# Can Implicit Bias Imply Adversarial Robustness?

**Hancheng Min** [1] **René Vidal** [1]

## Abstract

The implicit bias of gradient-based training algorithms has been considered mostly beneficial as it leads to trained networks that often generalize well. However, Frei et al. (2023) show that such implicit bias can harm adversarial robustness. Specifically, they show that if the data consists of clusters with small inter-cluster correlation, a shallow (two-layer) ReLU network trained by gradient flow generalizes well, but it is not robust to adversarial attacks of small radius. Moreover, this phenomenon occurs despite the existence of a much more robust classifier that can be explicitly constructed from a shallow network. In this paper, we extend recent analyses of neuron alignment to show that a shallow network with a polynomial ReLU activation (pReLU) trained by gradient flow not only generalizes well but is also robust to adversarial attacks. Our results highlight the importance of the interplay between data structure and architecture design in the implicit bias and robustness of trained networks.

## 1. Introduction

Behind the success of deep neural networks in many application domains lies their vulnerability to *adversarial attacks*, i.e., small and human-imperceptible perturbations to the input data. Such a phenomenon was observed in the seminal paper of Szegedy et al. (2014) and has motivated a large body of work on building defenses against such attacks (Shafahi et al., 2019; Papernot et al., 2016; Wong et al., 2019; Guo et al., 2018; Cohen et al., 2019; Levine & Feizi, 2020; Yang et al., 2020; Sulam et al., 2020; Kinfu & Vidal, 2022).

However, many defense strategies have been shown to fail against new adaptive attacks (Athalye et al., 2018; Carlini et al., 2019), and understanding these failures seems to be a fundamental challenge. For example, Fawzi et al. (2018); Dohmatob (2019); Shafahi et al. (2018) show the non-existence of robust classifiers for certain data distributions. Recently, Pal et al. (2023) show that having a data distribution that concentrates within a small-volume subset of the ambient space is necessary for the existence of a robust classifier. These results highlight the importance of understanding and exploiting data structure in the process of finding classifiers with certified robustness, yet almost none of the existing defense strategies do so.

Besides data distribution, additional issues arise from the training algorithms. For example, Vardi et al. (2022) and Frei et al. (2023) show that when the data consists of clusters with small inter-cluster correlation, a shallow (two-layer) ReLU network trained by gradient flow generalizes well but fails to be robust against adversarial attacks of small radius, despite the existence of a much more robust classifier that can be explicitly constructed from a shallow network. This study unveils new challenges in the search for robust classifiers: Even if we know robust classifiers exist for certain data distribution, the *implicit bias* from our training algorithm (the choice of network architecture, optimization algorithm, etc.) may prevent us from finding it.

**Paper contributions.** In this paper, we show that under the same setting studied in Frei et al. (2023), the implicit bias of gradient flow that leads to non-robust networks can be altered to favor robust networks by modifying the ReLU activation. Specifically, we consider a data distribution consisting of a mixture of $K$ Gaussians, referred to as *subclasses*, which have small inter-subclass correlation and are grouped into two *superclasses/classes*. When training a two-layer binary classification network, we show, with formal theorems, and with conjectures validated experimentally, that (also illustrated in Figure 1):

- If the activation is a ReLU, neurons (rows of the first layer weight matrix) tend to learn only the average direction of each class, leading to a classifier that generalizes well on clean data, but is vulnerable to an adversarial attack with $l_2$ radius $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$, i.e. the trained network is non-robust with many subclasses. This leads to a new neural alignment perspective on the nonrobustness of ReLU networks identified by Frei et al. (2023).
- If the activation is replaced by a novel polynomial ReLU

[1]University of Pennsylvania, Philadelphia, PA, USA. Correspondence to: Hancheng Min <hanchmin@seas.upenn.edu>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

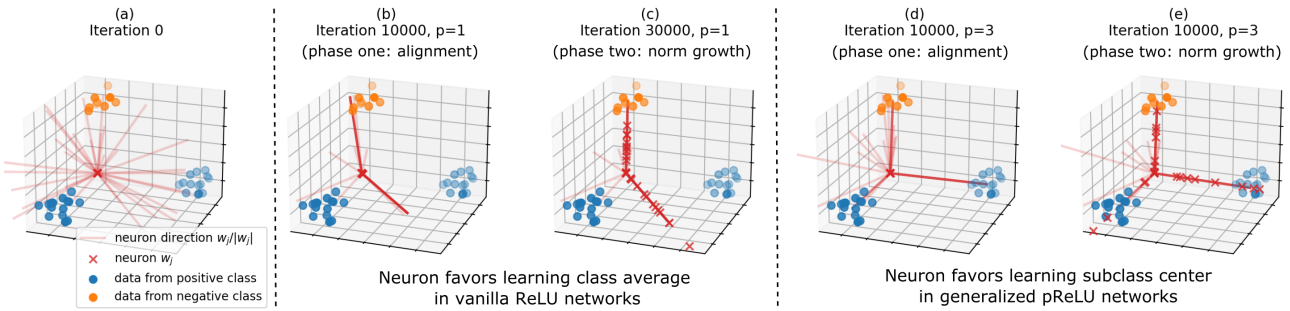


*Figure 1.* Visualizing the training of a pReLU network under small initialization (Details explained in later sections). The dataset has its positive class sampled from two subclasses. **(a)** At initialization, all neurons have small norms and point toward random directions; **When** $p = 1$ (vanilla ReLU network), **(b)** During the alignment phase, the neuron directions are aligned with either of the average class centers $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$; **(c)** During the second phase, neurons keep the alignment with $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$ while growing their norms; **When** $p = 3$, **(d)** neurons learn subclass centers during alignment phase and **(e)** keep the alignment in the second phase. Note: the neurons pointing toward directions other than class/subclass centers are not activated by any data point and have small norms throughout training.

activation, proposed based on recent advances in understanding the neuron alignment in shallow networks, neurons tend to learn the direction of each subclass center, leading to a classifier that generalizes well on clean data and can sustain any adversarial attack with $\mathcal{O}(1)$ radius.

Our analysis **(1)** highlights the importance of the interplay between data structure and network architecture in determining the robustness of the trained network, **(2)** explains how the implicit bias (regularization) of training a ReLU network fails to exploit the data structure and leads to non-robust networks, and **(3)** shows how the issue is resolved by using a polynomial ReLU activation function. Moreover, numerical experiments on real datasets show that shallow networks with our generalized ReLU activation functions are much more robust than those with a ReLU activation.

**Relation to existing analysis on implicit bias of neural networks.** Our discussion is theory-centric. It builds upon the analysis of the implicit bias of training algorithms, but it is significantly different from prior theoretical analyses. The implicit bias of training algorithms has been studied extensively over past years for various architectures, including both linear networks (Saxe et al., 2014; Gunasekar et al., 2017; Ji & Telgarsky, 2019; Woodworth et al., 2020; Min et al., 2021; Stöger & Soltanolkotabi, 2021; Jacot et al., 2021; Wang & Jacot, 2023), and nonlinear networks (Lyu & Li, 2019; Ji & Telgarsky, 2020; Maennel et al., 2018; Chizat & Bach, 2020; Boursier et al., 2022; Min et al., 2024; Wang & Ma, 2023; Frei et al., 2022; 2023; Kumar & Haupt, 2024; Abbe et al., 2023; Tarzanagh et al., 2023). Moreover, this phenomenon has been studied from different perspectives, including max-margin (Lyu & Li, 2019; Chizat & Bach, 2020; Tarzanagh et al., 2023), min-norm (Gunasekar et al., 2017; Min et al., 2021), sparsity/low-rankness (Saxe et al., 2014; Woodworth et al., 2020; Wang & Jacot, 2023; Abbe et al., 2023), and alignment (Maennel et al., 2018; Min et al., 2024; Kumar & Haupt, 2024). While most works consider these implicit biases beneficial for the success of neural networks in practice, and some existing work (Faghri et al., 2021) has even shown that such biases help robustness in the cases of linear regression; few works (Frei et al., 2023; Boursier & Flammarion, 2024) discuss the potential harm caused by such biases. Our work takes one step further by proposing fixes to the harms identified by prior work (Frei et al., 2023), which sheds light on the potential of using deep learning theory to not only understand but also improve neural networks in practice.

**Paper organization.** Our main formal results regarding adversarial robustness will be shown (in Section 3) for explicitly constructed classifiers that can be realized by shallow networks with different activation functions. Then we connect these robustness results to those of actual trained shallow networks by a conjecture that shallow networks, when initialized properly, can learn such constructions via gradient flow training. The rationale behind this conjecture is carefully explained in Section 4, together with a preliminary theoretical analysis that supports this conjecture. Lastly, we empirically verify our conjecture in Section 5 via experiments on synthetic data, and then we conduct numerical experiments on real datasets, showing that our proposed new activation improves the robustness of shallow networks.

**Notation**: We denote the the inner product between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y}$, and the cosine of the angle between them as $\cos(\boldsymbol{x}, \boldsymbol{y}) = \langle \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \rangle$. For an $n \times m$ matrix $\mathbf{A}$, we let $\|\boldsymbol{A}\|$ and $\|\boldsymbol{A}\|_F$ denote the spectral and Frobenius norm of $\boldsymbol{A}$, respectively. We also define $\mathbb{1}_A$ as the indicator for a statement $A$: $\mathbb{1}_A = 1$ if $A$ is true and $\mathbb{1}_A = 0$ otherwise. We also let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)$ denote the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}^2$, and $\mathrm{Unif}(S)$ denote the uniform distribution over a set $S$. Lastly, we let $[N]$ denote the integer set $\{1, \cdots, N\}$.

## 2. Problem Setting

We consider a binary classification problem where within each class, the data consists of subclasses with small inter-subclass correlations, formally defined as follows.

**Data distribution.** We assume $(X, Y, Z) \sim \mathcal{D}$, where a sample $(\boldsymbol{x}, y, z) \in \mathbb{R}^D \times \{-1, 1\} \times \mathbb{N}_+$ drawn from $\mathcal{D}$ consists of the *observed data* and *class label* $(\boldsymbol{x}, y)$, and a latent (unobserved) variable $z$ denoting the *subclass membership*. We denote the marginal distribution of the observed part $(X, Y)$ as $\mathcal{D}_{X,Y}$. In particular, given $1 \le K_1 < K$, we let

- $Z \sim \mathcal{D}_Z = \text{Unif}([K])$; $Y = \mathbb{1}_{\{Z \le K_1\}} - \mathbb{1}_{\{Z > K_1\}}$;
- $(X \mid Z = k) \sim \mathcal{N}\left(\boldsymbol{\mu}_k, \frac{\alpha^2}{D}\boldsymbol{I}\right)$, $\forall 1 \le k \le K$, where $\alpha > 0$ is the *intra-subclass variance*, and $\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$ are *subclass centers* that forms an orthonormal basis of a $K$-dimensional subspace in $\mathbb{R}^D$, where $K < D$.

This data distribution has $K$ subclasses with small inter-subclass correlation if $\alpha$ is small (since the subclass centers are orthonormal to each other), and the observed labels only reveal the superclass/class membership: the first $K_1$ subclasses belong to the positive class ($y = +1$) and the remaining $K_2 := K - K_1$ belong to the negative class ($y = -1$). One such a dataset is depicted in Figure 1.

**Classification task and robustness of a classifier.** The learning task we consider is to find a binary classifier $f(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}$ such that given a randomly drawn $(\boldsymbol{x}, y)$, $\text{sign}(f(\boldsymbol{x}))$ predicts the correct class label $y$, i.e., $f(\boldsymbol{x})y > 0$, with high probability. Moreover, we are interested in the robustness of $f(\boldsymbol{x})$ against additive adversarial attacks/perturbations on $\boldsymbol{x}$ with $l_2$-norm bounded by some constant $r > 0$ (called the *attack radius*). Specifically, given a randomly drawn $(\boldsymbol{x}, y)$ from $\mathcal{D}_{X,Y}$, we would like $\inf_{\|\boldsymbol{d}\|=1} f(\boldsymbol{x}+r\boldsymbol{d})y > 0$ to hold with high probability (formal statement in Section 3) for $r$ as large as possible.

**Gradient flow training.** A candidate classifier can be realized from a neural network $f(\boldsymbol{x}; \boldsymbol{\theta})$ parameterized by its network weights $\boldsymbol{\theta}$. Given a size-$n$ training dataset $\{(\boldsymbol{x}_i, y_i), i = 1, \cdots, n\}$, where each training sample is randomly drawn from $\mathcal{D}_{X,Y}$, we define a loss function

$$\mathcal{L}(\boldsymbol{\theta}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n) = \sum\nolimits_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}))\,, \qquad (1)$$

where $\ell(y, \hat{y}) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is either the exponential loss $\ell(y, \hat{y}) = \exp(-y\hat{y})$ or the logistic loss $\ell(y, \hat{y}) = \log(1 + \exp(-y\hat{y}))$. We train the network using *gradient flow* (GF), $\dot{\boldsymbol{\theta}}(t) \in \partial\mathcal{L}(\boldsymbol{\theta}(t))$, where $\partial\mathcal{L}$ denotes the *Clarke subdifferential* (Clarke, 1990) w.r.t. $\boldsymbol{\theta}$. With proper initialization $\boldsymbol{\theta}(0)$, one expects the trained network weights $\boldsymbol{\theta}(T)$ (for some large $T > 0$) to be close to some minimizer of $\mathcal{L}(\boldsymbol{\theta})$ and $f(\boldsymbol{x}; \boldsymbol{\theta}(T))$ to be a good classifier for $\mathcal{D}_{X,Y}$.

