{a}^T\boldsymbol{w}_1 + \frac{3\mu}{4}v_1\right)^2 + \left(\left\|(\boldsymbol{I}-\boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1\right\| + \frac{2}{c_5^2}\mu v_1^{(0)}\right)^2 \\ &\le \|\boldsymbol{w}_1\|^2 + 2\left(\frac{3\mu}{4}v_1 + \frac{2}{c_5^2}\mu v_1^{(0)}\right)\|\boldsymbol{w}_1\| + \left(\frac{3\mu}{4}v_1\right)^2 + \left(\frac{2}{c_5^2}\mu v_1\right)^2 \\ &\le \left(\|\boldsymbol{w}_1\| + \frac{3\mu}{4}v_1 + \frac{2}{c_5^2}\mu v_1^{(0)}\right)^2,\end{aligned}$$

Taking the square root and using the inductive hypothesis $v_1^{(\tau)}$, $\left\|\boldsymbol{w}_1^{(\tau)}\right\| \le 2(1+\mu)^\tau v_1^{(0)}$, along with the fact that $\frac{2}{\hat{c}_5^2} \le \frac{1}{4}$, we conclude $\left\|\boldsymbol{w}_1^{(\tau+1)}\right\| \le 2(1+\mu)^{\tau+1}v_1^{(0)}$, which establishes the upper bound for $\|\boldsymbol{w}_1\|$ in (23) for iteration $\tau+1$.

We now turn to the evolution of v_1. the update for v_1 is given by:

$$\begin{aligned}&v_1^{(\tau+1)} \\ =&v_1 - \mu\nabla_{v_1}Loss \\ =&v_1 - \mu\frac{\boldsymbol{w}_1^T}{v_1}\nabla_{\boldsymbol{w}_1}Loss \\ =&v_1 - \frac{\mu}{2}\boldsymbol{w}_1^T\left(v_1\boldsymbol{w}_1 - \left(1-\frac{\theta}{\pi}\right)v_2\boldsymbol{w}_2 - \frac{\sin\theta}{\pi}v_2\|\boldsymbol{w}_2\|\overline{\boldsymbol{w}}_1 - \boldsymbol{a} + \Delta\mathcal{G}_1\right) \\ =&v_1 + \frac{\mu}{2}\boldsymbol{a}^T\boldsymbol{w}_1 - \frac{\mu}{2}\boldsymbol{w}_1^T\boldsymbol{w}_1v_1 - \frac{\mu}{2}\boldsymbol{w}_1^T\Delta\mathcal{G}_1 + \frac{\mu}{2}v_2\left(\left(1-\frac{\theta}{\pi}\right)\boldsymbol{w}_1^T\boldsymbol{w}_2 + \frac{\sin\theta}{\pi}\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|\right).\end{aligned}$$

By triangle inequalities, we have

$$\begin{aligned}&\left|-\frac{\mu}{2}\boldsymbol{w}_1^T\boldsymbol{w}_1v_1 - \frac{\mu}{2}\boldsymbol{w}_1^T\Delta\mathcal{G}_1 + \frac{\mu}{2}v_2\left(\left(1-\frac{\theta}{\pi}\right)\boldsymbol{w}_1^T\boldsymbol{w}_2 + \frac{\sin\theta}{\pi}\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|\right)\right| \\ \le&\frac{\mu}{2}\|\boldsymbol{w}_1\|^2v_1 + \frac{\mu}{2}\|\boldsymbol{w}_1\|\|\mathcal{G}_1\| + \frac{\mu}{2}v_2\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\| + \frac{\mu}{2}v_2\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\| \\ \le&\frac{\mu}{2}\cdot 4\frac{\hat{c}_3}{\hat{c}_5^2}v_1^{(0)}.\end{aligned}$$

For the upper bound, we have

$$\begin{aligned}v_1^{(\tau+1)} &\le v_1 + \frac{\mu}{2}\boldsymbol{a}^T\boldsymbol{w}_1 + \frac{\mu}{2}\cdot 4\frac{\hat{c}_3}{\hat{c}_5^2}v_1^{(0)} \\ &\le v_1 + \frac{\mu}{2}\boldsymbol{a}^T\boldsymbol{w}_1 + \frac{\mu}{2}v_1^{(0)} \\ &\le \left(1+\frac{\mu}{2}+\frac{\mu}{2}\right)\cdot 2(1+\mu)^\tau v_1^{(0)} \\ &= 2(1+\mu)^{\tau+1}v_1^{(0)},\end{aligned}$$

which shows the bound of v_1 in (23) for iteration $\tau+1$.

For the lower bound, we have

$$\begin{aligned} v_1^{(\tau+1)} &\geq v_1 + \frac{\mu}{2}\left(-\hat{c}_4 v_1^{(0)}\right) - \frac{\mu}{2}\cdot 4\frac{\hat{c}_3}{\hat{c}_5^2}v_1^{(0)} \\ &\geq v_1 - \frac{1}{2T_1}v_1^{(0)}. \end{aligned}$$

This implies that v_1 decrease at most one half in the first T_1 iterations, which shows (21) for iteration $\tau+1$. Here we use the fact that $\hat{c}_4 + 4\frac{\hat{c}_3}{\hat{c}_5^2} \leq \frac{1}{\mu T_1} \leq 1$.

Resuming the proof of Lemma 6. Inequalities (21) and (23) directly imply that at iteration T_1:

$$v_1^{(T_1)} \geq \frac{1}{2}v_1^{(0)} \geq \frac{1}{4}\sigma,$$

and

$$v_1^{(T_1)} \leq \frac{1}{\hat{c}_5} \leq 2\sqrt{\|\boldsymbol{a}\|}.$$

Additionally, (23) yields an upper bound on the norm of the imbalance term $\left|b_1^{(T_1)}\right|$:

$$\left|b_1^{(T_1)}\right| = \left|\left(v_1^{(T_1)}\right)^2 - \left\|\boldsymbol{w}_1^{(T_1)}\right\|^2\right| \leq \frac{1}{\hat{c}_5^2}$$

Next, we estimate the alignment angle $\theta_1^{(T_1)}$. Summing the updates in (24) and (25) over T_1 iterations, we have

$$\begin{aligned} \boldsymbol{a}^T\boldsymbol{w}_1^{(T_1)} &\geq \boldsymbol{a}^T\boldsymbol{w}_1^{(0)} + \frac{\mu}{8}v_1^{(0)}\cdot T_1 \\ &\geq -\hat{c}_4 v_1^{(0)} + \frac{\mu}{8}T_1 v_1^{(0)} \end{aligned}$$

and

$$\begin{aligned} \left\|(\boldsymbol{I}-\boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1^{(T)}\right\| &\leq \left\|(\boldsymbol{I}-\boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1^{(0)}\right\| + \frac{2c_3}{c_5^2}\mu v_1^{(0)}\cdot T_1 \\ &\leq 2v_1^{(0)} + \frac{2\hat{c}_3}{\hat{c}_5^2}\mu T_1 v_1^{(0)}. \end{aligned}$$

It follows that $\boldsymbol{a}^T\boldsymbol{w}_1^{(T)}\tan c_4 \geq \left\|(\boldsymbol{I}-\boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{w}_1^{(T)}\right\|$, which confirms

$$\theta_1^{(T_1)} = \angle(\boldsymbol{w}_1^{(T_1)}, \boldsymbol{a}) \leq c_4.$$

Here we use the fact that $\mu T\left(\frac{\tan c_4}{8} - \frac{2\hat{c}_3}{\hat{c}_5^2}\right) \geq \hat{c}_4\tan c_4 + 2$. By symmetry, identical bounds hold for $v_2^{(T_1)}, b_2^{(T_1)}, \theta_2^{(T_1)}$. This completes the proof of Lemma 6 with constants $c_0 \leq 1, c_4 \leq \frac{\pi}{4}, c_9 = \frac{64}{\tan c_4}, c_2 = 2, c_3 = \frac{1}{4}, c_{10} = \frac{1}{4e^{2\alpha}}, \sigma_0 = \frac{1}{8e^{2\alpha}}$, where $\alpha = \frac{65}{\tan c_4}$.

C.5. Proof of Phase 2 (Lemma 7)

We prove the lemma with the following constants: $c_0 \leq \frac{1}{10^8}, c_7 = \frac{1}{4}, c_8 = 32, c_{11} = \frac{1}{50}$.

We first show that for any iteration $\tau \in [T_1, T_1 + T_2]$, the following bounds hold:

$$0 < v_1^{(\tau)}, v_2^{(\tau)} \leq 2\sqrt{\|\boldsymbol{a}\|} \tag{26}$$

$$\theta_1^{(\tau)}, \theta_2^{(\tau)} \leq c_4 \leq \frac{\pi}{10} \tag{27}$$

$$\left|b_1^{(\tau)}\right|, \left|b_2^{(\tau)}\right| \leq c_{10} \|\boldsymbol{a}\| + (\tau - T_1)\mu \|\boldsymbol{a}\|^2 \leq c_{11} \|\boldsymbol{a}\| \tag{28}$$

We proceed by induction. According to Lemma 6, equations (26), (27) and (28) hold for $\tau = T_1$ with probability at least $1 - Ce^{-cd}$. Now assume that we have (26), (27) and (28) for $\tau = T_1 + t$ with $t \in [0, T_2 - 1]$. For $\tau = T_1 + t + 1$, observe that

$$\left|b_1^{(\tau)}\right|, \left|b_2^{(\tau)}\right| \leq c_{10} \|\boldsymbol{a}\| + t\mu \|\boldsymbol{a}\|^2 \leq \gamma \|\boldsymbol{a}\| .$$

By invoking Lemma 4, Lemma 5 with $\beta = 0$, we establish that (26) and (27) hold for $\tau = T_1 + t + 1$. To prove (28), we apply the Lemma 3 with the constant $c_6 = 6$, which yields:

$$\begin{aligned}
\left|b_1^{(\tau+1)} - b_1^{(\tau)}\right| &\leq c_6\mu^2 \left(\left(v_1^{(\tau)}\right)^2 + \left\|\boldsymbol{w}_1^{(\tau)}\right\|^2\right)\left(\left\|v_1^{(\tau)}\boldsymbol{w}_1^{(\tau)} - \boldsymbol{a}\right\|^2 + \left\|v_2^{(\tau)}\boldsymbol{w}_2^{(\tau)} + \boldsymbol{a}\right\|^2\right) \\
&\leq c_6\mu^2 \left(4\|\boldsymbol{a}\| + 5\|\boldsymbol{a}\|\right)\left(\left(2\sqrt{5}+1\right)^2 \|\boldsymbol{a}\|^2 + \left(2\sqrt{5}+1\right)^2 \|\boldsymbol{a}\|^2\right) \\
&\leq 3300\mu^2 \|\boldsymbol{a}\|^3 .
\end{aligned}$$

Here, we use the assumptions $v_1^{(\tau)}, v_2^{(\tau)} \leq 2\sqrt{\|\boldsymbol{a}\|}$ and the fact that $\left\|\boldsymbol{w}_1^{(\tau)}\right\|^2 \leq \left(v_1^{(\tau)}\right)^2 + b_1^{(\tau)} \leq (4 + c_{11})\|\boldsymbol{a}\| \leq 5\|\boldsymbol{a}\|$, $\left\|\boldsymbol{w}_2^{(\tau)}\right\|^2 \leq \left(v_2^{(\tau)}\right)^2 + b_2^{(\tau)} \leq (4 + c_{11})\|\boldsymbol{a}\| \leq 5\|\boldsymbol{a}\|$. Substituting these into the recursive relation for b_1 and using the Inequality (28) for $\tau = T_1 + t$ we have:

$$\left|b_1^{(\tau+1)}\right| \leq \left|b_1^{(\tau)}\right| + \left|b_1^{(\tau+1)} - b_1^{(\tau)}\right| \leq c_{10}\|\boldsymbol{a}\| + (\tau + 1 - T_1) \cdot 3300\mu^2 \|\boldsymbol{a}\|^3 \leq c_{11}\|\boldsymbol{a}\| .$$

Here we use the assumption that $\tau + 1 - T_1 \leq T_2 = \left\lceil \frac{c_8}{\mu\|\boldsymbol{a}\|} \ln\left(\frac{\sqrt{\|\boldsymbol{a}\|}}{\sigma}\right)\right\rceil$ and $\mu \leq \frac{c_0}{\|\boldsymbol{a}\| \ln\left(\frac{\sqrt{\|\boldsymbol{a}\|}}{\sigma}\right)}$. By substituting the constants $c_8 = 32$, $c_{10} \leq \frac{1}{100}$, $c_{11} = \frac{1}{50}$, and $c_0 \leq \frac{1}{10^8}$, one can verify that the requirement for c_{11} is satisfied. By symmetry, we also have $\left|b_2^{(\tau+1)}\right| \leq c_{11}\|\boldsymbol{a}\|$. Thus, we have (28) for $\tau = T_1 + t + 1$.

By symmetry, we only need to focus on $v_1, \boldsymbol{w}_1$. We first establish a lower bound on $\left\|\boldsymbol{w}_1^{(\tau+1)}\right\|$. Since $\boldsymbol{w}_1^T \nabla_{\boldsymbol{w}_1} \widehat{\mathcal{L}} = v_1 \nabla_{v_1} \widehat{\mathcal{L}}$ (due to Eq. 13), we have

$$\begin{aligned}
\left\|\boldsymbol{w}_1^{(\tau+1)}\right\|^2 &= \left\|\boldsymbol{w}_1 - \mu\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 \\
&= \|\boldsymbol{w}_1\|^2 - 2\mu\boldsymbol{w}_1^T\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} + \mu^2\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 \\
&= \|\boldsymbol{w}_1\|^2 - 2\mu v_1\nabla_{v_1}\widehat{\mathcal{L}} + \mu^2\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2 ,
\end{aligned}$$

and

$$\begin{aligned}\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\| &\geq \frac{|\boldsymbol{w}_1^T\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}|}{\|\boldsymbol{w}_1\|}\\ &= \frac{|v_1\nabla_{v_1}\widehat{\mathcal{L}}|}{\|\boldsymbol{w}_1\|}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\left\|\boldsymbol{w}_1^{(\tau+1)}\right\|^2 &= \|\boldsymbol{w}_1\|^2 - 2\mu v_1\nabla_{v_1}\widehat{\mathcal{L}} + \mu^2\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}}\right\|^2\\ &\geq \|\boldsymbol{w}_1\|^2 - 2\mu v_1\nabla_{v_1}\widehat{\mathcal{L}} + \mu^2\left(\frac{|v_1\nabla_{v_1}\widehat{\mathcal{L}}|}{\|\boldsymbol{w}_1\|}\right)^2\\ &= \left(\|\boldsymbol{w}_1\| - \mu\frac{v_1\nabla_{v_1}\widehat{\mathcal{L}}}{\|\boldsymbol{w}_1\|}\right)^2.\end{aligned}$$

It follows that

$$\left\|\boldsymbol{w}_1^{(\tau+1)}\right\| \geq \|\boldsymbol{w}_1\| - \mu\frac{v_1\nabla_{v_1}\widehat{\mathcal{L}}}{\|\boldsymbol{w}_1\|}. \tag{29}$$

We continue by estimating $\nabla_{v_1}\widehat{\mathcal{L}}$. Note that

$$\begin{aligned}\nabla_{v_1}\widehat{\mathcal{L}} &= \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - v_2\frac{(\pi-\theta)\cos\theta+\sin\theta}{\pi}\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\| - \boldsymbol{w}_1^T\boldsymbol{a} + \boldsymbol{w}_1^T\Delta\mathcal{G}_1\right)\\ &\leq \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \boldsymbol{w}_1^T\boldsymbol{a} + \boldsymbol{w}_1^T\Delta\mathcal{G}_1\right)\\ &\leq \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \cos c_4\|\boldsymbol{w}_1\|\|\boldsymbol{a}\| + \|\boldsymbol{w}_1\|\|\Delta\mathcal{G}_1\|\right)\\ &\leq \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \frac{5}{6}\|\boldsymbol{w}_1\|\|\boldsymbol{a}\| + \frac{1}{3}\|\boldsymbol{w}_1\|\|\boldsymbol{a}\|\right)\\ &\leq \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \frac{1}{2}\|\boldsymbol{w}_1\|\|\boldsymbol{a}\|\right).\end{aligned}$$

In the penultimate line, we use the assumptions $c_4 = \frac{\pi}{10} \leq \arccos\left(\frac{5}{6}\right)$ and $\|\Delta\mathcal{G}_1\| \leq \frac{1}{3}\|\boldsymbol{a}\|$. To establish the lower bound for $v_1^{(T_1+T_2)}$, we will show that in phase 2, $v_1\|\boldsymbol{w}_1\|$ increases when it is small and will never decrease too much when it is large. Specifically, we consider the following three cases based on the magnitude of $v_1\|\boldsymbol{w}_1\|$:

- **Case 1:** $v_1\|\boldsymbol{w}_1\| \leq \frac{1}{4}\|\boldsymbol{a}\|$.
  We have

$$\nabla_{v_1}\widehat{\mathcal{L}} \leq \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \frac{1}{2}\|\boldsymbol{w}_1\|\|\boldsymbol{a}\|\right) \leq -\frac{1}{8}\|\boldsymbol{a}\|\|\boldsymbol{w}_1\|.$$

By Inequality (29), we have

$$\begin{aligned}\left\|\boldsymbol{w}_1^{(\tau+1)}\right\| \geq& \|\boldsymbol{w}_1\| - \mu\frac{v_1\nabla_{v_1}\widehat{\mathcal{L}}}{\|\boldsymbol{w}_1\|}\\ \geq& \|\boldsymbol{w}_1\| - \mu\frac{v_1(-\frac{1}{8}\|\boldsymbol{w}_1\|\|\boldsymbol{a}\|)}{\|\boldsymbol{w}_1\|}\\ =& \|\boldsymbol{w}_1\| + \frac{1}{8}\mu\|\boldsymbol{a}\| v_1.\end{aligned}$$

Summing the two inequalities we conclude that

$$v_1^{(\tau+1)} + \left\|\boldsymbol{w}_1^{(\tau+1)}\right\| \geq \left(1+\frac{1}{8}\mu\|\boldsymbol{a}\|\right)\left(v_1^{(\tau)} + \left\|\boldsymbol{w}_1^{(\tau)}\right\|\right).$$

- **Case 2:** $\frac{1}{4}\|\boldsymbol{a}\| < v_1\|\boldsymbol{w}_1\| \leq \frac{1}{2}\|\boldsymbol{a}\|$.
We have

$$\begin{aligned}\nabla_{v_1}\widehat{\mathcal{L}} \leq& \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \frac{1}{2}\|\boldsymbol{w}_1\|\|\boldsymbol{a}\|\right)\\ \leq& 0,\end{aligned}$$

which means $v_1^{(\tau+1)} = v_1 - \mu\nabla_{v_1}\widehat{\mathcal{L}} \geq v_1$. By Inequality (29), we have

$$\begin{aligned}\left\|\boldsymbol{w}_1^{(\tau+1)}\right\| \geq& \|\boldsymbol{w}_1\| - \mu\frac{v_1\nabla_{v_1}\widehat{\mathcal{L}}}{\|\boldsymbol{w}_1\|}\\ \geq& \|\boldsymbol{w}_1\|.\end{aligned}$$

Thus, we have $v_1^{(\tau+1)}\|\boldsymbol{w}_1\|^{(\tau+1)} \geq v_1\|\boldsymbol{w}_1\| \geq \frac{1}{4}\|\boldsymbol{a}\|$.

- **Case 3:** $v_1\|\boldsymbol{w}_1\| > \frac{1}{2}\|\boldsymbol{a}\|$.
We have

$$\begin{aligned}\nabla_{v_1}\widehat{\mathcal{L}} \leq& \frac{1}{2}\left(v_1\|\boldsymbol{w}_1\|^2 - \frac{1}{2}\|\boldsymbol{w}_1\|\|\boldsymbol{a}\|\right)\\ \leq& \frac{1}{2}v_1\|\boldsymbol{w}_1\|^2,\end{aligned}$$

which means $v_1^{(\tau+1)} = v_1 - \mu\nabla_{v_1}\widehat{\mathcal{L}} \geq v_1 - \frac{\mu v_1}{2}\|\boldsymbol{w}_1\|^2 \geq v_1 - \frac{1}{4}v_1 = \frac{3}{4}v_1$. Here we use the fact that $\mu\|\boldsymbol{w}_1\|^2 \leq c_0(c_2^2+\gamma) \leq \frac{1}{2}$. Using Inequality (29), we have

$$\begin{aligned}\left\|\boldsymbol{w}_1^{(\tau+1)}\right\| \geq& \|\boldsymbol{w}_1\| - \mu\frac{v_1\nabla_{v_1}\widehat{\mathcal{L}}}{\|\boldsymbol{w}_1\|}\\ \geq& \|\boldsymbol{w}_1\| - \frac{\mu v_1^2}{2}\|\boldsymbol{w}_1\|\\ \geq& \|\boldsymbol{w}_1\| - \frac{1}{4}\|\boldsymbol{w}_1\|\\ =& \frac{3}{4}\|\boldsymbol{w}_1\|.\end{aligned}$$

Here we use the fact that $\mu v_1^2 \le c_0 c_2^2 \le \frac{1}{2}$. Thus, we have $v_1^{(\tau+1)} \left\|\boldsymbol{w}_1\right\|^{(\tau+1)} \ge \frac{3}{4} v_1 \cdot \frac{3}{4} \left\|\boldsymbol{w}_1\right\| \ge \frac{1}{2} v_1 \left\|\boldsymbol{w}_1\right\| \ge \frac{1}{4} \left\|\boldsymbol{a}\right\|$.