**Shallow networks with pReLU activation.** We specifically study the following shallow (two-layer) network, parametrized by $\boldsymbol{\theta} := \{(\boldsymbol{w}_j, v_j) \in \mathbb{R}^D \times \mathbb{R}, j = 1, \cdots, h\}$, as a candidate classifier:

$$f_p(\boldsymbol{x}; \{\boldsymbol{w}_j, v_j\}_{j=1}^h) = \sum\nolimits_{j=1}^h v_j \frac{[\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle)]^p}{\|\boldsymbol{w}_j\|^{p-1}}\,,$$

(*Shallow pReLU Network*)

where $\sigma(\cdot) = \text{ReLU}(\cdot) = \max\{\cdot, 0\}$ and $p \ge 1$. When $p = 1$, this is exactly a shallow ReLU network. When $p \ge 1$, $f_p$ can be loosely viewed as a shallow network with a polynomial ReLU activation and a special form of weight normalization (Salimans & Kingma, 2016) on each $\boldsymbol{w}_j$. One of the most important reasons for this particular generalization of the ReLU activation is the "extra penalty" on angle separation between the input $\boldsymbol{x}$ and *neurons* $\boldsymbol{w}_j$s. To see this, assume $\|\boldsymbol{x}\| = 1$, then

$$\begin{aligned}\frac{[\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle)]^p}{\|\boldsymbol{w}_j\|^{p-1}} &= \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle) \frac{\cos^{p-1}(\boldsymbol{x}, \boldsymbol{w}_j)\,\|\boldsymbol{w}_j\|^{p-1}}{\|\boldsymbol{w}_j\|^{p-1}} \\ &= \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle) \cos^{p-1}(x, \boldsymbol{w}_j)\,,\end{aligned}$$

that is, for each neuron $\boldsymbol{w}_j$, when compared to ReLU activation ($p = 1$), the post-activation value is much smaller (penalized) if the angle separation between $\boldsymbol{x}$ and $\boldsymbol{w}_j$ is large. When $p > 1$, such penalties, as we will see later, promote the alignment between training data samples and neurons, and result in trained networks that capture well the intrinsic structure of the data. In addition, such generalized ReLU activation makes the function $f_p$ *positively homogeneous* of degree two w.r.t. its parameters $\boldsymbol{\theta}$, i.e., $f_p(\boldsymbol{x}; \gamma\boldsymbol{\theta}) = \gamma^2 f_p(\boldsymbol{x}; \boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ and $\gamma > 0$. Many existing analyses (Du et al., 2018; Lyu et al., 2021; Kumar & Haupt, 2024) on positively homogeneous networks apply to the generalized pReLU networks. Lastly, it is worth noting that the concept of promoting alignment between data points and neurons via cosine penalty has been explored for improving interpretability (Böhle et al., 2022) of the trained network.

## 3. Main Results on Adversarial Robustness

This section considers two distinct classifiers for $\mathcal{D}_{X,Y}$:

$$F^{(p)}(\boldsymbol{x}) = \sum_{k=1}^{K_1} \sigma^p(\langle \boldsymbol{\mu}_k, \boldsymbol{x} \rangle) - \sum_{k=K_1+1}^{K} \sigma^p(\langle \boldsymbol{\mu}_k, \boldsymbol{x} \rangle) \quad (2)$$

and

$$F(\boldsymbol{x}) = \sqrt{K_1}\sigma(\langle \bar{\boldsymbol{\mu}}_+, \boldsymbol{x} \rangle) - \sqrt{K_2}\sigma(\langle \bar{\boldsymbol{\mu}}_-, \boldsymbol{x} \rangle)\,, \qquad (3)$$

where $\bar{\boldsymbol{\mu}}_+ = \frac{1}{\sqrt{K_1}} \sum_{k=1}^{K_1} \boldsymbol{\mu}_k$ and $\bar{\boldsymbol{\mu}}_- = \frac{1}{\sqrt{K_2}} \sum_{k=K_1+1}^{K} \boldsymbol{\mu}_k$ are, respectively, the average direction of the positive and negative subclass centers. Here the average direction is

computed by taking the sum of all $\boldsymbol{\mu}_k$s, then normalize it, thus we have $\bar{\boldsymbol{\mu}}_+ \in \mathbb{S}^{D-1}$ and $\bar{\boldsymbol{\mu}}_- \in \mathbb{S}^{D-1}$.

Based on existing analyses for the implicit bias of gradient-based training, we conjecture that both classifiers can be learned by gradient flow on shallow networks with pReLU activations and proper initialization. Before studying this conjecture, we analyze these two classifiers in more detail.

**Realizability of $F(\boldsymbol{x})$ and $F^{(p)}(\boldsymbol{x})$ by pReLU networks.** We first notice that both $F(\boldsymbol{x})$ and $F^{(p)}(\boldsymbol{x})$ can be realized (non-uniquely) by a pReLU network $f_p(\boldsymbol{x}; \{\boldsymbol{w}_j, v_j\}_{j=1}^h)$ for suitable choices of the weights $\{\boldsymbol{w}_j, v_j\}_{j=1}^h$, as stated in the following claim. (Verifying this claim is straightforward and we leave it to Appendix C).

**Claim.** *The following two statements are true:*

- *When $p = 1$, let $I_+$ and $I_-$ be any two non-empty and disjoint subsets of $[h]$, let $\boldsymbol{\lambda} := (\lambda_j) \in \mathbb{R}_{\geq 0}^h$ be any vector such that $\sum_{j \in I_+} \lambda_j^2 = \sqrt{K_1}$ and $\sum_{j \in I_-} \lambda_j^2 = \sqrt{K_2}$, and let*

$$\begin{aligned} &\boldsymbol{w}_j = \lambda_j \bar{\boldsymbol{\mu}}_+,\ v_j = \lambda_j, && \forall j \in I_+\,, \\ &\boldsymbol{w}_j = \lambda_j \bar{\boldsymbol{\mu}}_-,\ v_j = -\lambda_j, && \forall j \in I_-\,, \\ &\boldsymbol{w}_j = 0,\ v_j = 0, && \textit{otherwise}. \end{aligned}$$

  *Then $f_p(x; \{\boldsymbol{w}_j, v_j\}_{j=1}^h) \equiv F(\boldsymbol{x})$.*

- *When $p \geq 1$, let $I_1, \cdots, I_K$ be any $K$ non-empty and disjoint subsets of $[h]$, let $\boldsymbol{\lambda} := (\lambda_j) \in \mathbb{R}_{\geq 0}^h$ be any vector such that $\sum_{j \in I_k} \lambda_j^2 = 1, \forall k \in [K]$, and let*

$$\begin{aligned} &\boldsymbol{w}_j = \lambda_j \boldsymbol{\mu}_k,\ v_j = \lambda_j, && \forall j \in I_k, k \leq K_1\,, \\ &\boldsymbol{w}_j = \lambda_j \boldsymbol{\mu}_k,\ v_j = -\lambda_j, && \forall j \in I_k, k > K_1\,, \\ &\boldsymbol{w}_j = 0,\ v_j = 0 && \textit{otherwise}. \end{aligned}$$

  *Then $f_p(x; \{\boldsymbol{w}_j, v_j\}_{j=1}^h) \equiv F^{(p)}(\boldsymbol{x})$.*

**Remark 1.** *As shown in the claim, there are infinitely many parameterizations of $f_p$ (determined by the choice of subsets of $[h]$ and $\boldsymbol{\lambda}$) that lead to the same function $F(\boldsymbol{x})$ (or $F^{(p)}(\boldsymbol{x})$), due to the* symmetry *in the network weights. That is, the resulting network represents the same function if one re-indexes (permutes) the neurons, or if one does the following rescaling on some weights $(\boldsymbol{w}_j, v_j) \to (\gamma \boldsymbol{w}_j, v_j/\gamma)$ for some $j \in [j]$ and some $\gamma > 0$.*

Notice that the classifier $F$ corresponds to a vanilla shallow ReLU network whose non-zero neurons are either aligned with $\bar{\boldsymbol{\mu}}_+$, the average direction of the positive subclass centers, or $\bar{\boldsymbol{\mu}}_-$, the negative counterpart. On the other hand, $F^{(p)}$ corresponds to a pReLU network whose non-zero neurons are aligned with one of the subclass centers. As we will see soon, both $F(\boldsymbol{x})$ and $F^{(p)}(\boldsymbol{x})$ achieve high prediction accuracy on clean samples $(\boldsymbol{x}, y)$ from $D_{X,Y}$, i.e., learning subclass centers is not necessary for good generalization. However, we will also show that $F^{(p)}$ is much more robust against adversarial perturbation than $F$. This is why we argue that *learning subclass centers significantly improve robustness*.

**Generalization and Robustness of $F(\boldsymbol{x})$ and $F^{(p)}(\boldsymbol{x})$.** Having established that both $F$ and $F^{(p)}$ are realizable by a pReLU network $f_p$, we now study the generalization performance and robustness of $F$ and $F^{(p)}$. Under the conjecture that both $F$ and $F^{(p)}$ can be learned by gradient flow training on a pReLU network $f_p$, we expect that such generalization and robustness properties extend to $f_p$.

The following result states that both $F$ and $F^{(p)}$ achieve good generalization performance on $\mathcal{D}_{X,Y}$ when the dimension $D$ of the input data $X$ is sufficiently large and the number of subclasses $K$ is not too large.

**Proposition 1** (Generalization on clean data)**.** *Given a test sample $(\boldsymbol{x}, y) \sim \mathcal{D}_{X,Y}$, and classifiers $F$ and $F^{(p)}$, $p \geq 1$, defined in* (3) *and* (2)*, resp., there is a constant $C$ such that*

$$\mathbb{P}\left(F(\boldsymbol{x})y > 0\right) \geq 1 - 4\exp\left(-\frac{CD}{4\alpha^2 K}\right),$$
$$\mathbb{P}\left(F^{(p)}(\boldsymbol{x})y > 0\right) \geq 1 - 2(K+1)\exp\left(-\frac{CD}{\alpha^2 K^2}\right).$$

However, the following theorem shows a significant difference between $F$ and $F^{(p)}$ regarding their adversarial robustness, when the number of subclasses $K$ is large.

**Theorem 1** ($l_2$-adversarial robustness)**.** *Given a test sample $(\boldsymbol{x}, y) \sim \mathcal{D}_{X,Y}$, and classifiers $F$ and $F^{(p)}$, defined in* (3) *and* (2)*, resp., the following statement is true for the same constant $C$ in Proposition 1:*

- *(Non-robustness of $F$ against $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$-attack) Let $d_0 := \frac{\sqrt{K_1}\bar{\boldsymbol{\mu}}_+ - \sqrt{K_2}\bar{\boldsymbol{\mu}}_-}{\sqrt{K}} \in \mathbb{S}^{D-1}$. Then for any $\rho \geq 0$,*

$$\mathbb{P}\left(F\left(\boldsymbol{x} - \frac{y(1+\rho)}{\sqrt{K}}\boldsymbol{d}_0\right)y > 0\right) \leq 2\exp\left(-\frac{CD\rho^2}{K\alpha^2}\right).$$

- *(Robustness of $F^{(p)}$ against $\mathcal{O}(1)$-attack) Let $p \geq 2$. Then for any $0 \leq \delta \leq \sqrt{2}$,*

$$\mathbb{P}\left(\min_{\|\boldsymbol{d}\| \leq 1} F^{(p)}\left(\boldsymbol{x} + \frac{\sqrt{2}-\delta}{2}\boldsymbol{d}\right)y > 0\right) \geq 1 - 2(K+1)\exp\left(-\frac{CD\delta^2}{2K^2\alpha^2}\right).$$

We refer the reader to Appendix C for the proofs for Proposition 1 and Theorem 1. Our theoretical results should be understood in the high-dimensional or low-intra-class-variance regime so that $\frac{D}{\alpha^2} \gg K^2 \log K$. On the one hand,

Theorem 1 shows that $F$, a classifier that corresponds to shallow ReLU networks whose neurons are aligned with either $\bar{\boldsymbol{\mu}}_+$ or $\bar{\boldsymbol{\mu}}_-$, is vulnerable to an adversarial attack of radius $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$, which diminishes as the number of subclasses grows. On the other hand, Theorem 1 shows that $F^{(p)}$, $p \geq 2$, a classifier that corresponds to shallow pReLU networks whose neurons are aligned with one of the subclass centers, is $\mathcal{O}(1)$-robust against adversarial attacks.

**Remark 2.** *Complementary results to Theorem 1 in Appendix C.4 show that $F$ is robust against any attack with $l_2$-norm slightly smaller than $\frac{1}{\sqrt{K}}$, and $F^{(p)}$ is not robust to some attack with $l_2$-norm slightly larger than $\frac{\sqrt{2}}{2}$. Therefore, $\frac{1}{\sqrt{K}}$ and $\frac{\sqrt{2}}{2}$ are the "critical" attack radius for $F$ and $F^{(p)}$, resp. We conjecture this is also the case for the associated trained networks, which we verify in Section 5.1.*

**Conjecture on the outcome of gradient flow training.** We now study the conjecture that both classifiers can be learned by gradient flow on a shallow network with pReLU activations and proper initialization. Specifically, we pose the following:

**Conjecture 1.** *Suppose that the intra-subclass variance $\alpha > 0$ is sufficiently small, that one has a training dataset of sufficiently large size, and that we run gradient flow training on $f_p(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{\theta} = \{\boldsymbol{w}_j, v_j\}_{j=1}^h$ of sufficiently large width $h$ for sufficiently long time $T$, starting from random initialization of the weights with a sufficiently small initialization scale. If $p = 1$, then we have $\inf_{c>0}\sup_{\boldsymbol{x}\in\mathbb{S}^{D-1}} |cf_p(\boldsymbol{x};\boldsymbol{\theta}(T)) - F(\boldsymbol{x})| \ll 1$; If $p \in [3, \bar{p})$ for some $\bar{p} > 3$, then we we have $\inf_{c>0}\sup_{\boldsymbol{x}\in\mathbb{S}^{D-1}} |cf_p(\boldsymbol{x};\boldsymbol{\theta}(T)) - F^{(p)}(\boldsymbol{x})| \ll 1$.*

Our conjecture states that with proper initialization and sufficiently long training time, gradient flow finds a network that is, up to a scaling factor $c > 0$[1], close to $F$ in $l_\infty$-distance over $\mathbb{S}^{D-1}$, if $p = 1$. That is, when training a vanilla ReLU network, the neurons learn average directions $\bar{\boldsymbol{\mu}}_+$, and $\bar{\boldsymbol{\mu}}_-$ of subclass centers. However, when $p \geq 3$, gradient flow finds a network close to $F^{(p)}$, i.e., the neurons learn individual subclass centers. Note that $p$ can not be too large, as the post activation $\frac{[\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j\rangle)]^p}{\|\boldsymbol{w}_j\|^{p-1}}$ converges to $\mathbb{1}_{\cos(\boldsymbol{x},\boldsymbol{w}_j)=1} \cdot \langle \boldsymbol{x}, \boldsymbol{w}_j\rangle$ when $p$ grows ($\|\boldsymbol{x}\| = 1$), effectively zeroing out post activation values almost everywhere and also the gradient, staggering gradient flow training.