As a result, the sum $v_1 + \|\boldsymbol{w}_1\|$ will increase by a factor $\left(1 + \frac{1}{8}\mu \left\|\boldsymbol{a}\right\|\right)$ as long as $v_1 \left\|\boldsymbol{w}_1\right\| < \frac{1}{4} \left\|\boldsymbol{a}\right\|$. Once we have $v_1 \left\|\boldsymbol{w}_1\right\| \ge \frac{1}{4} \left\|\boldsymbol{a}\right\|$ at some iteration, it remains bounded below by $\frac{1}{4} \left\|\boldsymbol{a}\right\|$. By Lemma 6, $v_1^{(T_1)} + \left\|\boldsymbol{w}_1^{(T_1)}\right\| \ge c_3 \sigma \ge \frac{1}{4}\sigma$. Consequently, after $T_2 = \lceil \frac{c_8}{\mu \|\boldsymbol{a}\|} \ln\left(\frac{\sqrt{\|\boldsymbol{a}\|}}{\sigma}\right) \rceil$ iterations, where we have used the constant $c_8 = 32$, we have $v_1^{(T_1+T_2)} \left\|\boldsymbol{w}_1^{(T_1+T_2)}\right\| \ge \frac{1}{4} \left\|\boldsymbol{a}\right\|$. Combining this with the imbalance bound $\left| \left(v_1^{(T_1+T_2)}\right)^2 - \left\|\boldsymbol{w}_1^{(T_1+T_2)}\right\|^2 \right| \le c_{11} \left\|\boldsymbol{a}\right\| \le \frac{1}{2} \left\|\boldsymbol{a}\right\|$, we have $v_1^{(T_1+T_2)} \ge \frac{1}{4}\sqrt{\|\boldsymbol{a}\|}$. This completes the proof with the constant $c_{11} = \frac{1}{2}$.

C.6. Uniform Concentration (Lemma 10)

We prove the result in 5 steps.

Step 1: Standard net reduction for operator norm.

Lemma 14 (Operator norm on a net) *Let $A \in \mathbb{R}^{d\times d}$ be symmetric and let $U \subset \mathbb{S}^{d-1}$ be an ε-net with $\varepsilon \in (0, 1/2)$. Then*

$$\|A\| \;\le\; \frac{1}{1-2\varepsilon} \max_{u\in\mathcal{U}} |u^\top A u|.$$

Moreover, there exists a $1/4$-net U with $|U| \le 9^d$.

Hence it suffices to control, uniformly over $\boldsymbol{w}, \boldsymbol{w}^*$,

$$\max_{u\in\mathcal{U}} \left| u^\top \Big(M(\boldsymbol{w}, \boldsymbol{w}^*) - \mathbb{E}M(\boldsymbol{w}, \boldsymbol{w}^*) \Big) u \right|.$$

Step 2: Truncation decomposition. Fix $\tau \ge 1$ and define truncation and remainder for $t \ge 0$:

$$T_\tau(t) := \min\{t, \tau\}, \qquad R_\tau(t) := (t-\tau)_+.$$

Fix $u \in \mathbb{S}^{d-1}$ and define $W_u(x) := \langle x, u\rangle^2$ and

$$W_u^{(\tau)}(x) := T_\tau(W_u(x)), \qquad G_u^{(\tau)}(x) := R_\tau(W_u(x)).$$

For any $(\boldsymbol{w}, \boldsymbol{w}^*)$ and any x,

$$W_u(x)\, \mathbb{1}_{\{\langle x,\boldsymbol{w}\rangle \ge 0\}} \mathbb{1}_{\{\langle x,\boldsymbol{w}^*\rangle \ge 0\}} \;\le\; W_u^{(\tau)}(x)\, \mathbb{1}_{\{\langle x,\boldsymbol{w}\rangle \ge 0\}} \mathbb{1}_{\{\langle x,\boldsymbol{w}^*\rangle \ge 0\}} \;+\; G_u^{(\tau)}(x), \tag{30}$$

since $W_u = W_u^{(\tau)} + G_u^{(\tau)}$ and indicators are ≤ 1.

For fixed u, define the *bounded* function class

$$\mathcal{F}_{u,\tau} := \Big\{ f_{\boldsymbol{w},\boldsymbol{w}^*}(x) := W_u^{(\tau)}(x)\, \mathbb{1}_{\{\langle x,\boldsymbol{w}\rangle \ge 0\}} \mathbb{1}_{\{\langle x,\boldsymbol{w}^*\rangle \ge 0\}} \;:\; \boldsymbol{w}, \boldsymbol{w}^* \in \mathbb{S}^{d-1} \Big\}.$$

Then every $f \in \mathcal{F}_{u,\tau}$ satisfies $0 \le f \le \tau$ pointwise.

Step 3: Tail remainder bound

Lemma 15 (Chi-square tail moments) *Let $Z \sim \mathcal{N}(0,1)$ and $Y = (Z^2 - \tau)_+$. There exist universal constants $c, C > 0$ such that for all $\tau \geq 1$,*

$$\mathbb{E}Y \leq Ce^{-c\tau}, \qquad \mathbb{E}Y^2 \leq Ce^{-c\tau}.$$

Proof Standard: $\mathbb{P}(Z^2 > t) \leq 2e^{-t/2}$ and $\mathbb{E}(Z^2 - \tau)_+ = \int_\tau^\infty \mathbb{P}(Z^2 > t)\, dt$, $\mathbb{E}(Z^2 - \tau)_+^2 = 2\int_\tau^\infty (t-\tau)\mathbb{P}(Z^2 > t)\, dt$. ■

Lemma 16 (Uniform control of the truncation tail over a $1/4$-net) *Let U be a $1/4$-net with $|U| \leq 9^d$. There exist universal $c, C > 0$ such that for all $\tau \geq 1$, with probability at least $1 - 2e^{-cd}$,*

$$\max_{u \in \mathcal{U}} \left| \frac{1}{n} \sum_{i=1}^n G_u^{(\tau)}(x_i) - \mathbb{E}G_u^{(\tau)}(X) \right| \;\leq\; C\sqrt{\frac{e^{-c\tau}d}{n}} + C\frac{d}{n}.$$

Moreover, $\max_{u \in \mathcal{U}} \mathbb{E}G_u^{(\tau)}(X) \leq Ce^{-c\tau}$.

Proof Fix $u \in \mathcal{U}$. Then $G_u^{(\tau)}(X) \overset{d}{=} (Z^2 - \tau)_+$ with $Z \sim \mathcal{N}(0,1)$. By Lemma 15, $\mathbb{E}G_u^{(\tau)} \lesssim e^{-c\tau}$ and $\mathbb{E}(G_u^{(\tau)})^2 \lesssim e^{-c\tau}$, and $G_u^{(\tau)}$ is sub-exponential (dominated by Z^2). More specifically, it is a (σ^2, b) subexponential with $\sigma^2 \leq Ce^{-c\tau}$ and b a constant.

Definition 17 (Sub-exponential random variable with parameters (σ^2, b)) *A real-valued random variable X is said to be* sub-exponential *with parameters (σ^2, b) if, for all $|\lambda| \leq \frac{1}{b}$,*

$$\mathbb{E}\exp\big(\lambda(X - \mathbb{E}X)\big) \;\leq\; \exp\left(\frac{\lambda^2\sigma^2}{2}\right). \tag{31}$$

For such subexponential random variables we have the following standard refined Bernstein-type inequality.

Theorem 18 (Bernstein's-type inequality for sub-exponential sums) *Let $X_1, \ldots, X_n$ be independent mean-zero random variables, where each X_i are sub-exponential with parameters (σ^2, b) in the sense of (31). Then, for every $s \geq 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq s\right) \;\leq\; 2\exp\left(-\frac{1}{2}n \cdot \min\left\{\frac{s^2}{\sigma^2},\; \frac{s}{b}\right\}\right). \tag{32}$$

Thus using using this Bernstein's inequality for sub-exponential variables above with $s = C\sqrt{\frac{e^{-c\tau}t}{n}} + C\frac{t}{n}$ yields for all $t \geq 1$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n G_u^{(\tau)}(x_i) - \mathbb{E}G_u^{(\tau)}(x) \geq C\sqrt{\frac{e^{-c\tau}t}{n}} + C\frac{t}{n}\right) \leq e^{-t}.$$

Set $t = c_1 d$ and union bound over $|\mathcal{U}| \leq 9^d$; choosing c_1 large enough makes the union-bound failure probability $\leq e^{-cd}$. ■

Step 4: Uniform control of the truncated term. We will use the following result on covering numbers for VC-subgraph classes.

Theorem 19 (Theorem 2.6.7 in Van der Vaart and Wellner (1996)) *Let $\mathcal{H}$ be a VC-subgraph class of real-valued functions on $\mathbb{R}^d$ with VC-subgraph dimension at most V and envelope bound $|h| \le B$ pointwise. Then there exist absolute constants $A, C > 0$ such that for every probability measure Q and every $0 < \eta \le B$,*

$$\log N(\eta, \mathcal{H}, L_2(Q)) \;\le\; CV \log\Big(\frac{AB}{\eta}\Big).$$

We will also use the following Dudley-type bound for Rademacher averages.

Theorem 20 (Equation (5.48) in Wainwright (2019)) *There exists an absolute constant $C > 0$ such that, for any function class $\mathcal{F}$, conditionally on $x_1, \dots, x_n$,*

$$\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\Big] \le \frac{C}{\sqrt{n}}\int_0^{2r}\sqrt{\log N(\eta, \mathcal{F}, L_2(\mathbb{P}_n))}\,d\eta,$$

where $\mathbb{P}_n$ is the empirical measure of $x_1, \dots, x_n$ and $r := \sup_{f\in\mathcal{F}}\|f\|_{L_2(\mathbb{P}_n)}$.

Lemma 21 (Expected supremum for the truncated class) *There exists a universal constant $C > 0$ such that for each fixed u,*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}_{u,\tau}} |(\mathbb{P}_n - \mathbb{P})f|\Big] \;\le\; C\tau\sqrt{\frac{d}{n}}.$$

Proof By symmetrization,

$$\mathbb{E}\sup_{f\in\mathcal{F}_{u,\tau}} |(\mathbb{P}_n - \mathbb{P})f| \le 2\mathbb{E}\sup_{f\in\mathcal{F}_{u,\tau}}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|.$$

Let $r := \sup_{f\in\mathcal{F}_{u,\tau}}\|f\|_{L_2(\mathbb{P}_n)}$. Since $0 \le f \le \tau$ we have $r \le \tau$. Applying Dudley's entropy integral bound for Rademacher averages per Theorem 20 and the covering bound from Theorem 19 with $Q = \mathbb{P}_n$ yields

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big|\Big] \;\le\; \frac{C_3}{\sqrt{n}}\int_0^{2r}\sqrt{\log N\big(\eta, \mathcal{F}, L_2(\mathbb{P}_n)\big)}\,d\eta \;\le\; \frac{C_3}{\sqrt{n}}\int_0^{2r}\sqrt{V\log\Big(\frac{A\tau}{\eta}\Big)}\,d\eta.$$

Using the change of variables $\eta = re^{-s}$ and the elementary bound

$$\int_0^{2r}\sqrt{\log\Big(\frac{A\tau}{\eta}\Big)}\,d\eta \;\le\; C_4\, r\sqrt{\log\Big(\frac{eA\tau}{r}\Big)},$$

we obtain

$$\mathbb{E}\Big[\sup_{f\in F}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big|\Big] \;\le\; C_5\,\frac{r}{\sqrt{n}}\sqrt{V}\sqrt{\log\Big(\frac{eA\tau}{r}\Big)}.$$

Substituting $r \le \tau$ and $V \le c_2 d$, and absorbing constants (including A and C_0) into a single absolute constant C, gives

$$\mathbb{E}\Big[\sup_{f\in F}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\Big|\Big] \;\le\; C\tau\sqrt{\frac{d}{n}}.$$

Finally, multiplying by 2 from symmetrization proves the claim. ■

Now we focus on establishing a high-probability bound. For this we will use Bousquet's concentration inequality for suprema of bounded empirical processes.

Theorem 22 (Theorem 2.3 in Bousquet (2002)) *Let $\mathcal{F}$ be a class of measurable functions with $0 \le f \le b$. Let*

$$Z := \sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\big(f(X_i)-\mathbb{E}f(X)\big), \qquad \sigma^2 := \sup_{f\in\mathcal{F}}\mathrm{Var}(f(X)).$$

Then for all $t \ge 0$, with probability at least $1-e^{-t}$,

$$Z \le \mathbb{E}Z + \sqrt{2t(n\sigma^2 + 2b\mathbb{E}Z)} + \frac{b\,t}{3}. \tag{33}$$

In particular, we will further upper bound the RHS of (33) as

$$\mathbb{E}Z + \sqrt{2t(n\sigma^2 + 2b\mathbb{E}Z)} + \frac{b\,t}{3} \le 2\mathbb{E}Z + \sqrt{2tn\sigma^2} + \frac{4b\,t}{3}, \tag{34}$$

where we have used that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$.

Lemma 23 (Uniform high-probability deviation for $\mathcal{F}_{u,\tau}$) *There exist universal constants $C, c > 0$ such that for each fixed u and each $t \ge 1$, with probability at least $1-e^{-t}$,*

$$\sup_{f\in\mathcal{F}_{u,\tau}} |(\mathbb{P}_n - \mathbb{P})f| \le C\tau\sqrt{\frac{d}{n}} + C\sqrt{\frac{\tau t}{n}} + C\frac{\tau t}{n}.$$

Proof Apply (34) to $\mathcal{F} = \mathcal{F}_{u,\tau}$ with $b = \tau$. We have $\sigma^2 \le \sup_f \mathbb{E}f^2 \le \tau/2$ as above. Also, $\mathbb{E}Z = n \cdot \mathbb{E}\sup_{f\in\mathcal{F}_{u,\tau}} |(\mathbb{P}_n - \mathbb{P})f|$ which is bounded by Lemma 21. Divide by n to conclude. ■

Step 5: Complete the operator-norm bound Fix $\tau > 0$ and recall the decomposition

$$W_u(x) \;=\; W_u^{(\tau)}(x) + G_u^{(\tau)}(x), \qquad W_u^{(\tau)}(x) := \min\{\langle x,u\rangle^2, \tau\}, \qquad G_u^{(\tau)}(x) := (\langle x,u\rangle^2 - \tau)_+.$$

For $(\boldsymbol{w}, \boldsymbol{w}^\star)$ define the ReLU sign indicator

$$\mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}(x) := \mathbb{1}\{\langle x, \boldsymbol{w}\rangle \ge 0\}\,\mathbb{1}\{\langle x, \boldsymbol{w}^\star\rangle \ge 0\} \in \{0,1\}.$$

Then for every x,

$$W_u(x)\,\mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}(x) = W_u^{(\tau)}(x)\,\mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}(x) + G_u^{(\tau)}(x)\,\mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}(x),$$

and hence, by the triangle inequality,

$$\sup_{\boldsymbol{w},\boldsymbol{w}^\star} \left|(\mathbb{P}_n - \mathbb{P})\big(W_u \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}\big)\right| \;\le\; \sup_{\boldsymbol{w},\boldsymbol{w}^\star} \left|(\mathbb{P}_n - \mathbb{P})\big(W_u^{(\tau)} \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}\big)\right| \;+\; \sup_{\boldsymbol{w},\boldsymbol{w}^\star} \left|(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)} \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}\big)\right|. \tag{35}$$

We control the second term in (35) uniformly over $(\boldsymbol{w}, \boldsymbol{w}^\star)$ without any VC/covering argument. Using the identity

$$G_u^{(\tau)} \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star} = G_u^{(\tau)} - G_u^{(\tau)}(1 - \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}),$$

we have for every $(\boldsymbol{w}, \boldsymbol{w}^\star)$,

$$(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)} \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}\big) = (\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)}\big) - (\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)}(1 - \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star})\big),$$

so

$$\left|(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)} \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}\big)\right| \;\le\; \left|(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)}\big)\right| + \left|(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)}(1 - \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star})\big)\right|. \tag{36}$$

Since $G_u^{(\tau)} \ge 0$ and $0 \le 1 - \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star} \le 1$, we have pointwise

$$0 \le G_u^{(\tau)}(1 - \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}) \le G_u^{(\tau)},$$

and therefore

$$\mathbb{P}_n\big(G_u^{(\tau)}(1 - \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star})\big) \le \mathbb{P}_n\big(G_u^{(\tau)}\big), \qquad \mathbb{P}\big(G_u^{(\tau)}(1 - \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star})\big) \le \mathbb{P}\big(G_u^{(\tau)}\big).$$

Consequently,

$$\left|(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)}(1 - \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star})\big)\right| \le \mathbb{P}_n\big(G_u^{(\tau)}\big) + \mathbb{P}\big(G_u^{(\tau)}\big) = (\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)}\big) + 2\,\mathbb{P}\big(G_u^{(\tau)}\big),$$

and hence

$$\sup_{\boldsymbol{w},\boldsymbol{w}^\star} \left|(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)} \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}\big)\right| \;\le\; 2\left|(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)}\big)\right| \;+\; 2\,\mathbb{P}\big(G_u^{(\tau)}\big). \tag{37}$$

Plugging (37) into (35) yields the corrected completion bound:

$$\sup_{\boldsymbol{w},\boldsymbol{w}^\star} \left|(\mathbb{P}_n - \mathbb{P})\big(W_u \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}\big)\right| \;\le\; \sup_{\boldsymbol{w},\boldsymbol{w}^\star} \left|(\mathbb{P}_n - \mathbb{P})\big(W_u^{(\tau)} \mathbb{1}_{\boldsymbol{w},\boldsymbol{w}^\star}\big)\right| \;+\; 2\left|(\mathbb{P}_n - \mathbb{P})\big(G_u^{(\tau)}\big)\right| \;+\; 2\,\mathbb{P}\big(G_u^{(\tau)}\big). \tag{38}$$

Now apply Lemma 23 and Lemma 16 and union bound over $u \in \mathcal{U}$. Taking $t = c_0 d$ with c_0 large enough absorbs the 9^d factor, giving with probability $\ge 1 - 3e^{-cd}$:

$$\max_{u \in \mathcal{U}} \sup_{\boldsymbol{w},\boldsymbol{w}^*} \left|u^\top (M - \mathbb{E}M) u\right| \le C\tau\sqrt{\frac{d}{n}} + C\sqrt{\frac{\tau d}{n}} + C\frac{\tau d}{n} + C\sqrt{\frac{e^{-c\tau} d}{n}} + Ce^{-c\tau} + C\frac{d}{n}.$$

Multiplying by 2 from the net lemma gives the same bound for $\sup_{\boldsymbol{w},\boldsymbol{w}^*} \|M - \mathbb{E}M\|$.

Finally choose $\tau = C\log(e/\delta)$ so that $e^{-c\tau} \ll \delta$. If $n \ge Cd\,\frac{\log^2(1/\delta)}{\delta^2}$, then each term on the right is $\le \delta$ (after increasing the universal constants), which proves Theorem 10.

C.7. Concentration of Gradient Component Deviations (Lemma 11)

For the proof, we utilize the variational characterization of the Euclidean norm: $\|\boldsymbol{z}\| = \sup_{\|\mathbf{u}\|=1} \langle \mathbf{u}, \boldsymbol{z} \rangle$.

Bound for $\boldsymbol{w}_1$: Let $\mathbf{u} \in \mathbb{S}^{d-1}$ be an arbitrary unit vector. Using the Fundamental Theorem of Calculus, the definition of the ReLU gradient, and the residual $r(\boldsymbol{x}_i) = v_1\phi(\boldsymbol{w}_1^\top \boldsymbol{x}_i) - v_2\phi(\boldsymbol{w}_2^\top \boldsymbol{x}_i) - (\phi(\boldsymbol{a}^\top \boldsymbol{x}_i) - \phi(-\boldsymbol{a}^\top \boldsymbol{x}_i))$, we observe:

$$\left\langle \mathbf{u}, \nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L} \right\rangle = v_1 \int_0^1 \mathbf{u}^\top \left(M(t(v_1\boldsymbol{w}_1) + (1-t)\boldsymbol{a}, \boldsymbol{w}_1) - \mathbb{E}M(\dots)\right) \mathbf{h}_1 \, dt + v_1 \int_0^1 \mathbf{u}^\top \left(M(-t\boldsymbol{a} + (1-t)(v_2\boldsymbol{w}_2), \boldsymbol{w}_1) - \mathbb{E}M(\dots)\right) (-\mathbf{h}_2) \, dt,$$

where $M(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})$ is defined as in Eq. 4. The term v_1 appears due to the chain rule derivative with respect to $\boldsymbol{w}_1$. By Lemma 10, given the sample complexity $n \geq Cd\frac{\log(1/\delta)^2}{\delta^2}$, the following spectral deviation bound holds with probability at least $1 - 3e^{-cd}$:

$$\sup_{\tilde{\mathbf{u}}, \tilde{\mathbf{v}} \in \mathbb{S}^{d-1}} \|M(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}) - \mathbb{E}\left[M(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})\right]\| \leq \delta. \tag{39}$$

Since the indicator functions $\mathbb{1}_{\{\langle \boldsymbol{x}, \mathbf{u}\rangle \geq 0\}}$ are scale-invariant, this uniform bound applies to all directions $t\mathbf{u} + (1-t)\mathbf{v}$ appearing in the integrals. Then,

$$\begin{aligned}\left|\left\langle \mathbf{u}, \nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L} \right\rangle\right| &\leq v_1\delta \|\mathbf{u}\| \|\mathbf{h}_1\| + v_1\delta \|\mathbf{u}\| \|\mathbf{h}_2\| \\ &= v_1\delta(\|\mathbf{h}_1\| + \|\mathbf{h}_2\|).\end{aligned}$$

Taking the supremum over $\mathbf{u}$ proves the first bound.