Given Conjecture 1, Theorem 1 can be used to infer the robustness of the networks $f_p(\boldsymbol{x};\boldsymbol{\theta}(T))$ obtained from gradient flow training with small initialization. When training a vanilla ReLU network in this regime, we expect the trained network to be close to $F$, thus non-robust against $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$-attacks. When training a pReLU network with $p \geq 3$, we expect the trained network to be close to $F^{(p)}$, which is more robust than its ReLU counterpart.

Our conjecture is based on existing work on the implicit bias of gradient flow on shallow networks with small initialization, and we carefully explain the rationale behind it in Section 4. However, formally showing such convergence results is challenging as the data distribution $\mathcal{D}_{X,Y}$ considered here is more complicated than those for which convergence results can be successfully shown (Boursier et al., 2022; Min et al., 2024; Chistikov et al., 2023; Wang & Ma, 2023). Instead, we provide a preliminary analysis of this conjecture. Moreover, in Section 5, we provide numerical evidence that supports our conjecture.

**Comparison with prior work.** Frei et al. (2023) consider a similar setting as ours with three minor differences. Firstly, their data distribution differs from ours by a scaling factor of $\sqrt{D}$ on the input data. Second, they allow for tiny correlations between the two subclass centers. Third, they consider shallow ReLU networks with bias. They show that any network trained by gradient flow is non-robust against $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ attacks, which covers any initialization, but their analysis does not characterize what classifier the trained network corresponds to. While we consider specifically small initialization, we can at least conjecture, and numerically verify, what the corresponding classifier is at the end of the training. In addition, Frei et al. (2023) show the existence of $\mathcal{O}(1)$-robust ReLU network if neurons are aligned with subclasses and there is some carefully chosen bias. While it already sheds light on the need for learning subclasses, their proposed network can not be found by gradient flow. On the contrary, our proposed robust $F^{(p)}$ can be obtained by gradient flow.

## 4. Discussion on Gradient Flow Training

In Section 3, we conjectured that under gradient flow starting from a small initialization, a vanilla ReLU network favors learning average directions $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$ of subclass centers, while a pReLU network favors learning every subclass centers $\boldsymbol{\mu}_k$, $k \in [K]$, resulting in a significant difference between these two networks in terms of adversarial robustness. In this section, we provide a theoretical explanation of such alignment preferences in the scope of implicit bias of gradient flow training with small initialization (Maennel et al., 2018; Boursier & Flammarion, 2024). We remind the reader that gradient flow training is described in Section 2.

[1] Since $f_p(x;\boldsymbol{\theta})$ is positively homogeneous of degree two w.r.t. its parameter $\boldsymbol{\theta}$, $\|\boldsymbol{\theta}(t)\| \to \infty$ as training time $t \to \infty$ (Lyu & Li, 2019; Ji & Telgarsky, 2020), the output of $f_p(\boldsymbol{x};\boldsymbol{\theta}(T))$ can never match that of $F(\boldsymbol{x})$ or $F^{(p)}(\boldsymbol{x})$ without a proper choice of scaling factor $c > 0$. However, we note, that such a scaling factor will not change the prediction $\text{sign}(f_p(\boldsymbol{x};\boldsymbol{\theta}(T)))$.

### 4.1. Gradient flow with small initialization

We first briefly describe the training trajectory of gradient flow on shallow networks with a small initialization. With a sufficiently small initialization scale (to be defined later), the gradient flow training is split into two phases with distinct dynamic behaviors of the neurons. The first phase is often referred to as the initial/early *alignment phase* (Boursier & Flammarion, 2024; Min et al., 2024), during which the neurons keep small norms while changing their directions towards one of the *extremal vectors* (Maennel et al., 2018), which are jointly determined by the training dataset and the activation function. The second phase is often referred to as the fitting/convergence phase. Within the fitting phase, neurons keep a good alignment with the extremal vectors and start to grow their norms, and the loss keeps decreasing until convergence, upon which one obtains a trained network whose dominant neurons (those with large norms) are all aligned with one of the extremal vectors.

Notably, the neurons favor different extremal vectors depending on the activation function, precisely depicted by Figure 1. With the same dataset and with the same initialization of the weights, pReLU activation makes a significant difference in neuron alignment: When $p = 1$ (vanilla ReLU network), the average class centers $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$ are those extremal vectors "attracting" neurons during the alignment phase, leading to a trained ReLU network that has effectively two neurons (one aligned with $\bar{\boldsymbol{\mu}}_+$ and another with $\bar{\boldsymbol{\mu}}_-$) at the end of the training. However, when $p = 3$, the subclass centers $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k$ become extremal vectors that are "attracting" neurons , the resulting pReLU networks successfully learn every subclass center, which, we have argued in Section 3, substantially improves the robustness (over vanilla ReLU net) against adversarial attack. While the case visualized in Figure 1 is a simple one in 3-dimension with 3 subclasses, we will provide experiments in higher dimension and with more subclasses in Section 5.

Despite the seemingly simple dynamic behavior of the neurons, the rigorous theoretical analysis is much more challenging due to the complicated dependence (Maennel et al., 2018; Boursier et al., 2022; Wang & Ma, 2023; Kumar & Haupt, 2024) of the extremal vectors on the dataset and the activation function. Hence prior works (Boursier et al., 2022; Min et al., 2024; Chistikov et al., 2023; Wang & Ma, 2023) assume simple training datasets and their analyses are for ReLU activation. Here, we are faced with a relatively more complex data distribution that generates our dataset, and at the same time deals with a pReLU activation, thus a full theoretical analysis of the convergence, that would prove our Conjecture 1, still has many challenges and deserves a careful treatment in a separate future work. For now, we provide some preliminary theoretical analysis of the neural alignment in pReLU networks that supports our conjecture.

### 4.2. Preliminary theoretical analysis on neuron alignment in pReLU networks

We first make the following two simplifying assumptions:

**Small and balanced initialization.** First, we obtain i.i.d. samples $w_{j0}, j = 1, \cdots, h$ drawn from some random distribution such that almost surely $\|w_{j0}\| \leq M, \forall j$ for some $M > 0$[2], and then and then initialize the weights as

$$\boldsymbol{w}_j(0) = \epsilon w_{j0},\ \ v_j(0) \in \{\|\boldsymbol{w}_j(0)\|, -\|\boldsymbol{w}_j(0)\|\}, \forall j \in [h]\,. \tag{4}$$

That is, $w_{j0}$ determines the initial direction of the neurons and we use $\epsilon$ to control the initialization scale. This balanced assumption is standard in the analysis of shallow networks with small initialization (Soltanolkotabi et al., 2023; Boursier et al., 2022; Min et al., 2024).

**Simplified training dataset.** The training dataset $\{(\boldsymbol{x}_k, y_k), k = 1, \cdots, K\}$ is the collection of exact subclass centers, together with their class label. That is, $\boldsymbol{x}_k = \boldsymbol{\mu}_k, \quad y_k = \mathbb{1}_{\{k \leq K_1\}} - \mathbb{1}_{\{k > K_1\}}, \forall k \in [K]$. Similar datasets with orthogonality among data points $\boldsymbol{x}_k$s are studied in Boursier et al. (2022) for ReLU networks.

**Neuron alignment in pReLU networks.** The key to the theoretical understanding of neuron alignment is the following lemma.

**Lemma 1.** *Given some initialization from* (4)*, if* $\epsilon = \mathcal{O}(\frac{1}{\sqrt{h}})$*, then there exists* $T = \boldsymbol{\theta}(\frac{1}{K}\log\frac{1}{\sqrt{h}\epsilon})$ *such that the trajectory under gradient flow training with the simplified training dataset almost surely satisfies that* $\forall t \leq T$*,*

$$\max_j \left\| \frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \mathrm{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)}\boldsymbol{x}^{(p)}(\boldsymbol{w}_j(t)) \right\| = \mathcal{O}\left(\epsilon k\sqrt{h}\right)\,,$$

*where* $\mathcal{P}^{\perp}_{\boldsymbol{w}} = I - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}$ *and*

$$\boldsymbol{x}^{(p)}(w) = \sum\nolimits_{k=1}^{K} \gamma_k(w) y_k \boldsymbol{x}_k \cdot p[\cos(\boldsymbol{x}_k, w)]^{p-1}, \tag{5}$$

*with* $\gamma_k(\boldsymbol{w})$ *being a subgradient of* $\sigma(z)$ *at* $z = \langle \boldsymbol{x}_k, \boldsymbol{w}\rangle$ *when* $p = 1$ *and* $\gamma_k(\boldsymbol{w}) = \mathbb{1}_{\cos(\boldsymbol{x}_k, w) \geq 0}$ *when* $p > 1$.

Lemma 1 suggests that during the alignment phase $t \leq T$, one have the following approximation

$$\frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} \simeq \mathrm{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t))\,, \tag{6}$$

[2]For example, if we sample every entries of $w_{j0}$ by a uniform distribution over $\left(-\frac{1}{\sqrt{D}}, -\frac{1}{\sqrt{D}}\right)$.

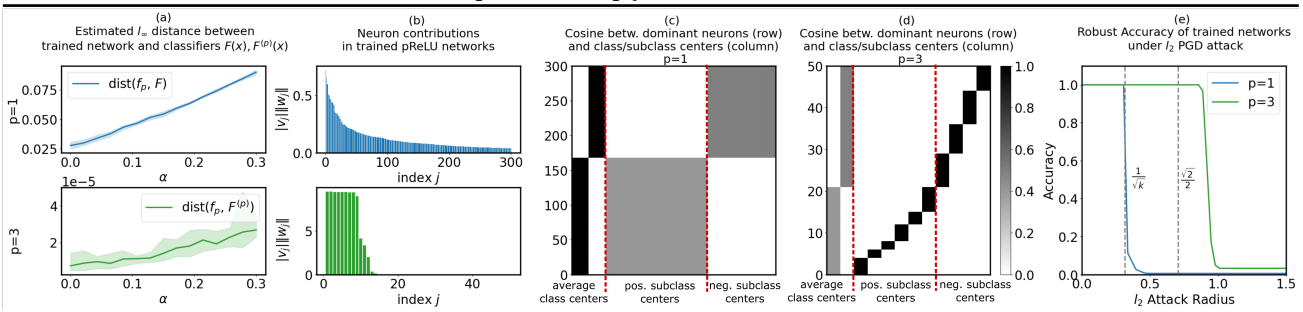


*Figure 2.* Numerical experiment ($K = 10, K_1 = 6$) validates Conjecture 1. **(a)** We train pReLU networks using SGD with small initialization, then estimate the distance $\text{dist}(f_p, F)$ between the trained network $f_p$ and the classifier $F$, when $p = 1$ (top plot); When $p = 3$, we estimate $\text{dist}(f_p, F^{(p)})$ instead (bottom plot). The training is done under different choices of intra-subclass variance $\alpha$ and repeated 10 times per $\alpha$; the Solid line shows the average and the shade denotes the region between max and min values. **(b)** Given a trained network obtained from an instance of this training ($\alpha = 0.1$), we reorder the neurons w.r.t. their contributions $|v_j|\|\boldsymbol{w}_j\|$ and then plot the contributions in a bar plot; **(c)(d)** For neurons with large contributions, we plot a colormap, with each pixel represents some $\cos(\boldsymbol{w}_j, \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ could be either average class centers $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$ or subclass centers $\boldsymbol{\mu}_k, k \in [K]$. Note: For visibility, the neurons are reordered again so that neurons aligned with the same $\boldsymbol{\mu}$ are grouped together. **(e)** Lastly, we carry out $l_2$ PGD attack on a test dataset and plot the robust accuracy of the trained network under different choices of attack radius.

which essentially shows that when $\boldsymbol{w}_j$ is a *positive neuron* ($\text{sign}(v_j(0)) > 0$), then gradient flow dynamics during alignment phase pushes $\boldsymbol{w}_j/\|\boldsymbol{w}_j\|$ toward the direction of $\boldsymbol{x}^{(p)}(\boldsymbol{w}_j)$. Notably, $\boldsymbol{x}^{(p)}(\boldsymbol{w}_j)$ is a weighted combination of $\boldsymbol{x}_k$s and the mixing weights critically depend on $p$. Roughly speaking, when $p = 1$, $\boldsymbol{x}^{(p)}(\boldsymbol{w}_j)$, weighing equally on $\boldsymbol{x}_k$s that activate $\boldsymbol{w}_j$, are more aligned with $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$, while when $p \geq 3$, $\boldsymbol{x}^{(p)}(\boldsymbol{w}_j)$, weighing more on $\boldsymbol{x}_k$s that are close to $\boldsymbol{w}_j$ in angle, are more aligned with one of the subclass centers, thus by moving toward $\boldsymbol{x}^{(p)}(\boldsymbol{w}_j)$ in direction, the neuron $\boldsymbol{w}_j$ is likely to align with average class centers in the former case, and with subclass centers in the latter case (We illustrate this in an example in Appendix D.1). This trend leads to the following alignment bias.