Bound for $\boldsymbol{w}_2$: Similarly, for an arbitrary unit vector $\mathbf{v} \in \mathbb{S}^{d-1}$:

$$\left\langle \mathbf{v}, \nabla_{\boldsymbol{w}_2}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_2}\mathcal{L} \right\rangle = v_2 \int_0^1 \mathbf{v}^\top \left(M(t(v_2\boldsymbol{w}_2) - (1-t)\boldsymbol{a}, \boldsymbol{w}_2) - \mathbb{E}M(\dots)\right) \mathbf{h}_2 \, dt + v_2 \int_0^1 \mathbf{v}^\top \left(M(t\boldsymbol{a} + (1-t)(v_1\boldsymbol{w}_1), \boldsymbol{w}_2) - \mathbb{E}M(\dots)\right) (-\mathbf{h}_1) \, dt.$$

Applying the same uniform spectral bound yields:

$$\left|\left\langle \mathbf{v}, \nabla_{\boldsymbol{w}_2}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_2}\mathcal{L} \right\rangle\right| \leq v_2\delta(\|\mathbf{h}_1\| + \|\mathbf{h}_2\|).$$

This completes the proof of Lemma 11.

Corollary 24 (Bounds in terms of $\|\boldsymbol{a}\|$) *Further assume that $\|v_1\boldsymbol{w}_1\| \leq C\|\boldsymbol{a}\|$ and $\|v_2\boldsymbol{w}_2\| \leq C\|\boldsymbol{a}\|$. Then:*

$$\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L}\right\| \leq \tilde{c}v_1\|\boldsymbol{a}\|, \tag{40}$$

$$\left\|\nabla_{\boldsymbol{w}_2}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_2}\mathcal{L}\right\| \leq \tilde{c}v_2\|\boldsymbol{a}\|, \tag{41}$$

where $\tilde{c} = 2\delta(C+1)$.

Proof We bound the norms of the error vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ using the triangle inequality:

$$\|\mathbf{h}_1\| = \|v_1\boldsymbol{w}_1 - \boldsymbol{a}\| \le \|v_1\boldsymbol{w}_1\| + \|\boldsymbol{a}\| \le (C+1)\|\boldsymbol{a}\|,$$
$$\|\mathbf{h}_2\| = \|v_2\boldsymbol{w}_2 + \boldsymbol{a}\| \le \|v_2\boldsymbol{w}_2\| + \|\boldsymbol{a}\| \le (C+1)\|\boldsymbol{a}\|.$$

Substituting these into the theorem's bounds:

$$\left\|\nabla_{\boldsymbol{w}_1}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_1}\mathcal{L}\right\| \le v_1\delta\left(2(C+1)\|\boldsymbol{a}\|\right) = 2\delta(C+1)v_1\|\boldsymbol{a}\|,$$
$$\left\|\nabla_{\boldsymbol{w}_2}\widehat{\mathcal{L}} - \nabla_{\boldsymbol{w}_2}\mathcal{L}\right\| \le v_2\delta\left(2(C+1)\|\boldsymbol{a}\|\right) = 2\delta(C+1)v_2\|\boldsymbol{a}\|.$$

■

Appendix D. Proof of Main Theorems

D.1. Proof of Theorem 1 for Landscape Characterization

To prove this theorem, we first show that $v_1\boldsymbol{w}_1 = \boldsymbol{a}$, $\boldsymbol{v}_2\boldsymbol{w}_2 = -\boldsymbol{a}$ is the global optima. Since $v_1, v_2 > 0$,

$$v_1\phi\left(\boldsymbol{w}_1^T\boldsymbol{x}\right) - v_2\phi\left(\boldsymbol{w}_2^T\boldsymbol{x}\right) = \phi\left(v_1\boldsymbol{w}_1^T\boldsymbol{x}\right) - \phi\left(v_2\boldsymbol{w}_2^T\boldsymbol{x}\right) = \phi\left(\boldsymbol{a}^T\boldsymbol{x}\right) - \phi\left(-\boldsymbol{a}^T\boldsymbol{x}\right) = \boldsymbol{a}^T\boldsymbol{x}.$$

Hence, the given weights implement the planted model exactly. Next, we verify that all $v_1, v_2 > 0$, and $\boldsymbol{w}_1, \boldsymbol{w}_2$ that satisfy

$$v_1\boldsymbol{w}_1 - v_2\boldsymbol{w}_2 = \boldsymbol{a}, \quad \text{and} \quad \theta = 0$$

are indeed non-strict saddle points of our optimization problem when $k = 2$. We first show that the gradient vanishes. Plugging such $v_1, v_2, \boldsymbol{w}_1, \boldsymbol{w}_2$ into (2):

$$\begin{aligned}
\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{\theta}) =& \frac{1}{2\pi} diag(\boldsymbol{v})\left(\left(\pi\mathbb{1}\mathbb{1}^T - \boldsymbol{\Theta}\right) diag(\boldsymbol{u}) + diag\left(\sin(\boldsymbol{\Theta})\boldsymbol{u}\right)\right)\bar{\boldsymbol{W}} - \frac{1}{2}\boldsymbol{v}\boldsymbol{a}^T \\
\overset{(a)}{=}& \frac{1}{2} diag(\boldsymbol{v})\,\mathbb{1}\mathbb{1}^T diag(\boldsymbol{v})\,\boldsymbol{W} - \frac{1}{2}\boldsymbol{v}\boldsymbol{a}^T \\
=& \frac{1}{2}\boldsymbol{v}\left(\boldsymbol{W}^T\boldsymbol{v} - \boldsymbol{a}\right)^T \\
=& \frac{1}{2}\boldsymbol{v}\left(v_1\boldsymbol{w}_1 - v_2\boldsymbol{w}_2 - \boldsymbol{a}\right)^T = \frac{1}{2}\boldsymbol{v}\left(\boldsymbol{a} - \boldsymbol{a}\right)^T = \boldsymbol{0}.
\end{aligned}$$

where (a) follows from the fact that $\boldsymbol{\Theta} = \boldsymbol{0}$ at these points. Furthermore, due to (13), $\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{\theta})$ is also $\boldsymbol{0}$. Next we show that the Hessian at these points are PSD. Plugging the values into (14) we get:

$$\nabla^2\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\begin{bmatrix}
\|\boldsymbol{w}_1\|^2 & -\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\| & v_1\boldsymbol{w}_1^T & -v_2\boldsymbol{w}_1^T \\
-\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\| & \|\boldsymbol{w}_2\|^2 & -v_1\boldsymbol{w}_2^T & v_2\boldsymbol{w}_2^T \\
v_1\boldsymbol{w}_1 & -v_1\boldsymbol{w}_2 & v_1^2\boldsymbol{I} & -v_1v_2\boldsymbol{I} \\
-v_2\boldsymbol{w}_1 & v_2\boldsymbol{w}_2 & -v_1v_2\boldsymbol{I} & v_2^2\boldsymbol{I}
\end{bmatrix}$$

which follows from the fact that $\theta_{\ell,i} = 0$ and $\bar{\boldsymbol{w}}_{m,\ell^\perp} = \bar{\boldsymbol{w}}_{\ell,m^\perp} = \mathbf{0}$ for any choice of $\ell, m, i \in [2]$. This $(2d+2) \times (2d+2)$ matrix has eigenvalues 0, $\frac{v_1^2+v_2^2}{2}$, and $\frac{\|\boldsymbol{w}_1\|^2+\|\boldsymbol{w}_2\|^2+v_1^2+v_2^2}{2}$ (all non-negative) with multiplicities $d+2, d-1$, and 1 respectively. Therefore, all the stationary points are in fact non-strict saddle points of the problem.

Finally, we show that there are no other stationary points besides the ones identified above. A necessary condition for $\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$ is that any linear combination of the gradient rows must vanish. Specifically, for $v_1, v_2 > 0$, we have:

$$\left[\tfrac{1}{v_1}, \quad \tfrac{1}{v_2}\right] \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}.$$

By substituting the gradient expression, this implies:

$$\theta \left(v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2\right) = \sin\theta \left(v_1 \|\boldsymbol{w}_1\| \bar{\boldsymbol{w}}_2 + v_2 \|\boldsymbol{w}_2\| \bar{\boldsymbol{w}}_1\right).$$

Note that the vectors $v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2$ and $v_1 \|\boldsymbol{w}_1\| \bar{\boldsymbol{w}}_2 + v_2 \|\boldsymbol{w}_2\| \bar{\boldsymbol{w}}_1$ have identical norms. Taking the norm of both sides, the equality holds only if $|\theta| = |\sin\theta|$, which implies $\theta = 0$, or if the vectors themselves are zero ($v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2 = \mathbf{0}$).

The case $\theta = 0$ corresponds to the non-strict saddle points previously identified. The case $v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2 = \mathbf{0}$ corresponds to the global optima where the two neurons are anti-aligned ($\theta = \pi$) such that their combined contribution exactly implements the target $\boldsymbol{a}$. Consequently, there are no other stationary points in the optimization landscape. This completes the proof of the theorem.

D.2. Proof of Theorem 2 for Convergence of the GD Trajectory

To prove this theorem first we note that after $T_1 = \lceil \frac{c_9}{\mu \|\boldsymbol{a}\|} \rceil$ iterations of GD (i.e. *alignment* phase), using Lemma 6 from Section 5.4 we have with high probability

$$\theta_1^{(T_1)}, \theta_2^{(T_1)} \le c_4, \quad \text{and} \quad c_1\sigma \le v_1^{(T_1)}, v_2^{(T_1)} \le c_2\sqrt{\|\boldsymbol{a}\|}.$$

Using Lemma 7, after $T = T_1 + T_2$ iterations, we have

$$\theta_1^{(T)}, \theta_2^{(T)} \le c_4, \quad v_1^{(T)}, v_2^{(T)} \ge c_7\sqrt{\|\boldsymbol{a}\|}, \quad \text{and} \quad \left|b_1^{(T)}\right|, \left|b_2^{(T)}\right| \le c_{11}\|\boldsymbol{a}\|.$$

Using the definition of the imbalance term, $\left\|\boldsymbol{w}_i^{(T)}\right\|^2 = (v_i^{(T)})^2 + b_i^{(T)}$, we evaluate the weights at the end of the growth phase ($T = T_1 + T_2$). From Lemma 7, we have $v_i^{(T)} \ge c_7\sqrt{\|\boldsymbol{a}\|}$ and $|b_i^{(T)}| \le c_{11}\|\boldsymbol{a}\|$. Noting that $c_7^2 > c_{11}$ for $c_7 = \frac{1}{4}, c_{11} = \frac{1}{50}$; this implies:

$$(c_7^2 - c_{11})\|\boldsymbol{a}\| \le \left\|\boldsymbol{w}_1^{(T)}\right\|^2, \left\|\boldsymbol{w}_2^{(T)}\right\|^2 \le (c_2^2 + c_{11})\|\boldsymbol{a}\|. \tag{42}$$

We now establish that these bounds hold uniformly for all $\tau \ge T$. Lemma 4 ensures that since the angles $\theta_1^{(T)}, \theta_2^{(T)}$ are small, they remain bounded by c_4 for all subsequent iterations. Lemma 5 ensures that if the norms v_i start in the interval $[c_7\sqrt{\|\boldsymbol{a}\|}, c_2\sqrt{\|\boldsymbol{a}\|}]$, they remain within a fixed range $[c_7\sqrt{\|\boldsymbol{a}\|}, c_2\sqrt{\|\boldsymbol{a}\|}]$ for all $\tau > T$. Finally, as established in the convergence analysis below, the error $\|\boldsymbol{h}_1\|^2 + \|\boldsymbol{h}_2\|^2$ decays geometrically. By Lemma 3, the total drift in the imbalance terms is summable, keeping $|b_i^{(\tau)}|$ uniformly bounded by a constant $\gamma\|\boldsymbol{a}\|$ for all $\tau \ge T$.

Consequently, there exist universal constants $c_{\min}$ and $c_{\max}$ such that for all $\tau \geq T$:

$$c_{\min}\sqrt{\|\boldsymbol{a}\|} \leq v_1^{(\tau)}, v_2^{(\tau)}, \left\|\boldsymbol{w}_1^{(\tau)}\right\|, \left\|\boldsymbol{w}_2^{(\tau)}\right\| \leq c_{\max}\sqrt{\|\boldsymbol{a}\|}. \tag{43}$$

From (43), the conditions for the PL inequality (Lemma 8) and smoothness (Lemma 9) hold uniformly for all $\tau \geq T$, where the constants α and L now depend on $c_{\min}$ and $c_{\max}$. Specifically, we have:

$$\left\|\nabla_{\boldsymbol{w}_1}\mathcal{L}\left(\boldsymbol{v}^{(\tau)}, \boldsymbol{W}^{(\tau)}\right)\right\|^2 + \left\|\nabla_{\boldsymbol{w}_2}\mathcal{L}\left(\boldsymbol{v}^{(\tau)}, \boldsymbol{W}^{(\tau)}\right)\right\|^2 \geq \alpha c_{\min}^2 \|\boldsymbol{a}\| \mathcal{L}\left(\boldsymbol{v}^{(\tau)}, \boldsymbol{W}^{(\tau)}\right),$$

and

$$\left\|\nabla^2\mathcal{L}\left(\boldsymbol{v}^{(\tau)}, \boldsymbol{W}^{(\tau)}\right)\right\|_F \leq L\|\boldsymbol{a}\|.$$

Next, we keep track of the *population* loss while performing GD updates on the *empirical* loss. Let $\boldsymbol{\theta}$ denote $\begin{bmatrix}\boldsymbol{v}\\ \text{vect}(\boldsymbol{W})\end{bmatrix}$, and also define errors vectors $\boldsymbol{h}_1 = v_1\boldsymbol{w}_1 - \boldsymbol{a}$, $\boldsymbol{h}_2 = v_2\boldsymbol{w}_2 + \boldsymbol{a}$. We note that $\left\|\text{diag}\left(\begin{bmatrix}v_1\\ v_2\end{bmatrix}\right)\boldsymbol{W} - \boldsymbol{W}^*\right\|_F^2 = \|\boldsymbol{h}_1\|^2 + \|\boldsymbol{h}_2\|^2$. For all $\tau \geq T$ we have

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{\theta}^{(\tau+1)}\right) =& \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)} - \mu\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right)\right)\\
\overset{(a)}{\leq}& \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) - \mu\left\langle\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right), \nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\rangle + \frac{L\|\boldsymbol{a}\|}{2}\mu^2\left\|\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\\
=& \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) - \mu\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2 - \mu\left\langle\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right), \nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right) - \nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\rangle\\
&+ \frac{L\|\boldsymbol{a}\|}{2}\mu^2\left\|\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\\
\overset{(b)}{\leq}& \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) - \mu\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2 + \eta\mu\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2 + \frac{\mu}{\eta}\left\|\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right) - \nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\\
&+ \frac{L\|\boldsymbol{a}\|}{2}\mu^2\left\|\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\\
\overset{(c)}{\leq}& \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) - \mu(1-\eta)\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2 + \frac{\mu}{\eta}\left\|\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right) - \nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\\
&+ L\|\boldsymbol{a}\|\mu^2\left(\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2 + \left\|\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right) - \nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\right)\\
=& \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) - \mu(1-\eta-\mu L\|\boldsymbol{a}\|)\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\\
&+ \left(\frac{\mu}{\eta} + \mu^2 L\|\boldsymbol{a}\|\right)\left\|\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right) - \nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\\
\overset{(d)}{\leq}& \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) - \alpha c_1^2\|\boldsymbol{a}\|\mu(1-\eta-\mu L\|\boldsymbol{a}\|)\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\\
&+ \left(\frac{\mu}{\eta} + \mu^2 L\|\boldsymbol{a}\|\right)\left\|\nabla\widehat{\mathcal{L}}\left(\boldsymbol{\theta}^{(\tau)}\right) - \nabla\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)\right\|^2\\
\overset{(e)}{\leq}& \left(1 - \alpha c_{\min}^2\|\boldsymbol{a}\|\mu(1-\eta-\mu L\|\boldsymbol{a}\|)\right)\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)
\end{aligned}$$

$$
\begin{aligned}
&\quad + \left(\frac{\mu}{\eta} + \mu^2 L \left\|\boldsymbol{a}\right\|\right) \left(\left(v_1^{(\tau)}\right)^2 + \left(v_2^{(\tau)}\right)^2 + \left\|\boldsymbol{w}_1^{(\tau)}\right\|^2 + \left\|\boldsymbol{w}_2^{(\tau)}\right\|^2\right) \delta^2 \left(\left\|\boldsymbol{h}_1^{(\tau)}\right\|^2 + \left\|\boldsymbol{h}_2^{(\tau)}\right\|^2\right) \\
&\overset{(f)}{\leq} \left(1 - \alpha c_{\min}^2 \left\|\boldsymbol{a}\right\| \mu \left(1 - \eta - \mu L \left\|\boldsymbol{a}\right\|\right)\right) \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) \\
&\quad + \left(\frac{\mu}{\eta} + \mu^2 L \left\|\boldsymbol{a}\right\|\right) 4 c_{\max}^2 \left\|\boldsymbol{a}\right\| \delta^2 \left(\left\|\boldsymbol{h}_1^{(\tau)}\right\|^2 + \left\|\boldsymbol{h}_2^{(\tau)}\right\|^2\right) \\
&\overset{(g)}{\leq} \left(1 - \alpha c_{\min}^2 \left\|\boldsymbol{a}\right\| \mu \left(1 - \eta - \mu L \left\|\boldsymbol{a}\right\|\right) + 20 \left(\frac{\mu}{\eta} + \mu^2 L \left\|\boldsymbol{a}\right\|\right) 4 c_{\max}^2 \left\|\boldsymbol{a}\right\| \delta^2\right) \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right)
\end{aligned}
$$

where (a) follows from the quadratic upper bound and smoothness of $\mathcal{L}$, (b) follows from $\left\langle \sqrt{\eta}\boldsymbol{a}, \frac{1}{\sqrt{\eta}}\boldsymbol{b} \right\rangle \leq \eta \left\|\boldsymbol{a}\right\|^2 + \frac{1}{\eta}\left\|\boldsymbol{b}\right\|^2$ for any $\eta > 0$, (c) follows from triangle inequality, (d) follows from PL inequality, (e) follows from Lemma 10 and gradient identity (13), (f) follows from upper bounds on the norms, and finally (g) follows from applying the population loss lower bound (Lemma 13). Set $\mu = \frac{\bar{\mu}}{\left\|\boldsymbol{a}\right\|}$, $\eta = \frac{1}{4}$. We have

$$
\begin{aligned}
\mathcal{L}\left(\boldsymbol{\theta}^{(\tau+1)}\right) &\leq \left(1 - \alpha c_{\min}^2 \bar{\mu} \left(\frac{3}{4} - \bar{\mu} L\right) + \left(\frac{4\bar{\mu}}{3} + \bar{\mu}^2 L\right) 80 c_{\max}^2 \delta^2\right) \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) \\
&= \left(1 - \left(\frac{3\alpha c_{\min}^2}{4} - \frac{320 c_{\max}^2 \delta^2}{3}\right) \bar{\mu} + \left(\alpha c_{\min}^2 + 80 c_{\max}^2 \delta^2\right) L \bar{\mu}^2\right) \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right).
\end{aligned}
$$

We now choose δ and $\bar{\mu}$ so that the quadratic factor above yields a strict contraction. First require that the linear coefficient is positive, i.e.

$$
\delta^2 \leq \frac{9\alpha c_{\min}^2}{1280 c_{\max}^2}.
$$

Fix any such δ (this is ensured by Lemma 10 by taking n sufficiently large). Next, define,

$$
a := \frac{3\alpha c_{\min}^2}{4} - \frac{320 c_{\max}^2 \delta^2}{3} \quad \text{and} \quad b := \left(\alpha c_{\min}^2 + 80 c_{\max}^2 \delta^2\right) L.
$$