**Theorem 2** (Alignment bias of positive neurons)**.** *Given some sufficiently small $\delta > 0$ and a fixed choice of $p \geq 1$, then $\exists \epsilon_0 := \epsilon_0(\delta, p) > 0$ such that for any solution of the gradient flow on $f_p(\boldsymbol{x}; \boldsymbol{\theta})$ with the simplified training dataset, starting from some initialization from* (4) *with initialization scale $\epsilon < \epsilon_0$, almost surely we have that at any time $t \leq T = \boldsymbol{\theta}(\frac{1}{n} \log \frac{1}{\sqrt{h}\epsilon})$ and $\forall j$ with $\text{sign}(v_j(0)) > 0$,*

$$\frac{d}{dt}\cos\big(\boldsymbol{w}_j(t), \bar{\boldsymbol{\mu}}_+\big)\Big|_{\cos(\boldsymbol{w}_j(t), \bar{\boldsymbol{\mu}}_+)=1-\delta} \begin{cases} > 0, & \textit{when } p = 1 \\ < 0, & \textit{when } p \geq 3 \end{cases}, \tag{7}$$

*and*

$$\frac{d}{dt}\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_k)\Big|_{\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_k)=1-\delta} > 0, \forall k \leq K_1 \,. \tag{8}$$

This theorem shows how different choice of $p$ alters the neurons' preferences on which directions to align. If we define the neighborhood of some $\boldsymbol{\mu}$ direction as $\mathcal{B}(\boldsymbol{\mu}, \delta) := \{\boldsymbol{z} \in \mathbb{S}^{D-1} : \cos(\boldsymbol{\mu}, \boldsymbol{z}) \geq 1 - \delta\}$, then specifically for a positive neuron: When $p = 1$, then any neuron with $\boldsymbol{w}_j(t_0) \in \mathcal{B}(\bar{\boldsymbol{\mu}}_+, \delta)$ necessarily has $\boldsymbol{w}_j(t) \in \mathcal{B}(\bar{\boldsymbol{\mu}}_+, \delta), \forall t \in [t_0, T]$, i.e. it keeps at least $1 - \delta$ alignment with $\bar{\boldsymbol{\mu}}_+$ until the end of the alignment phase (at time $T$). Therefore, there is a preference of staying close to $\bar{\boldsymbol{\mu}}_+$ for positive neurons. Interestingly, such preference no longer exists when $p \geq 3$. In particular, any positive neuron with $\boldsymbol{w}_j(t_0) \notin \mathcal{B}(\bar{\boldsymbol{\mu}}_+, \delta)$ necessarily has $\boldsymbol{w}_j(t) \notin \mathcal{B}(\bar{\boldsymbol{\mu}}_+, \delta), \forall t \in [t_0, T]$, i.e., any neuron initialized with some angular distance to $\bar{\boldsymbol{\mu}}_+$ will not get any closer to $\bar{\boldsymbol{\mu}}_+$. Instead, the neurons are now more likely to stay close to some subclass centers, as shown in (8). Similar results can be said for negative neurons (whose index $j$ has $\text{sign}(v_j(0)) < 0$): they favor average negative class center $\bar{\boldsymbol{\mu}}_-$ when $p = 1$ and favor subclass directions when $p \geq 3$. We refer the interested readers to Appendix D for a complete Theorem 2 (and its proof) that also includes the statement for negative neurons.

## 5. Numerical Experiments

Our numerical experiments[3] have two parts: First, we conduct experiments to validate Conjecture 1. Then, we provide preliminary experiments on real datasets to highlight the potential of using pReLU activation for obtaining more robust classifiers.

### 5.1. Numerical evidence supports our conjecture

We run SGD (batch size 100 and step size 0.2 for $2 \times 10^5$ epochs) with small initialization (all weights initialized as mean-zero Gaussian with standard deviation $10^{-7}$) to train a pReLU network with $h = 2000$ neurons on a dataset drawn from $\mathcal{D}_{X,Y}$ with $D = 1000$, $K = 10$, and

[3]Code available at `https://github.com/hanchmin/pReLU_ICML24`.

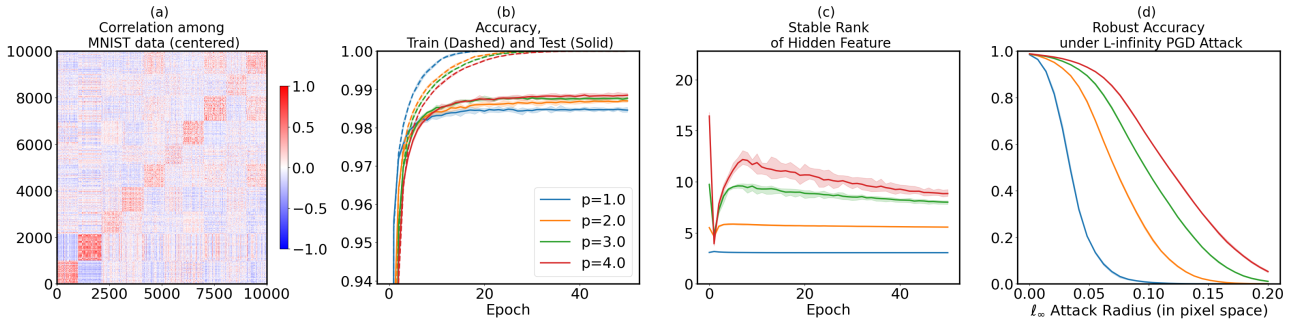


*Figure 3.* Parity prediction on MNIST dataset with pReLU networks. **(a)** We plot the data correlation as a colormap, where each pixel represents some $\cos(x_i, x_j)$ between two centered data $x_i, x_j$ from MNIST training dataset; **(b)** We run Adam with batch size 1000 to train a pReLU network under Kaiming initialization (repeated 10 times), then plot the training/testing accuracy during training for different choice of $p$ (The shade region indicates the range between the minimum and maximum values over 10 randomized runs); **(c)** We stack the hidden post-activation representation of each training sample into a matrix and compute its stable rank, and plot the evolution of this stable rank during training; **(d)** After training for 50 epoch, we carry out APGD $l_\infty$-attack on MNIST test dataset (in pixel space) and plot the robust accuracy of the trained pReLU network under different choice of attack radius.

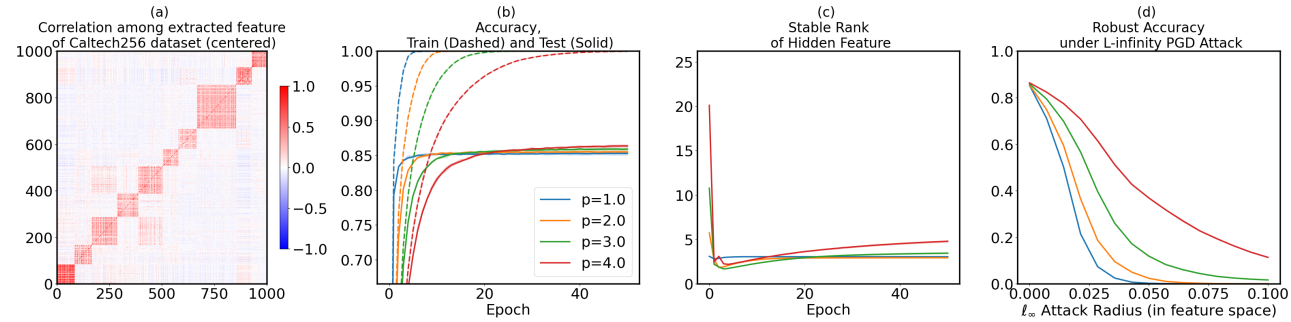


*Figure 4.* Classification on Caltech256 dataset (relabeled into 10 superclasses) with a pre-trained ResNet152 as a fixed feature extractor.

$K_1 = 6$. With different choices of intra-subclass variance $\alpha$, we take the network $f_p$ at the end of the training and estimate (we refer the readers to Appendix A for the estimation algorithm) its distance $\mathrm{dist}(f_p, F) = \inf_{c>0} \sup_{\boldsymbol{x}\in\mathbb{S}^{D-1}} |cf_p(\boldsymbol{x}) - F(\boldsymbol{x})|$ to the classifier $F(\boldsymbol{x})$, or the distance $\mathrm{dist}(f_p, F^{(p)})$ to $F^{(p)}(\boldsymbol{x})$, depending on the value of $p$. As one sees from Figure 2, when $p = 1$, $f_p$ is close to $F$, and the estimated distance slightly increases as $\alpha$ gets larger. Similarly, when $p = 3$, $f_p$ is close to $F^{(p)}$. Furthermore, we investigate the alignment between the dominant neurons in $f_p$ and class/subclass centers. Figure 2 shows that indeed neurons in $f_p$ learn only average class centers $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$ when $p = 1$ while learning every subclass center $\boldsymbol{\mu}_k, k \in [K]$ when $p = 3$. Lastly, as our Theorem 1 predicts, $f_p, p = 3$ is much more robust than the one with $p = 1$. This series of numerical evidence strongly support Conjecture 1 (with additional experiments in Appendix A).

### 5.2. Experiments on real datasets

Although we have shown good alignment between our theory and our numerical experiments in 5.1. The settings largely remain unrealistic. To show the potential value of pReLU networks in practical settings in finding a robust classifier, we now study classification (albeit slightly modified) problems on real datasets.

#### 5.2.1. Parity Classification on MNIST

We first consider training a pReLU network of width $h = 500$ to predict whether an MNIST digit is even or odd. This poses a problem similar to the one studied in our theoretical analyses: each digit is naturally a subclass and they form classes of even digits and odd digits. Moreover, once the data is centered, as shown in Figure 3, two data points of the same digit are likely to have a large positive correlation, and two points of different digits to have a small, or negative correlation, which resembles the our orthogonality assumption among subclass centers. Therefore, with appropriate data preprocessing, we expect the pReLU network to find a more robust classifier when $p$ is large.

**Data preprocessing.** We relabel MNIST data by parity, which leads to a binary classification. Then we center both the training and test set by subtracting off the mean image of all training data and then normalize the residue, resulting in a centered, normalized training/test set [4].

[4] When training pReLU networks, some normalization of the data is required to improve training stability. To see this, notice that the post-activation value for $i$th data scales as $\|x_i\|^p$; When $p$ is large, this term diminishes or explodes depending on where the

**Training pReLU.** We use Kaiming initialization (He et al., 2015) with non-small variance for all the weights and run Adam with cross-entropy loss and with batch size 1000 for 50 epochs and summarize the training results in Figure 3. First of all, as $p$ increases, the trained network becomes more robust against the adversarial $l_\infty$-attack computed from an *adaptive projected gradient ascent* (APGD) algorithm (Croce & Hein, 2020). Interestingly, pReLU with $p > 1$ even has a slight edge over vanilla ReLU net in terms of test accuracy on clean data. Given that the MNIST dataset does not satisfy our data distribution, there is no reason to expect that neurons in pReLU can learn each subclass (in this case, the individual digit) and indeed they cannot. However, we highlight that pReLU empirically is more capable of learning distinct patterns within each superclass/class: We stack the hidden post-activation representation of each training sample into a matrix and compute its stable rank (defined as $\frac{\|\boldsymbol{A}\|_F^2}{\|\boldsymbol{A}\|^2}$ for a matrix $\boldsymbol{A}$, as an approximation of $\mathrm{rank}(\boldsymbol{A})$). From Figure 3, we see that the hidden feature matrix of MNIST obtained from pReLU network has a much larger stable rank than the one from vanilla ReLU net, i.e. the hidden features collapse less when $p$ is large, and we conjecture it to be the reason why pReLU still gains much more adversarial robustness than vanilla ReLU. Theoretically investigating such a phenomenon is an interesting future direction.

**Additional experiments.** To further illustrate the effectiveness of pReLU activation in improving the adversarial robustness of the trained network, we conduct additional experiments, in Appendix B, of training pReLU networks for the original digit classification task on MNIST and test the robustness of the trained network with different types of attacks. Despite that our theoretical results only suggest robustness gain can be achieved under datasets with subclasses, the additional experimental results show that the pReLU networks are still much more robust than their ReLU counterpart even for the original digit classification task, which clearly does not follow our data assumption.

#### 5.2.2. Image classification with pre-trained feature extractor

Our next experiment considers a transfer learning setting where we use some intermediate layer (more precisely, the feature layer before the fully-connected layers) of a pre-trained ResNet152 on ImageNet as a feature extractor and classify the extracted feature from Caltech256 (Griffin et al., 2007) object classification dataset with pReLU networks. The main reason behind this setting is that we expect the extracted feature representation of each class to be clustered, and this is verified in Figure 4.

value $\|x\|$ is smaller or larger than one.

We intend to have a classification task with subclasses, thus we group the original 256 classes in the dataset into 10 superclasses (each superclass has no semantic meaning in this case) and train pReLU networks of width $h = 2000$ that predict the superclass label. Then we obtain the feature representation of the dataset from our feature extractor, center the feature, and normalize, same as we did for the MNIST dataset. The training process is exactly the same as for the MNIST dataset, and we summarize the results in Figure 4. Even for this multi-class classification task, still pReLU achieves higher test accuracy and is more robust when $p$ gets larger.

## 6. Conclusion

By introducing a generalized pReLU activation, we resolve the non-robustness issue, caused by the implicit bias of gradient flow, of ReLU networks when trained on a classification task in the presence of latent subclasses with small inter-subclass correlations. Future work includes formal analyses on neural alignment in pReLU networks as well as more empirical evaluation of pReLU activation for its ability to improve the robustness of neural networks.

## Acknowledgement

The authors thank the support of the NSF-Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning (NSF grant 2031985), and the ONR MURI Program (ONR grant 503405-78051). The authors thank the feedback from anonymous reviewers, which has greatly improved our experimental results. The authors thank Enrique Mallada, Jeremias Sulam, and Ambar Pal for their suggestions and comments at various stages of this research project, and also thank Kyle Poe for carefully reading the manuscript.

## Impact Statement

While this work is mostly theoretical, it tackles the issue of nonrobustness of neural networks trained by gradient-based algorithms, which potentially leads to more robust, reliable, and trustworthy neural networks in many application domains.

## References

Abbe, E., Bengio, S., Boix-Adsera, E., Littwin, E., and Susskind, J. Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems*, 36, 2023.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing de-

fenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Böhle, M., Fritz, M., and Schiele, B. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10329–10338, 2022.

Boursier, E. and Flammarion, N. Early alignment in two-layer networks training is a two-edged sword. *arXiv preprint arXiv:2401.10791*, 2024.

Boursier, E., Pullaud-Vivien, L., and Flammarion, N. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20105–20118, 2022.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Chistikov, D., Englert, M., and Lazic, R. Learning a neuron by a shallow reLU network: Dynamics and implicit bias for correlated inputs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 1305–1338. PMLR, 09–12 Jul 2020.

Clarke, F. H. *Optimization and nonsmooth analysis*. SIAM, 1990.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Dohmatob, E. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*, pp. 1646–1654. PMLR, 2019.

Du, S. S., Hu, W., and Lee, J. D. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Faghri, F., Gowal, S., Vasconcelos, C., Fleet, D. J., Pedregosa, F., and Roux, N. L. Bridging the gap between adversarial robustness and optimization bias. *arXiv preprint arXiv:2102.08868*, 2021.

Fawzi, A., Fawzi, H., and Fawzi, O. Adversarial vulnerability for any classifier. *Advances in neural information processing systems*, 31, 2018.

Frei, S., Vardi, G., Bartlett, P., Srebro, N., and Hu, W. Implicit bias in leaky relu networks trained on high-dimensional data. In *The Eleventh International Conference on Learning Representations*, 2022.

Frei, S., Vardi, G., Bartlett, P., and Srebro, N. The double-edged sword of implicit bias: Generalization vs. robustness in reLU networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. 2007.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6152–6160, 2017.

Guo, C., Rana, M., Cisse, M., and van der Maaten, L. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Jacot, A., Ged, F., Şimşek, B., Hongler, C., and Gabriel, F. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Ji, Z. and Telgarsky, M. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Kinfu, K. A. and Vidal, R. Analysis and extensions of adversarial training for video classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3416–3425, 2022.

Kumar, A. and Haupt, J. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *arXiv preprint arXiv:2402.09226*, 2024.

Levine, A. and Feizi, S. (de) randomized smoothing for certifiable defense against patch attacks. *Advances in Neural Information Processing Systems*, 33:6465–6475, 2020.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

Lyu, K., Li, Z., Wang, R., and Arora, S. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

Maennel, H., Bousquet, O., and Gelly, S. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

Min, H., Tarmoun, S., Vidal, R., and Mallada, E. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7760–7768. PMLR, 18–24 Jul 2021.

Min, H., Mallada, E., and Vidal, R. Early neuron alignment in two-layer relu networks with small initialization. In *International Conference on Learning Representations*, pp. 1–8, 5 2024.

Pal, A., Sulam, J., and Vidal, R. Adversarial examples might be avoidable: The role of data concentration in adversarial robustness. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pp. 582–597. IEEE, 2016.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Saxe, A. M., Mcclelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural network. In *International Conference on Learning Representations*, 2014.

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2018.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Soltanolkotabi, M., Stöger, D., and Xie, C. Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pp. 5140–5142. PMLR, 12–15 Jul 2023.

Stöger, D. and Soltanolkotabi, M. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021.

Sulam, J., Muthukumar, R., and Arora, R. Adversarial robustness of supervised sparse coding. *Advances in neural information processing systems*, 33:2110–2121, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations*, 2014.

Tarzanagh, D. A., Li, Y., Zhang, X., and Oymak, S. Max-margin token selection in attention mechanism. *Advances in Neural Information Processing Systems*, 36, 2023.

Vardi, G., Shamir, O., and Srebro, N. On margin maximization in linear and relu networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022.

Wang, M. and Ma, C. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of reLU networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wang, Z. and Jacot, A. Implicit bias of sgd in $l_\{2\}$-regularized linear dnns: One-way jumps from high to low rank. *arXiv preprint arXiv:2305.16038*, 2023.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Yang, G., Duan, T., Hu, J. E., Salman, H., Razenshteyn, I., and Li, J. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pp. 10693–10705. PMLR, 2020.

# A. Additional Discussion on Validating Conjecture 1

## A.1. Estimating $\mathrm{dist}(f_p, F)$

To estimate the distance $\mathrm{dist}(f_p, F) = \inf_{c>0} \sup_{\mathbb{S}^{D-1}} |cf_p(\boldsymbol{x}) - F(\boldsymbol{x})|$ between a trained network $f_p$ and a classifier $F$ (or $F^{(p)}$), we first pick an $\hat{c} > 0$ that yields the least-square match between $\hat{c}f_p(\boldsymbol{x})$ and $F(\boldsymbol{x})$ over a set of points sampled from $\mathrm{Unif}(\mathbb{S}^{D-1})$. Then given this choice of $\hat{c}$, we run normalized projected gradient ascent on $|\hat{c}f_p(\boldsymbol{x}) - F(\boldsymbol{x})|^2$ starting from an initial $x$ sampled from $\mathrm{Unif}(\mathbb{S}^{D-1})$, and repeat a large number of times, the maximum value of $|\hat{c}f_p(\boldsymbol{x}) - F(\boldsymbol{x})|$ at the end of Normalized PGA over all runs give an estimate of $\mathrm{dist}(f_p, F)$.

**Algorithm 1** Estimating $\mathrm{dist}(f_p, F)$

**Input**: Network $f_p$; Classifier $F$ (or $F^{(p)}$); step size $\alpha$; sample numbers $N_1$, $N_2$;
**Do**:

1. Sample $\boldsymbol{x}, i = 1, \cdots, N_1$ from $\mathrm{Unif}(\mathbb{S}^{D-1})$; $\quad \hat{c} \leftarrow \arg\min_{c>0} \sum_{i=1}^{N_1} |cf_p(\boldsymbol{x}) - F(\boldsymbol{x})|^2$
2. Sample new $\boldsymbol{x}, i = 1, \cdots, N_2$ from $\mathrm{Unif}(\mathbb{S}^{D-1})$;
   $\ell(\cdot) \leftarrow |\hat{c}f_p(\cdot) - F(\cdot)|^2$;
   **For** $i = 1, \cdots, N_2$:
      $\boldsymbol{x}^{(0)} \leftarrow \boldsymbol{x}$;
      **For** $t = 1, \cdots, \mathsf{max_iter}$:
         $\boldsymbol{x}^{(t)} \leftarrow \boldsymbol{x}^{(t-1)} + \alpha \frac{\nabla_x \ell(x^{(t-1)})}{\|\nabla_x \ell(x^{(t-1)})\|}$; `% normalized gradient ascent`
         $\boldsymbol{x}^{(t)} \leftarrow \frac{\boldsymbol{x}^{(t)}}{\|\boldsymbol{x}^{(t)}\|}$; `% projection onto unit sphere`

**Return** $\max_i |\hat{c}f_p(\boldsymbol{x}_i^{(\mathsf{max_iter})}) - F(\boldsymbol{x}_i^{(\mathsf{max_iter})})|$

## A.2. Additional experiment

We conduct the same experiment as in Section 5.1, with more subclasses $K = 20, K_1 = 8$, the results are still well aligned with Conjecture 1. Notably, as $K$ the total number of subclasses increases, the trained ReLU network is more vulnerable to $l_2$ attacks, compared to the case in Section 5.1, while generalized pReLU still is robust to these attacks.

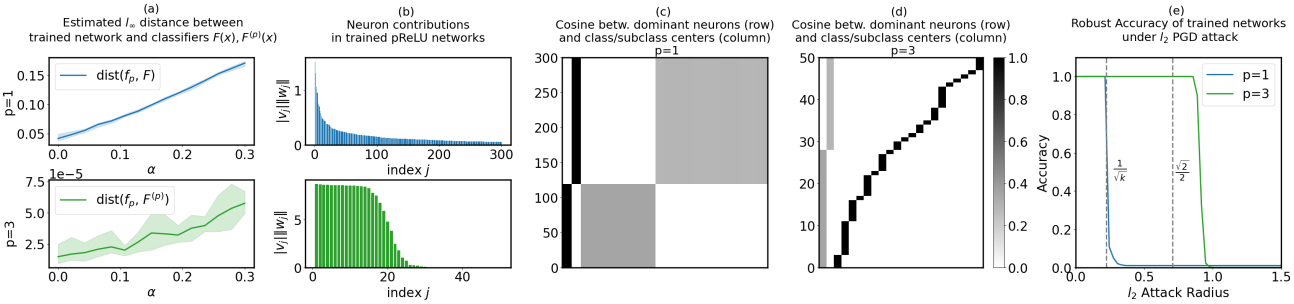


*Figure 5.* Additional Numerical experiment ($K = 20, K_1 = 8$) validates Conjecture 1. **(a)** We train pReLU networks using SGD with small initialization, then estimate the distance $\mathrm{dist}(f_p, F)$ between the trained network $f_p$ and the classifier $F$, when $p = 1$ (top plot); When $p = 3$, we estimate $\mathrm{dist}(f_p, F^{(p)})$ instead (bottom plot). The training is done under different choices of intra-subclass variance $\alpha$ and repeated 10 times per $\alpha$; the Solid line shows the average and the shade denotes the region between max and min values. **(b)** Given a trained network obtained from an instance of this training ($\alpha = 0.1$), we reorder the neurons w.r.t. their contributions $|v_j|\|\boldsymbol{w}_j\|$ and then plot the contributions in a bar plot; **(c)(d)** Given neurons with large contributions, we plot a colormap, with each pixel represents some $\cos(\boldsymbol{w}_j, \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ could be either average class centers $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$ or subclass centers $\boldsymbol{\mu}_k, k \in [K]$. Note: For visibility, the neurons are reordered again so that neurons aligned with the same $\boldsymbol{\mu}$ are grouped together. **(e)** Lastly, we carry out $l_2$ PGD attack on a test dataset and plot the robust accuracy of the trained network under different choices of attack radius.

## B. Additional experiments on MNIST dataset

To further illustrate the effectiveness of pReLU activation in improving the adversarial robustness of the trained network, we conduct the following additional experiments on MNIST dataset:

**MNIST digits classification.** We follow the same experiment setting in Section 5.2.1 and train the network to classify the digits instead of their parity. Figure 6 reports the training results. Despite that our theoretical results only suggest robustness gain can be achieved under datasets with subclasses, these additional experimental results show that the pReLU networks are still much more robust than their ReLU counterpart even for the original digit classification task, which clearly does not follow our data assumption.

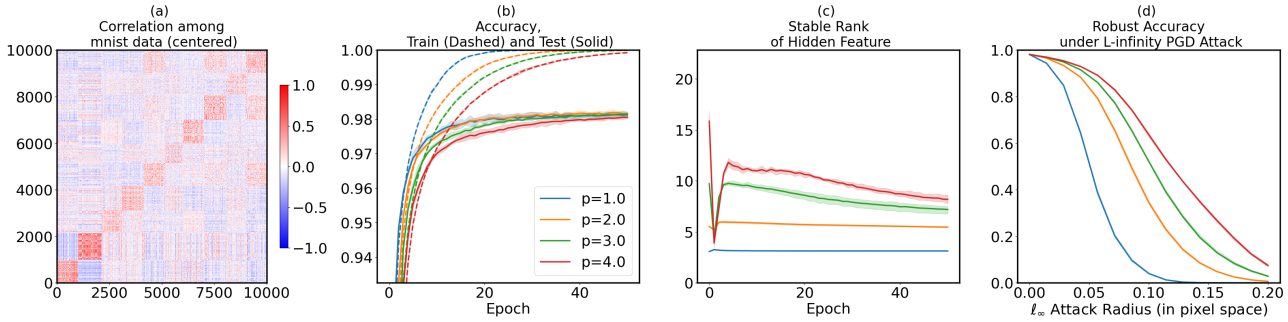


*Figure 6.* Digits classification on MNIST dataset with pReLU networks. **(a)** We plot the data correlation as a colormap, where each pixel represents some $\cos(x_i, x_j)$ between two centered data $x_i, x_j$ from MNIST training dataset; **(b)** We run Adam with batch size 1000 to train a pReLU network under Kaiming initialization (repeated 10 times), then plot the training/testing accuracy during training for different choice of $p$; **(c)** We stack the hidden post-activation representation of each training sample into a matrix and compute its stable rank, and plot the evolution of this stable rank during training; **(d)** After training for 50 epoch, we carry out APGD $l_\infty$-attack on MNIST test dataset (in pixel space) and plot the robust accuracy of the trained pReLU network under different choice of attack radius.

**Evaluating robustness under different attacks.** Lastly, we check the adversarial robustness of the trained networks above under attacks of different norms. The results are summarized in the table below, and they show that the robustness gained from pReLU activation is agnostic to the choice of attacks.