If we further choose

$$
\bar{\mu} \leq \frac{a}{2b},
$$

then the quadratic term is dominated by the linear term, and we have

$$
1 - a\bar{\mu} + b\bar{\mu}^2 \leq 1 - \frac{a}{2}\bar{\mu}.
$$

Consequently, for all $\tau \geq T$,

$$
\mathcal{L}\left(\boldsymbol{\theta}^{(\tau+1)}\right) \leq \left(1 - c\bar{\mu}\right) \mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right),
$$

where $c := \frac{a}{2} > 0$ is a numerical constant depending only on $\alpha, c_{\min}, c_{\max}$. Iterating this inequality for all $\tau > T$ yields geometric decrease of the population loss

$$
\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) \leq \left(1 - c\bar{\mu}\right)^{(\tau - T)} \mathcal{L}\left(\boldsymbol{\theta}^{(T)}\right).
$$

Finally, we apply the population loss lower bound (Lemma 13) to lower bound the left-hand side:

$$\mathcal{L}\left(\boldsymbol{\theta}^{(\tau)}\right) \geq \frac{1}{\tilde{c}} \left\| \operatorname{diag}\left(\begin{bmatrix} v_1^{(\tau)} \\ v_2^{(\tau)} \end{bmatrix} \right) \boldsymbol{W}^{(\tau)} - \boldsymbol{W} \right\|_F^2 .$$

By further upper bounding the right-hand side using the fact that the ReLU activation is 1-Lipschitz, we have $\mathcal{L}\left(\boldsymbol{\theta}^{(T)}\right) \leq \left\| \operatorname{diag}\left(\begin{bmatrix} v_1^{(T)} \\ v_2^{(T)} \end{bmatrix} \right) \boldsymbol{W}^{(T)} - \boldsymbol{W}^* \right\|_F^2$. Combining these yields:

$$\left\| \operatorname{diag}\left(\begin{bmatrix} v_1^{(\tau)} \\ v_2^{(\tau)} \end{bmatrix} \right) \boldsymbol{W}^{(\tau)} - \boldsymbol{W} \right\|_F^2 \leq \tilde{c}\left(1 - c\bar{\mu}\right)^{(\tau - T)} \left\| \operatorname{diag}\left(\begin{bmatrix} v_1^{(T)} \\ v_2^{(T)} \end{bmatrix} \right) \boldsymbol{W}^{(T)} - \boldsymbol{W} \right\|_F^2$$

for some numerical constant $\tilde{c} > 0$. This shows geometric convergence of the GD iterates, completing the proof of Theorem 2.

Appendix E. Additional Experimental Results

E.1. Pairing-up Behavior for $r \geq 3$

In this section, we present additional results on the pairing behavior of $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$ for different values of r. Although our theoretical analysis is limited to the scalar output setting, for our experiments we also consider multi-dimensional outputs. We only consider the case where the model is exactly parameterized i.e. $k = 2r$. We first show that an interesting pattern arises if both the inner and outer layers of the neural network are initialized sufficiently small.

For visualization purposes in Figure 6, we pick $r = 3$ and $k = 6$. As for the target function, we pick $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ to be $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ respectively which correspond to the standard basis vectors in $\mathbb{R}^d$. We plot the trajectory of both the inner and outer layer weights of the network across iterations and observe a peculiar pattern in both $\boldsymbol{v}_i$'s and $\boldsymbol{w}_i$'s. At convergence, weights can be grouped into pairs such that one of the weights is approximately negative of the other. As a concrete example, in Figure 6, we observe that $\boldsymbol{v}_3^{(\infty)} \approx -\boldsymbol{v}_4^{(\infty)}$, $\boldsymbol{v}_1^{(\infty)} \approx -\boldsymbol{v}_5^{(\infty)}$, and $\boldsymbol{v}_2^{(\infty)} \approx -\boldsymbol{v}_6^{(\infty)}$ which also holds similarly for $\boldsymbol{w}_i$'s as well. This suggests that after a permutation of the hidden units, we get

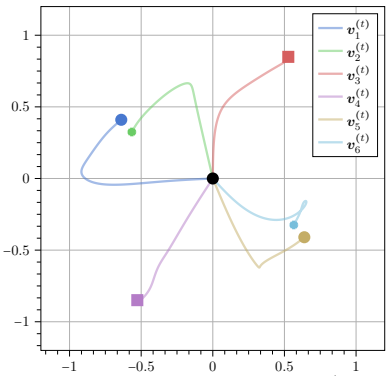
$$\boldsymbol{V}^{(\infty)} \approx \left[\boldsymbol{I}_r, -\boldsymbol{I}_r\right]^T \widetilde{\boldsymbol{V}}, \quad \boldsymbol{W}^{(\infty)} \approx \left[\boldsymbol{I}_r, -\boldsymbol{I}_r\right]^T \widetilde{\boldsymbol{W}}.$$

which can be considered as a natural extension to the $\boldsymbol{v}_i = \pm 1$ pattern in the single output setting.

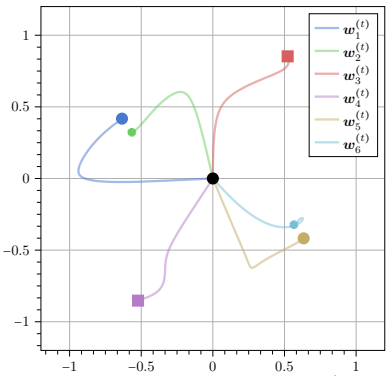
Beyond the $r = 3$ case, we illustrate the same behavior for $r = 5$ in Figure 7 and for $r = 10$ in Figure 8. While we also observe the pairing for $r > 10$, we omit those results here for visual clarity. In general, we note that the weights at convergence (indicated with *star* symbol in Figures 7 and 8) can be grouped into r pairs such that one of the weights is approximately negative of the other. To aid with detecting the pairs visually, we draw the line determined by each pair with dashed lines.

Appendix F. Related Work

There is a large body of work on developing global convergence guarantees for nonconvex problems. We review this literature and compare the differences with the setting discussed in this paper.



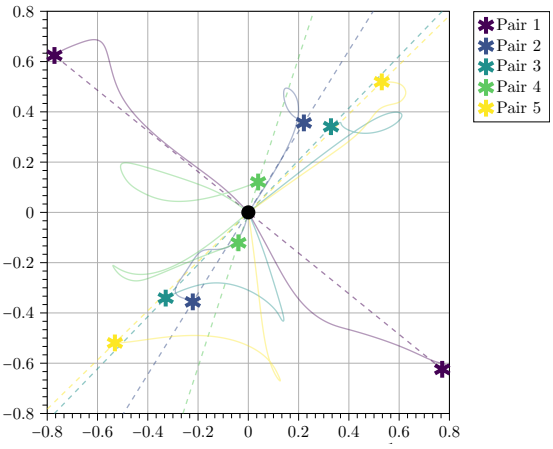
(a) Trajectory of $\boldsymbol{v}_i$'s and their pairing behavior.



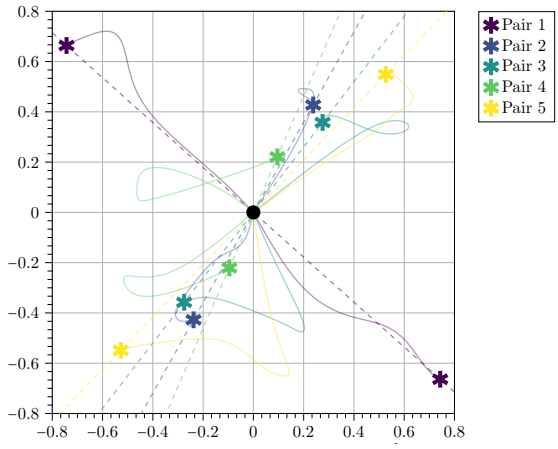
(b) Trajectory of $\boldsymbol{w}_i$'s and their pairing behavior.

Figure 6: **Pairing pattern in multi-dimensional setting.** We train the network from small initialization when exactly parameterized ($k = 6$ and $r = 3$). On left (a), we depict the trajectories of individual weights in the outer layer ($\boldsymbol{v}_i$'s) across iterations. We observe that the weights at convergence can be grouped into three pairs such that one of the weights is approximately negative of the other. For instance, we observe that $\boldsymbol{v}_3^{(\infty)} \approx -\boldsymbol{v}_4^{(\infty)}$. Which neurons end up pairing with each other is indicated by the usage of same symbol (square, circle, etc.). A similar pairing is observed for the inner layer weights as well (b). While these vectors all lie in a higher dimensional space, we pick an arbitrary two dimensional axis to plot them in 2D.

Nonconvex low-rank matrix recovery: In low-rank matrix recovery, numerous studies have shown that nonconvex gradient descent, when initiated with spectral initialization, can effectively solve low-rank reconstruction problems across various domains. This includes phase retrieval (Candès et al., 2015; Chen and Candès, 2017; Ma et al., 2020), matrix sensing Tu et al. (2016), blind deconvolution Li et al. (2019); Ling and Strohmer (2019), and matrix completion Chen et al. (2020). In practice, random initialization is frequently employed instead of specialized spectral initialization methods. As a result, more recent literature Sun et al. (2018); Ge et al. (2016); Zhang et al. (2019), have turned to analyzing the loss landscape. These studies demonstrate that, despite their nonconvex nature, these loss landscapes remain well-behaved under certain assumptions. Specifically, they contain no spurious local minima (i.e., all minimizers are global minima), and saddle points exhibit a strict direction of negative curvature (also known as strict saddle points) Sun et al. (2015). Then specialized truncation or saddle escaping algorithms such as trust region, cubic regularization Nesterov and Polyak (2006); Nocedal and Wright (2006) or noisy (stochastic) gradient-based methods Jin et al. (2017a); Ge et al. (2015); Raginsky et al. (2017); Zhang et al. (2017) are deployed to provably find a global optimum. In contrast to the above literature, the landscape of our loss contain non-strict saddle points. Furthermore, we do not seek any modification to the initialization or the GD updates. Indeed, our result holds with moderately small initialization. As mentioned earlier, we are able to establish this result by developing intricate control of the GD updates throughout the trajectory. This trajectory-level perspective (i.e. multi-phase analysis) is also explored in recent works on gradient descent dynamics and implicit bias under large learning rates (Wang et al., 2022, 2023), see also additional prior work (Stöger and Soltanolkotabi, 2021; Soltanolkotabi et al., 2023) on this topic. However, these works focus on matrix factorization and more general nonconvex objectives

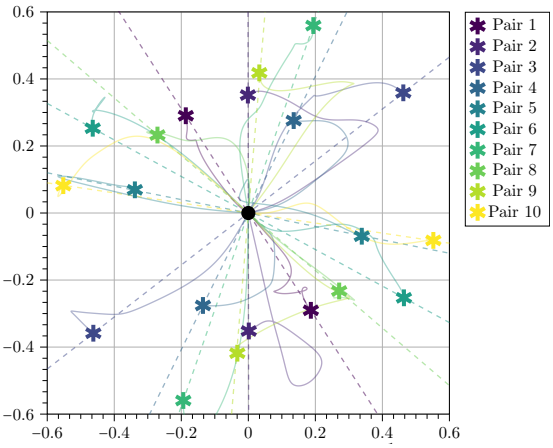


(a) Trajectory of $\boldsymbol{v}_i$'s and their pairing behavior.