*Table 1.* Robust accuracy of pReLU networks under different attacks. The reported accuracy is the mean value over 5 runs with random initialization. Bold text indicates the best accuracy within the same row.

| | ReLU ($p=1$) | ReLU ($p=2$) | ReLU ($p=3$) | ReLU ($p=4$) |
|---|---|---|---|---|
| Clean Acc. | 0.981($\pm$0.000) | **0.982** ($\pm$0.000) | **0.982** ($\pm$0.000) | 0.980 ($\pm$0.000) |
| Robust Acc. ($\ell_\infty$, radius= 0.05) | 0.512 ($\pm$0.006) | 0.844 ($\pm$0.002) | 0.892 ($\pm$0.002) | **0.913** ($\pm$0.001) |
| Robust Acc. ($\ell_\infty$, radius= 0.1) | 0.040 ($\pm$0.002) | 0.346 ($\pm$0.005) | 0.522 ($\pm$0.004) | **0.637** ($\pm$0.004) |
| Robust Acc. ($\ell_2$, radius= 1) | 0.301 ($\pm$0.004) | 0.640 ($\pm$0.006) | 0.726 ($\pm$0.004) | **0.775** ($\pm$0.002) |
| Robust Acc. ($\ell_2$, radius= 2) | 0.007 ($\pm$0.001) | 0.101 ($\pm$0.002) | 0.165 ($\pm$0.003) | **0.239** ($\pm$0.003) |
| Robust Acc. ($\ell_1$, radius= 5) | 0.500 ($\pm$0.007) | 0.744 ($\pm$0.007) | 0.780 ($\pm$0.004) | **0.807** ($\pm$0.002) |
| Robust Acc. ($\ell_1$, radius= 10) | 0.098 ($\pm$0.004) | 0.294 ($\pm$0.002) | 0.354 ($\pm$0.002) | **0.402** ($\pm$0.003) |

## C. Proofs for Proposition 1 and Theorem 1

### C.0. Verifying the claim regarding realizability of $F(\mathbf{x})$ and $F^{(p)}(\mathbf{x})$

Before we present our poofs for the main results, we start with verifying our claim regarding realizability of $F(\boldsymbol{x})$ and $F^{(p)}(\boldsymbol{x})$:

**Claim** (restated)**.** *The following two statements are true:*

- *When $p = 1$, let $I_+$ and $I_-$ be any two non-empty and disjoint subsets of $[h]$, let $\boldsymbol{\lambda} := (\lambda_j) \in \mathbb{R}^h_{\geq 0}$ be any vector such that $\sum_{j\in I_+} \lambda_j^2 = \sqrt{K_1}$ and $\sum_{j\in I_-} \lambda_j^2 = \sqrt{K_2}$, and let*

$$\begin{aligned} &\boldsymbol{w}_j = \lambda_j \bar{\boldsymbol{\mu}}_+,\ v_j = \lambda_j, && \forall j \in I_+\,, \\ &\boldsymbol{w}_j = \lambda_j \bar{\boldsymbol{\mu}}_-,\ v_j = -\lambda_j, && \forall j \in I_-\,, \\ &\boldsymbol{w}_j = 0,\ v_j = 0, && \textit{otherwise}. \end{aligned}$$

  *Then $f_p(x; \{\boldsymbol{w}_j, v_j\}_{j=1}^h) \equiv F(\boldsymbol{x})$.*

- *When $p \geq 1$, let $I_1, \cdots, I_K$ be any $K$ non-empty and disjoint subsets of $[h]$, let $\boldsymbol{\lambda} := (\lambda_j) \in \mathbb{R}^h_{\geq 0}$ be any vector such that $\sum_{j\in I_k} \lambda_j^2 = 1, \forall k \in [K]$, and let*

$$\begin{aligned} &\boldsymbol{w}_j = \lambda_j \boldsymbol{\mu}_k,\ v_j = \lambda_j, && \forall j \in I_k, k \leq K_1\,, \\ &\boldsymbol{w}_j = \lambda_j \boldsymbol{\mu}_k,\ v_j = -\lambda_j, && \forall j \in I_k, k > K_1\,, \\ &\boldsymbol{w}_j = 0,\ v_j = 0 && \textit{otherwise}. \end{aligned}$$

  *Then $f_p(x; \{\boldsymbol{w}_j, v_j\}_{j=1}^h) \equiv F^{(p)}(\boldsymbol{x})$.*

*Proof.* **Statement 1: When $p = 1$.**. Let $I_+, I_-, \boldsymbol{\lambda}$ be define as in the statement, then we have

$$\begin{aligned} f_p(\boldsymbol{x}; \{\boldsymbol{w}_j, v_j\}_{j=1}^h) &= \sum_{j=1}^h v_j \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle) \\ &= \sum_{j\in I_+} v_j \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle) + \sum_{j\in I_-} v_j \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle) \\ &= \sum_{j\in I_+} \lambda_j \sigma(\langle \boldsymbol{x}, \lambda_j \bar{\boldsymbol{\mu}}_+ \rangle) - \sum_{j\in I_-} \lambda_j \sigma(\langle \boldsymbol{x}, \lambda_j \bar{\boldsymbol{\mu}}_- \rangle) \\ &= \sum_{j\in I_+} \lambda_j^2 \sigma(\langle \boldsymbol{x}, \bar{\boldsymbol{\mu}}_+ \rangle) - \sum_{j\in I_-} \lambda_j^2 \sigma(\langle \boldsymbol{x}, \bar{\boldsymbol{\mu}}_- \rangle) \\ &= \sqrt{K_1} \sigma(\langle \boldsymbol{x}, \bar{\boldsymbol{\mu}}_+ \rangle) - \sqrt{K_2} \sigma(\langle \boldsymbol{x}, \bar{\boldsymbol{\mu}}_- \rangle) = F(\boldsymbol{x})\,. \end{aligned}$$

**Statement 2: When $p \geq 1$.**. Let $\{I_k\}_{k=1}^K, \boldsymbol{\lambda}$ be define as in the statement, then we have

$$\begin{aligned} f_p(\boldsymbol{x}; \{\boldsymbol{w}_j, v_j\}_{j=1}^h) &= \sum_{j=1}^h v_j \frac{\sigma^p(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle)}{\|\boldsymbol{w}_j\|^{p-1}} \\ &= \sum_{k\leq K_1} \sum_{j\in I_k} v_j \frac{\sigma^p(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle)}{\|\boldsymbol{w}_j\|^{p-1}} + \sum_{k> K_1} \sum_{j\in I_k} v_j \frac{\sigma^p(\langle \boldsymbol{x}, \boldsymbol{w}_j \rangle)}{\|\boldsymbol{w}_j\|^{p-1}} \\ &= \sum_{k\leq K_1} \sum_{j\in I_k} \lambda_j \frac{\sigma^p(\langle \boldsymbol{x}, \lambda_j \boldsymbol{\mu}_k \rangle)}{\|\lambda_j \boldsymbol{\mu}_k\|^{p-1}} - \sum_{k> K_1} \sum_{j\in I_k} \lambda_j \frac{\sigma^p(\langle \boldsymbol{x}, \lambda_j \boldsymbol{\mu}_k \rangle)}{\|\lambda_j \boldsymbol{\mu}_k\|^{p-1}} \\ &= \sum_{k\leq K_1} \sum_{j\in I_k} \lambda_j^2 \sigma^p(\langle \boldsymbol{x}, \boldsymbol{\mu}_k \rangle) - \sum_{k> K_1} \sum_{j\in I_k} \lambda_j^2 \sigma^p(\langle \boldsymbol{x}, \boldsymbol{\mu}_k \rangle) \\ &= \sum_{k\leq K_1} \sigma^p(\langle \boldsymbol{x}, \boldsymbol{\mu}_k \rangle) - \sum_{k> K_1} \sigma^p(\langle \boldsymbol{x}, \boldsymbol{\mu}_k \rangle) = F^{(p)}(\boldsymbol{x})\,. \end{aligned}$$

□

### C.1. Auxiliary lemmas

The proof will use some basic facts in probability theory and we list them below.

**Lemma 2.** *Let $\mathcal{E}_1, \mathcal{E}_2$ be two events defined on some probability space, then*

$$\mathbb{P}(\mathcal{E}_1) \leq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_2^c) \tag{9}$$

*Proof.* $\mathbb{P}(\mathcal{E}_1) = \mathbb{P}\left(\mathcal{E}_1 \cap (\mathcal{E}_2 \cup \mathcal{E}_2^c)\right) = \mathbb{P}\left((\mathcal{E}_1 \cap \mathcal{E}_2) \cup (\mathcal{E}_1 \cap \mathcal{E}_2^c)\right) \leq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2^c) \leq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_2^c)$. □

**Lemma 3** (Hoeffding inequality)**.** *For any unit vector $\boldsymbol{\mu} \in \mathbb{S}^{D-1}$, we have*

$$\mathbb{P}_{\boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle \boldsymbol{\mu}, \boldsymbol{\varepsilon} \rangle| > t\right) \leq 2\exp\left(-\frac{CDt^2}{\alpha^2}\right), \tag{10}$$

*for some constant $C > 0$.*

### C.2. Proof for Proposition 1

**Proposition 1** (Test error on clean data, restated)**.** *Given classifiers $F(\boldsymbol{x})$, $F^{(p)}(\boldsymbol{x})$ defined in* (3),(2)*, and a test sample $(x, y) \sim \mathcal{D}_{X,Y}$, we have, for some constant $C > 0$,*

$$\mathbb{P}\left(F(\boldsymbol{x})y > 0\right) \geq 1 - 4\exp\left(-\frac{CD}{4\alpha^2 K}\right)$$
$$\mathbb{P}\left(F^{(p)}(\boldsymbol{x})y > 0\right) \geq 1 - 2(K+1)\exp\left(-\frac{CD}{\alpha^2 K^2}\right), \ \forall p \geq 1\,.$$

*Proof.* The proof is split into parts: We first prove the bound for $F(\boldsymbol{x})$, then show the one for $F^{(p)}(\boldsymbol{x})$. But in both cases we have

$$\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(G(\boldsymbol{x})y > 0\right) = \sum_{k=1}^{K} \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(G(\boldsymbol{x})y > 0 \mid z = k\right)\mathbb{P}\left(z = k\right), \tag{11}$$

where $G$ can be either $F$ or $F^{(p)}$. Thus, it suffices to show

$$\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(G(\boldsymbol{x})y > 0 \mid z = k\right) \geq 1 - 4\exp\left(-\frac{CD}{\alpha^2 K}\right), \forall k = 1, \cdots, K\,. \tag{12}$$

**Bound for $F(\boldsymbol{x})$** we start with the case of $G$ being $F$. When $k \leq K_1$, we have

$$\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(F(\boldsymbol{x})y > 0 \mid z = k\right) = \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F(\boldsymbol{\mu}_k + \boldsymbol{\varepsilon}) > 0\right), \tag{13}$$

and then,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F(\boldsymbol{\mu}_k + \boldsymbol{\varepsilon}) > 0\right) \\
&= 1 - \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F(\boldsymbol{\mu}_k + \boldsymbol{\varepsilon}) < 0\right) \\
&= 1 - \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sqrt{K_1}\sigma\left(\frac{1}{\sqrt{K_1}} + \langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{\mu}}_+ \rangle\right) - \sqrt{K_2}\sigma\left(\langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{\mu}}_- \rangle\right) < 0, \mathcal{E}_1\right) \\
&\geq 1 - \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sqrt{K_1}\left(\frac{1}{\sqrt{K_1}} - \left|\langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{\mu}}_+ \rangle\right|\right) - \sqrt{K_2}\left|\langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{\mu}}_- \rangle\right| < 0\right) \qquad (14) \\
&= 1 - \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(1 - \sqrt{K_1}\left|\langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{\mu}}_+ \rangle\right| - \sqrt{K_2}\left|\langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{\mu}}_- \rangle\right| < 0\right) \\
&\geq 1 - \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\left|\langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{\mu}}_+ \rangle\right| > \frac{1}{2\sqrt{K_1}}\right) - \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0, \frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\left|\langle \boldsymbol{\varepsilon}, \bar{\boldsymbol{\mu}}_- \rangle\right| > \frac{1}{2\sqrt{K_2}}\right) \\
&\geq 1 - 2\exp\left(-\frac{CD}{4\alpha^2 K_1}\right) - 2\exp\left(-\frac{CD}{4\alpha^2 K_2}\right) \leq 1 - 4\exp\left(-\frac{CD}{4\alpha^2 K}\right), \qquad (15)
\end{aligned}$$

where (14) uses the fact that $\sigma(x)$ is non-decreasing w.r.t. $x$., and that $\sigma(x) \leq |x|$. The last line uses Lemma 3. The proof of the case $k > K_1$ is identical to the one above by the symmetry of the problem.