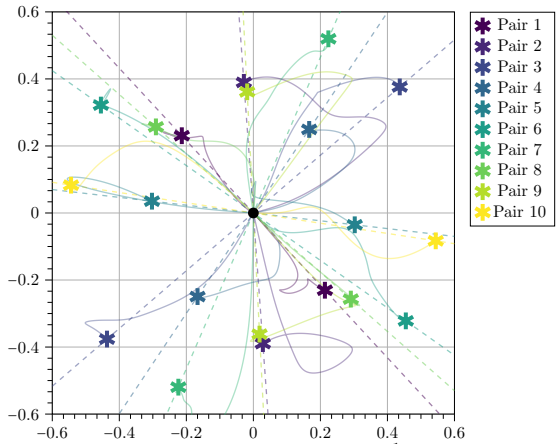


(b) Trajectory of $\boldsymbol{w}_i$'s and their pairing behavior.

Figure 7: **Pairing pattern for $r = 5$.** We train the network from small initialization when exactly parameterized ($k = 10$ and $r = 5$). On left (a), we depict the trajectories of individual weights in the outer layer ($\boldsymbol{v}_i$'s) across iterations. Each pair is indicated by the same color and the dashed line. A similar pairing is observed for the inner layer weights as well (b). While these vectors all lie in a higher dimensional space, we pick an arbitrary two dimensional axis to plot them in 2D.



(a) Trajectory of $\boldsymbol{v}_i$'s and their pairing behavior.



(b) Trajectory of $\boldsymbol{w}_i$'s and their pairing behavior.

Figure 8: **Pairing pattern for $r = 10$.** We train the network from small initialization when exactly parameterized ($k = 20$ and $r = 10$). On left (a), we depict the trajectories of individual weights in the outer layer ($\boldsymbol{v}_i$'s) across iterations. Each pair is indicated by the same color and the dashed line. A similar pairing is observed for the inner layer weights as well (b). While these vectors all lie in a higher dimensional space, we pick an arbitrary two dimensional axis to plot them in 2D.

rather than neural network training.

Gradient-based analysis for neural networks: A recent line of work is concerned with con-

necting the analysis of neural network training with the so-called neural tangent kernel (NTK) Jacot et al. (2018); Oymak and Soltanolkotabi (2019, 2020); Du et al. (2019); Arora et al. (2019). The core idea is that with sufficiently large initialization, a neural network can be approximated by its linearization around the origin. This approximation facilitates linking neural network analysis to the well-established theory of kernel methods. This approach is sometimes referred to as lazy training since, under such initialization, the network parameters remain close to their initial values throughout training. However, some research suggests that NTK-based analysis alone may not fully account for the practical success of neural networks. For instance, Chizat et al. (2019) presents empirical evidence indicating that reducing the initialization size can lead to lower test error. Similarly, Ghorbani et al. (2020) observes a performance gap between neural networks and their NTK counterparts, with the gap widening when the covariance matrix is isotropic. We note that in an NTK analysis the parameters stay close to the initialization which is not the case in our setting. Furthermore, an NTK analysis that relies on linearization can not deal with trajectory analysis that avoids local optima. Indeed, an NTK analysis will not yield the directional convergence established in this paper. So in this sense our result can be viewed as going beyond the lazy training in NTK theory.

Beyond NTK and learning of specific target functions. Recent work carries out analysis of neural networks beyond NTK regime including Damian et al. (2022); Ba et al. (2022); Lee et al. (2024); Xu and Du (2023). Many of these results also focus on learning specific target functions such as ReLUs (Xu and Du, 2023), (Soltanolkotabi, 2017) and polynomials (Damian et al., 2022). These results however typically exclude linear function classes and do not directly involve analysis that requires avoiding bad stationary points explicitly. In fact, many of the existing papers use a pre-processing step or alter the early optimization trajectory to avoid complications arising from the dynamics of learning linear functions (Damian et al., 2022). In contrast, our focus is directly dealing with such intricacies.

Among these papers, perhaps the closest to ours in spirit is (Xu and Du, 2023) which studies the problem of fitting an overparameterized ReLU network to a single ReLU target function with a one dimensional output. Our one-dimensional result can be viewed as a generalization of this work (in particular their exact parametrization result) where the target function has two ReLUs with a particular pattern. This is due to the fact that any linear function of the form $\boldsymbol{a}^T\boldsymbol{x}$ can also be written as a difference of two ReLUs: $v_1\text{ReLU}\left(\frac{1}{v_1}\boldsymbol{a}^T\boldsymbol{x}\right) - v_2\text{ReLU}\left(\frac{-1}{v_2}\boldsymbol{a}^T\boldsymbol{x}\right)$ for any $v_1, v_2 > 0$. The addition of this new ReLU with a negative sign introduces non-strict saddle points and various intricacies in the landscape necessitating a completely different analysis. However, compared to (Xu and Du, 2023) we do not study the effect of overparameterization theoretically. Our empirical results in Section 4 suggest that such an extension may be possible.

We highlight that besides Xu and Du (2023), there are several other works on learning a single neuron Yehudai and Shamir (2022); Vardi et al. (2022); Chistikov et al. (2023) and variants Brutzkus and Globerson (2017). As explained before, such results cannot be used to analyze linear targets due to the interaction terms between positive and negative ReLU neurons. Furthermore, we note that the landscape for fitting a single ReLU is fundamentally different as it contains only a single basin of attraction (albeitt a non-convex one). In contrast, as discussed earlier the landscape in our problem include non-strict saddle points significantly complicating gradient descent analysis.

We would also like to discuss the difference between our work and a few other papers Zhong et al. (2017); Zhang et al. (2018); Zhu et al. (2025); Ren et al. (2025) that have planted one-hidden layer models. These papers differ in at least one of three ways focusing on (1) local analysis,

(2) have sub-optimal sample complexity, and/or (3) assume non-negative outer layer weights. For instance, Zhong et al. (2017) utilize tensor initialization, performing a local analysis rather than a global GD analysis. This local analysis however can not be used to analyze the linear target setting. Indeed, as noted in Remark 4.3 of their work, their analysis requires $\boldsymbol{W}^*$ to be full-rank which does not hold in the linear setting (where the rows of the weight matrix are negatives of each other leading to a minimum singular value is zero). Furthermore, this result also requires resampling the data points at each iteration to ensure convergence of gradient descent where as we use the same samples across all iterations. On a related note, their sample complexity has polynomial dependency on many problem parameters (Theorem 4.2) whereas our proof only requires sample size linear in input dimension d.

Similarly, Zhang et al. (2018) provide a local analysis of GD when the outer layer weights are fixed to be all ones. They also utilize results of Zhong et al. (2017) and share similar limitations in terms of the rank requirement on $\boldsymbol{W}^*$. Thus this result can not be used in the linear target setting even for a local analysis. While they improve the sample complexity of (Zhang et al., 2018) by getting rid of the resampling trick, they still end up with a sample complexity polynomial in width of the network.

Wu et al. (2018) consider the setting when student and teacher networks both have 2 neurons. In particular, when the teachers are *orthogonal*, and the outer weights are all ones; they demonstrated an interesting result that the landscape is benign and all saddles are strict. In contrast, the landscape in our problem include non-strict saddle points significantly complicating gradient descent analysis. In more recent work, Ren et al. (2025) study the complexity of learning a planted model with *orthogonal* planted directions, quadratic activations, and non-negative outer weights. They obtain interesting results on the scaling laws of the MSE loss via a multi-phase analysis. However, this problem setting is substantially different due to the difference between the activation and the orthogonal weights in the planted model that makes the landscape benign per above discussion. More recently, Zhu et al. (2025) also consider learning multiple *orthogonal* ReLU neurons in a teacher-student framework with outer layer weights fixed to all ones. As just discussed, having orthogonal teacher weights leads to a much more benign landscape. Moreover, assumptions in the aforementioned works strictly exclude the linear target setting, where the outer layer must contain negative coefficients. Furthermore, they impose strong restrictions on the initialization. Specifically, they look at the convergence after "weak alignment" where for each student neuron there exists only one teacher neuron that is not near perpendicular. Our results on the other hand can handle random initializations where student neurons *could* be perpendicular to the target direction. That said, their analysis can handle over-parametrization ($k \gg k^*$) and teacher networks with more than 2 neurons.

In recent and independent work, Boursier and Flammarion (2025) also consider the problem of learning linear target functions. The authors demonstrate an interesting result: despite over-parametrization, the sum of positive (resp. negative) neurons aligns with the OLS estimator obtained from the "positive" (resp. negative) subset of the data. To prove this, the authors impose heavy restrictions on the data distribution (in particular, Conditions 3 and 4 in their paper) to essentially align the data with the target direction and avoid changes in the activation cone. We quote the authors:

"However, item 3 is quite restrictive: it is needed to ensure that the volume of the activation cone containing β^* does not vanish when $n \to \infty$. A similar assumption is considered by Chistikov et al. (2023); Tsoy and Konstantinov (2024), for similar reasons. Additionally, Condition 4 ensures that

$\mathbb{E}_x[xx^T]\beta^*$ and β^* are in the same activation cone. This assumption allows the training dynamics to remain within a single cone after the early alignment phase, significantly simplifying our analysis."

In contrast, we demonstrate feature learning in the linear target setting by performing a full characterization of GD dynamics with a generic data distribution and initialization without any of the restrictive assumptions mentioned above.

Finally, we note that recent interesting works by Shevchenko et al. (2022) and Kögler et al. (2024) provide a complementary perspective on nonlinear autoencoders, with the philosophical goal of understanding what happens when the autoencoder departs from PCA-like or linear-function learning. To isolate this genuinely nonlinear regime, these works study sign activations, and more generally odd activations, and characterize the optima of the corresponding population loss. They further show that these optima are achieved by suitable gradient-based procedures, including gradient flow with tied encoder/decoder weights and a projected/alternating gradient method in which the decoder is optimized to completion at each update of the encoder. In contrast, our setting studies standard empirical training of a ReLU network for a planted linear supervised target, where both layers are optimized simultaneously throughout gradient descent and the main challenge is controlling the full finite-sample trajectory through non-strict saddle regions.