**Bound for $F^{(p)}(\boldsymbol{x})$** When $k \leq K_1$, we have, again,

$$\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(F^{(p)}(\boldsymbol{x})y>0 \mid z=k\right)=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F^{(p)}(\boldsymbol{\mu}_k+\boldsymbol{\varepsilon})>0\right), \tag{16}$$

we define the event $\mathcal{E}_1 := \{|\langle \boldsymbol{\mu}_k, \boldsymbol{\varepsilon}\rangle| < 1\}$. Then by Lemma 2,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F^{(p)}(\boldsymbol{\mu}_k+\boldsymbol{\varepsilon})>0\right)\\
&=1-\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F^{(p)}(\boldsymbol{\mu}_k+\boldsymbol{\varepsilon})<0\right)\\
&\leq 1-\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F^{(p)}(\boldsymbol{\mu}_k+\boldsymbol{\varepsilon})<0,\mathcal{E}_1\right)-\mathbb{P}\left(\mathcal{E}_1^c\right)\\
&\leq 1-\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sigma^p\left(1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle\right)+\sum_{l\leq K_1,l\neq k}\sigma^p\left(\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle\right)-\sum_{K_1<l\leq K_2}\sigma^p\left(\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle\right)<0,\mathcal{E}_1\right)-\mathbb{P}\left(\mathcal{E}_1^c\right)\\
&\leq 1-\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left((1-|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|)^p-\sum_{l\neq k}|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|^p<0,\mathcal{E}_1\right)-\mathbb{P}\left(\mathcal{E}_1^c\right) \qquad (17)\\
&\leq 1-\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(1-|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|-\left(\sum_{l\neq k}|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|^p\right)^{1/p}<0,\mathcal{E}_1\right)-\mathbb{P}\left(\mathcal{E}_1^c\right)\\
&\leq 1-\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(1-|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|-\sum_{l\neq k}|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|<0,\mathcal{E}_1\right)-\mathbb{P}\left(\mathcal{E}_1^c\right) \qquad (18)\\
&\leq 1-\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(1-|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|-\sum_{l\neq k}|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|<0\right)-\mathbb{P}\left(\mathcal{E}_1^c\right)\\
&\leq 1-\sum_{k=1}^{K}\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|>\frac{1}{K}\right)-\mathbb{P}\left(\mathcal{E}_1^c\right)\\
&\leq 1-2K\exp\left(-\frac{CD}{\alpha^2K^2}\right)-2\exp\left(-\frac{CD}{\alpha^2}\right)\leq 1-2(K+1)\exp\left(-\frac{CD}{\alpha^2K^2}\right), \qquad (19)
\end{aligned}$$

where (17) uses the fact that under the event $\mathcal{E}_1$, one has $\sigma^p\left(1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle\right)\geq(1-|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|)^p$, and (18) uses the fact that $\|x\|_p\leq\|x\|_1$ for any vector $x$ and $p\geq 1$. The last line uses Lemma 3. The proof of the case $k>K_1$ is identical to the one above by the symmetry of the problem. □

### C.3. Proof for Theorem 1

**Theorem 1** ($l_2$-Adversarial Robustness, restated)**.** *Given classifiers $F(\boldsymbol{x})$, $F^{(p)}(\boldsymbol{x})$ defined in* (3),(2)*, and a test sample $(x,y)\sim\mathcal{D}_{X,Y}$, the following statement is true for some constant $C>0$:*

- *(Non-robustness of $F(\boldsymbol{x})$ against $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$-attack) We let $\boldsymbol{d}_0 := \frac{\sqrt{K_1}\bar{\boldsymbol{\mu}}_+-\sqrt{K_2}\bar{\boldsymbol{\mu}}_-}{\sqrt{K}}\in\mathbb{S}^{D-1}$, then for some $\rho>0$,*

$$\mathbb{P}\left(F\left(\boldsymbol{x}-\frac{y(1+\rho)}{\sqrt{K}}\boldsymbol{d}_0\right)y>0\right)\leq 2\exp\left(-\frac{CD\rho^2}{K\alpha^2}\right).$$

- *(Robustness of $F^{(p)}(\boldsymbol{x})$ against $\mathcal{O}(1)$-attack) We let $p\geq 2$, then for some $0\leq\delta<\sqrt{2}$,*

$$\mathbb{P}\left(\min_{\|\boldsymbol{d}\|\leq 1}F^{(p)}\left(\boldsymbol{x}+\frac{\sqrt{2}-\delta}{2}\boldsymbol{d}\right)y>0\right)\geq 1-2(K+1)\exp\left(-\frac{CD\delta^2}{2K^2\alpha^2}\right).$$

*Proof.* This proof is split into two parts. One for $F(\boldsymbol{x})$ and another for $F^{(p)}(\boldsymbol{x})$.

**Non-robustness of $F(\boldsymbol{x})$** Since

$$\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(F\left(\boldsymbol{x}-\frac{y(1+\rho)}{\sqrt{K}}\boldsymbol{d}_0\right)y>0\right)=\sum_{k=1}^{K}\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(F\left(\boldsymbol{x}-\frac{y(1+\rho)}{\sqrt{K}}\boldsymbol{d}_0\right)y>0\mid z=k\right)\mathbb{P}\left(z=k\right)\ , \tag{20}$$

It suffices to show

$$\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(F\left(\boldsymbol{x}-\frac{y(1+\rho)}{\sqrt{K}}\boldsymbol{d}_0\right)y>0\mid z=k\right)\leq 2\exp\left(-\frac{CD^2K}{(\rho-1)^2\alpha^2}\right),\forall k\leq K \tag{21}$$

When $k\leq K_1$, we have

$$\begin{aligned}
&\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(F\left(\boldsymbol{x}-\frac{y(1+\rho)}{\sqrt{K}}\boldsymbol{d}_0\right)y>0\mid z=k\right)\\
&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F\left(\boldsymbol{\mu}_k+\boldsymbol{\varepsilon}-\frac{1+\rho}{\sqrt{K}}\boldsymbol{d}_0\right)>0\right)\\
&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sqrt{K_1}\sigma\left(\frac{1}{\sqrt{K_1}}+\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\right\rangle-\frac{(1+\rho)\sqrt{K_1}}{K}\right)-\sqrt{K_2}\sigma\left(\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\right\rangle+\frac{(1+\rho)\sqrt{K_2}}{K}\right)>0\right).
\end{aligned} \tag{22}$$

To proceed, we define the event $\mathcal{E}_2:=\{\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\right\rangle-\frac{(1+\rho)\sqrt{K_1}}{\sqrt{K}}+\frac{1}{\sqrt{K_1}}\geq 0\}$, then by Lemma 2

$$\begin{aligned}
&(22)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sqrt{K_1}\sigma\left(\frac{1}{\sqrt{K_1}}+\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\right\rangle-\frac{(1+\rho)\sqrt{K_1}}{K}\right)-\sqrt{K_2}\sigma\left(\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\right\rangle+\frac{(1+\rho)\sqrt{K_2}}{K}\right)>0,\mathcal{E}_2\right)\\
&\qquad+\mathbb{P}\left(\sqrt{K_1}\sigma\left(\frac{1}{\sqrt{K_1}}+\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\right\rangle-\frac{(1+\rho)\sqrt{K_1}}{K}\right)-\sqrt{K_2}\sigma\left(\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\right\rangle+\frac{(1+\rho)\sqrt{K_2}}{K}\right)>0,\mathcal{E}_2^c\right)\\
&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sqrt{K_1}\sigma\left(\frac{1}{\sqrt{K_1}}+\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\right\rangle-\frac{(1+\rho)\sqrt{K_1}}{K}\right)-\sqrt{K_2}\sigma\left(\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\right\rangle+\frac{(1+\rho)\sqrt{K_2}}{K}\right)>0,\mathcal{E}_2\right) && (23)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(1+\sqrt{K_1}\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\right\rangle-\frac{(1+\rho)K_1}{K}-\sqrt{K_2}\left\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\right\rangle-\frac{(1+\rho)K_2}{K}>0,\mathcal{E}_2\right) && (24)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(1+\sqrt{K}\left|\langle\boldsymbol{\varepsilon},\boldsymbol{d}_0\rangle\right|-(1+\rho)>0\right)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\left|\langle\boldsymbol{\varepsilon},\boldsymbol{d}_0\rangle\right|>\frac{\rho}{\sqrt{K}}\right)\leq 2\exp\left(-\frac{CD\rho^2}{K\alpha^2}\right), && (25)
\end{aligned}$$

where (23) is because the second probability is 0, and (24) uses again the monotonicity of ReLU $\sigma(x)$. The last line uses Lemma 3. The proof of the case $k>K_1$ is identical to the one above by the symmetry of the problem.

**Robustness of $F^{(p)}(\boldsymbol{x})$** It suffices to show

$$\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(\min_{\|\boldsymbol{d}\|\leq 1}F^{(p)}\left(\boldsymbol{x}+\frac{\sqrt{2}-\delta}{2}d\right)y<0\mid z=k\right)\leq 2(K_2+2)\exp\left(-\frac{CD\delta^2}{2(K_2+1)^2\alpha^2}\right). \tag{26}$$

When $k \leq K_1$, we have

$$\begin{aligned}
&\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(\min_{\|\boldsymbol{d}\|\leq 1} F^{(p)}\left(\boldsymbol{x}+\frac{\sqrt{2}-\delta}{2}d\right) y<0 \mid z=k\right)\\
&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1} F^{(p)}\left(\boldsymbol{\mu}_k+\boldsymbol{\varepsilon}+\frac{\sqrt{2}-\delta}{2}d\right)<0\right)\\
&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1}\left[\sigma^p\left(1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_k\rangle\right)\right.\right.\\
&\qquad+\sum_{l\neq k,1\leq l\leq K_1}\sigma^p\left(\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle\right)\\
&\qquad\left.\left.-\sum_{K_1+1\leq l\leq K_2}\sigma^p\left(\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle\right)\right]<0\right) \qquad (27)
\end{aligned}$$

To proceed, we define the event

$$\mathcal{E}_3 := \left\{1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_k\rangle\geq 0, \forall d\in\mathbb{S}^{D-1}\right\}, \qquad (28)$$

and for ease of presentation, we write

$$\begin{aligned}
G(\boldsymbol{\varepsilon},d) \;:=\; &\sigma^p\left(1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_k\rangle\right)+\sum_{l\neq k,1\leq l\leq K_1}\sigma^p\left(\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle\right)\\
&-\sum_{l=K_1+1}^{K_2}\sigma^p\left(\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle\right).
\end{aligned}$$

Then, by Lemma 2,

$$\begin{aligned}
(27) \;&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1} G(\boldsymbol{\varepsilon},d)<0\right)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1} G(\boldsymbol{\varepsilon},d)<0,\mathcal{E}_3\right)+\mathbb{P}\left(\mathcal{E}_3^c\right). \qquad (29)
\end{aligned}$$

Now under the event $\mathcal{E}_3$, we can lower bound $G(\boldsymbol{\varepsilon},d)$ by

$$\begin{aligned}
&G(\boldsymbol{\varepsilon},d)\\
&=\sigma^p\left(1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_k\rangle\right)+\sum_{l\neq k,1\leq l\leq K_1}\sigma^p\left(\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle\right)\\
&\qquad-\sum_{l=K_1+1}^{K_2}\sigma^p\left(\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle\right). \qquad (30)\\
&\geq\left(1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_k\rangle\right)^p-\sum_{l=K_1+1}^{K_2}\left(|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|+\frac{\sqrt{2}-\delta}{2}|\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle|\right)^p. \qquad (31)
\end{aligned}$$

Thus, we have

(29)

$$
\begin{aligned}
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1}\left(1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_k\rangle\right)^p-\sum_{l=K_1+1}^{K_2}\left(|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|+\frac{\sqrt{2}-\delta}{2}|\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle|\right)^p<0,\mathcal{E}_3\right)\\
&\qquad+\mathbb{P}\left(\mathcal{E}_3^c\right)\\
&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1}1+\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_k\rangle-\left(\sum_{l=K_1+1}^{K_2}\left(|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|+\frac{\sqrt{2}-\delta}{2}|\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle|\right)^p\right)^{\frac{1}{p}}<0,\mathcal{E}_3\right)\\
&\qquad+\mathbb{P}\left(\mathcal{E}_3^c\right)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\Bigg(\underbrace{\min_{\|\boldsymbol{d}\|\leq 1}1+\frac{\sqrt{2}-\delta}{2}\langle\boldsymbol{d},\boldsymbol{\mu}_k\rangle-\frac{\sqrt{2}-\delta}{2}\left(\sum_{l=K_1+1}^{K_2}(|\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle|)^p\right)^{\frac{1}{p}}}_{M^*(\delta)}\\
&\qquad-\left(\sum_{l=K_1+1}^{K_2}(|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|)^p\right)^{\frac{1}{p}}-|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|<0,\mathcal{E}_3\Bigg)+\mathbb{P}\left(\mathcal{E}_3^c\right) \qquad (32)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|+\left(\sum_{l=K_1+1}^{K_2}(|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|)^p\right)^{\frac{1}{p}}>M^*(\delta)\right)+\mathbb{P}\left(\mathcal{E}_3^c\right)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|+\sum_{l=K_1+1}^{K_2}|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|>M^*(\delta)\right)+\mathbb{P}\left(\mathcal{E}_3^c\right) \qquad (33)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|>\frac{M^*(\delta)}{K_2+1}\right)+\sum_{l=K_1+1}^{K_2}\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|>\frac{M^*(\delta)}{K_2+1}\right)+\mathbb{P}\left(\mathcal{E}_3^c\right)\\
&\leq\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|>\frac{M^*(\delta)}{K_2+1}\right)+\sum_{l=K_1+1}^{K_2}\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}\rangle|>\frac{M^*(\delta)}{K_2+1}\right)\\
&\qquad+\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\mu}_k,\boldsymbol{\varepsilon}\rangle|>1-\frac{\sqrt{2}}{2}\right)\\
&\leq 2(K_2+1)\exp\left(-\frac{CD(M^*(\delta))^2}{(K_2+1)^2\alpha^2}\right)+2\exp\left(-\frac{CD}{4\alpha^2}\right)\leq 2(K_2+2)\exp\left(-\frac{CD(M^*(\delta))^2}{(K_2+1)^2\alpha^2}\right), \qquad (34)
\end{aligned}
$$

where (32) uses the sub-additivity of $\ell_p$ norm, and (33) uses again the inequality $\|\boldsymbol{x}\|_p\leq\|\boldsymbol{x}\|_1$ for any $\boldsymbol{x}$ and $p\geq 1$. The last line uses Lemma 3. The remaining thing is to show that $M^*(\delta)=\frac{\delta}{\sqrt{2}}$. First, by the property of $l_p$ norm (when $p\geq 2$), we have

$$
\left(\sum_{l=K_1+1}^{K_2}(|\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle|)^p\right)^{\frac{1}{p}}\leq\sqrt{\sum_{l=K_1+1}^{K_2}|\langle\boldsymbol{d},\boldsymbol{\mu}_l\rangle|^2}, \qquad (35)
$$

then

$$
\begin{aligned}
M^*(\delta) = & \min_{\|\boldsymbol{d}\|\le 1} 1 - \frac{\sqrt{2}-\delta}{2}\langle \boldsymbol{d}, \boldsymbol{\mu}_k\rangle - \frac{\sqrt{2}-\delta}{2}\left(\sum_{l=K_1+1}^{K_2} (|\langle \boldsymbol{d}, \boldsymbol{\mu}_l\rangle|)^p\right)^{\frac{1}{p}} \\
\ge & \min_{\|\boldsymbol{d}\|\le 1} 1 - \frac{\sqrt{2}-\delta}{2}|\langle \boldsymbol{d}, \boldsymbol{\mu}_k\rangle| - \frac{\sqrt{2}-\delta}{2}\sqrt{\sum_{l=K_1+1}^{K_2} |\langle \boldsymbol{d}, \boldsymbol{\mu}_l\rangle|^2} \\
\ge & \min_{\|\boldsymbol{d}\|\le 1} 1 - \frac{\sqrt{2}-\delta}{2}\sqrt{2}\sqrt{|\langle \boldsymbol{d}, \boldsymbol{\mu}_k\rangle|^2 + \sum_{l=K_1+1}^{K_2} |\langle \boldsymbol{d}, \boldsymbol{\mu}_l\rangle|^2} \qquad (36) \\
\ge & \min_{\|\boldsymbol{d}\|\le 1} 1 - \frac{\sqrt{2}-\delta}{2}\sqrt{2} = \frac{\delta}{\sqrt{2}}, \qquad (37)
\end{aligned}
$$

where (36) uses the fact that $\sqrt{a}+\sqrt{b} \le \sqrt{2(a+b)}$ for any $a, b \ge 0$, and (37) uses the fact that $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K$ are an orthonormal basis, thus $\sqrt{\sum_{l=1}^{K} |\langle \boldsymbol{d}, \boldsymbol{\mu}_l\rangle|^2} \le \|\boldsymbol{d}\| \le 1$.

We have proved $M^*(\delta) \ge \frac{\delta}{\sqrt{2}}$, and this lower bound can be attained by $\boldsymbol{d}^* = \frac{-\boldsymbol{\mu}_k + \boldsymbol{\mu}_l}{\sqrt{2}}$ for any $K_1+1 \le l \le K_2$. Therefore $M^*(\delta) = \frac{\delta}{\sqrt{2}}$. Finally, we have

$$
\begin{aligned}
\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(\min_{\|\boldsymbol{d}\|\le 1} F^{(p)}\left(\boldsymbol{x} + \frac{\sqrt{2}-\delta}{2}\boldsymbol{d}\right) y < 0 \mid z = k\right) \le & \; 2(K_2+2)\exp\left(-\frac{CD(M^*(\delta))^2}{(K_2+1)^2\alpha^2}\right) \\
= & \; 2(K_2+2)\exp\left(-\frac{CD\delta^2}{2(K_2+1)^2\alpha^2}\right) \\
\le & \; 2(K+1)\exp\left(-\frac{CD\delta^2}{2K^2\alpha^2}\right). \qquad (38)
\end{aligned}
$$

The proof of the case $k > K_1$ is identical to the one above by the symmetry of the problem. □

### C.4. Complementary results to Theorem 1

**Theorem 3** ($l_2$-Adversarial Robustness, complementary to Theorem 1)**.** *Given classifiers $F(\boldsymbol{x})$, $F^{(p)}(\boldsymbol{x})$ defined in* (3),(2)*, and a test sample $(x, y) \sim \mathcal{D}_{X,Y}$, the following statement is true for some constant $C > 0$:*

- *For some $0 \le \rho \le 1$,*

$$
\mathbb{P}\left(\min_{\|\boldsymbol{d}\|\le 1} F\left(\boldsymbol{x} + \frac{(1-\rho)}{\sqrt{K}}\boldsymbol{d}\right) y > 0\right) \ge 1 - 2\exp\left(-\frac{CD\rho^2}{K\alpha^2}\right).
$$

- *For some $0 < \delta$,*

$$
\mathbb{P}\left(\min_{\|\boldsymbol{d}\|\le 1} F^{(p)}\left(\boldsymbol{x} + \frac{\sqrt{2}+\delta}{2}\boldsymbol{d}\right) y > 0\right) \le 4\exp\left(\frac{CD\delta^2}{8K^2\alpha^2}\right).
$$

The proof has the same spirit as those for Theorem 1 so we state it briefly.

*Proof.* **Robustness of $F(\boldsymbol{x})$, complementary result** It suffices to show that

$$
\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(\min_{\|\boldsymbol{d}\|\le 1} F\left(\boldsymbol{x} + \frac{(1-\rho)}{\sqrt{K}}\boldsymbol{x}\right) y < 0 \mid z = k\right) \le 4\exp\left(-\frac{CD\rho^2}{K\alpha^2}\right), \forall k \le K. \qquad (39)
$$

When $k \leq K_1$, we have

$$
\begin{aligned}
&\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(\min_{\|\boldsymbol{d}\|\leq 1} F\left(\boldsymbol{x}+\frac{(1-\rho)}{\sqrt{K}}\boldsymbol{x}\right)y<0 \mid z=k\right)\\
&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1} F\left(\boldsymbol{\mu}_k+\boldsymbol{\varepsilon}+\frac{(1-\rho)}{\sqrt{K}}\boldsymbol{x}\right)<0\right)\\
&=\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1}\left[\sqrt{K_1}\sigma\left(\frac{1}{\sqrt{K_1}}+\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\rangle+\frac{1-\rho}{\sqrt{K}}\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_+\rangle\right)-\sqrt{K_2}\sigma\left(\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\rangle+\frac{1-\rho}{\sqrt{K}}\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_-\rangle\right)\right]<0\right) \quad (40)
\end{aligned}
$$

If we still let $d_0 := \frac{\sqrt{K_1}\bar{\boldsymbol{\mu}}_+-\sqrt{K_2}\bar{\boldsymbol{\mu}}_-}{\sqrt{K}} \in \mathbb{S}^{D-1}$, then by the fact that $|x| \geq \sigma(x) \geq x$ for any $x$, we have

(40)

$$
\begin{aligned}
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1}\left[1+\sqrt{K_1}\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\rangle+\sqrt{K_1}\frac{1-\rho}{\sqrt{K}}\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_+\rangle-\sqrt{K_2}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\rangle\right|-\sqrt{K_2}\frac{1-\rho}{\sqrt{K}}\left|\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_-\rangle\right|\right]<0\right)\\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1}\left[1-\sqrt{K_1}\frac{1-\rho}{\sqrt{K}}\left|\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_+\rangle\right|-\sqrt{K_2}\frac{1-\rho}{\sqrt{K}}\left|\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_-\rangle\right|\right]-\sqrt{K_2}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\rangle\right|-\sqrt{K_1}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\rangle\right|<0\right)\\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1}\left[1-\frac{1-\rho}{\sqrt{K}}\sqrt{K_1+K_2}\sqrt{\left|\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_+\rangle\right|^2+\left|\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_-\rangle\right|^2}\right]-\sqrt{K_2}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\rangle\right|-\sqrt{K_1}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\rangle\right|<0\right) \quad (41)\\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\min_{\|\boldsymbol{d}\|\leq 1}[1-(1-\rho)]-\sqrt{K_2}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\rangle\right|-\sqrt{K_1}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\rangle\right|<0\right) \quad (42)\\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sqrt{K_2}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\rangle\right|+\sqrt{K_1}\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\rangle\right|>\rho\right)\\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\rangle\right|+\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_+\rangle\right|>\frac{\rho}{\sqrt{K}}\right)\\
&\leq 2\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\left|\langle\boldsymbol{\varepsilon},\bar{\boldsymbol{\mu}}_-\rangle\right|>\frac{\rho}{2\sqrt{K}}\right)\leq 4\exp\left(-\frac{CD\rho^2}{4K\alpha^2}\right), \quad (43)
\end{aligned}
$$

where (41) uses the fact that $ab+cd \leq \sqrt{a^2+c^2}\sqrt{b^2+d^2}$ for any $a,b,c,d \in \mathbb{R}$, a simple application of Cauchy-Schwarz inequality, and (42) uses the fact that $\bar{\boldsymbol{\mu}}_+, \bar{\boldsymbol{\mu}}_-$ are orthonormal, thus $\sqrt{\left|\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_+\rangle\right|^2+\left|\langle\boldsymbol{d},\bar{\boldsymbol{\mu}}_-\rangle\right|^2} \leq \|d\| \leq 1$. The last line uses Lemma 3. The proof of the case $k > K_1$ is identical to the one above by the symmetry of the problem.

**Non-robustness of $F^{(p)}(\boldsymbol{x})$, complementary result** It suffices to show that

$$
\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(\min_{\|\boldsymbol{d}\|\leq 1} F^{(p)}\left(\boldsymbol{x}+\frac{\sqrt{2}+\delta}{2}\boldsymbol{d}\right)y>0 \mid z=k\right)\leq 4\exp\left(\frac{CD\delta^2}{8K^2\alpha^2}\right), \forall k\leq K\,. \quad (44)
$$

When $k \leq K_1$, we define $\boldsymbol{d}_k = \frac{-\boldsymbol{\mu}_k + \boldsymbol{\mu}_K}{\sqrt{2}}$, then

$$
\begin{aligned}
&\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(\min_{\|\boldsymbol{d}\|\leq 1} F^{(p)}\left(\boldsymbol{x}+\frac{\sqrt{2}+\delta}{2}\boldsymbol{d}\right) y > 0 \mid z = k\right) \\
&\leq \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}_{X,Y}}\left(F^{(p)}\left(\boldsymbol{x}+\frac{\sqrt{2}+\delta}{2}\boldsymbol{d}_k\right) y > 0 \mid z = k\right) \\
&= \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(F^{(p)}\left(\boldsymbol{\mu}_k+\boldsymbol{\varepsilon}+\frac{\sqrt{2}+\delta}{2}\boldsymbol{d}_k\right) > 0\right) \\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sigma^p\left(\frac{1-\delta/\sqrt{2}}{2}+\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle\right)-\sigma^p\left(\frac{1+\delta/\sqrt{2}}{2}+\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle\right)+\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|^p > 0\right) \\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sigma^p\left(\frac{1-\delta/\sqrt{2}}{2}+\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle\right)+\left(\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|\right)^p > \sigma^p\left(\frac{1+\delta/\sqrt{2}}{2}+\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle\right)\right) \\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sigma^p\left(\frac{1-\delta/\sqrt{2}}{2}+|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle|\right)+\left(\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|\right)^p > \sigma^p\left(\frac{1+\delta/\sqrt{2}}{2}-|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle|\right)\right) \quad (45) \\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\left(\sigma\left(\frac{1-\delta/\sqrt{2}}{2}+|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle|\right)+\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|\right)^p > \sigma^p\left(\frac{1+\delta/\sqrt{2}}{2}-|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle|\right)\right) \quad (46) \\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sigma\left(\frac{1-\delta/\sqrt{2}}{2}+|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle|\right)+\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle| > \sigma\left(\frac{1+\delta/\sqrt{2}}{2}-|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle|\right)\right) \\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sigma\left(\frac{1-\delta/\sqrt{2}}{2}+|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle|\right)+\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle| > \frac{1+\delta/\sqrt{2}}{2}-|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle|\right), \quad (47)
\end{aligned}
$$

where (45) uses the fact that $\sigma^p(x)$ is non-decreasing w.r.t. $x$, and (46) uses the fact that $(a+b)^p \geq a^p + b^p$ for any $a, b > 0$. We define the event

$$
\mathcal{E}_5 := \left\{\frac{1-\delta/\sqrt{2}}{2}+\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle > 0\right\}. \quad (48)
$$

Then by Lemma 2,

$$
\begin{aligned}
(47) &\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\frac{1-\delta/\sqrt{2}}{2}+|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle|+\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|>\frac{1+\delta/\sqrt{2}}{2}-|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle|,\mathcal{E}_5\right)\\
&\quad+\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle|+\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|>\frac{1+\delta/\sqrt{2}}{2}-|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle|,\mathcal{E}_5^c\right)\\
&\leq \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle|+|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle|+\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|>\frac{\delta}{\sqrt{2}}\right)\\
&\quad+\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_K\rangle|+|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_k\rangle|+\sum_{l\neq k,l\neq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|>\frac{1+\delta/\sqrt{2}}{2}\right)\\
&\leq 2\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(\sum_{1\leq l\leq K}|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_l\rangle|>\frac{\delta}{2\sqrt{2}}\right)\\
&\leq 2\mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\frac{\alpha^2}{D}\boldsymbol{I}_D\right)}\left(|\langle\boldsymbol{\varepsilon},\boldsymbol{\mu}_1\rangle|>\frac{\delta}{2\sqrt{2}K}\right)\leq 4\exp\left(\frac{CD\delta^2}{8K^2\alpha^2}\right).
\end{aligned}
\tag{49}
$$

The last line uses Lemma 3. When $k > K_1$, the proof is similar with $d_k := \frac{-\boldsymbol{\mu}_k+\boldsymbol{\mu}_1}{\sqrt{2}}$. □

## D. Proofs for Lemma 1 and Theorem 2

### D.1. Alignment bias illustrated

As we showed in Lemma 1, during the alignment phase $t \leq T$, one have the following approximation

$$\frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} \simeq \operatorname{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t))\,, \tag{50}$$

with

$$x^{(p)}(w) = \sum_{k=1}^{K} \gamma_k(w) y_k \boldsymbol{x}_k \cdot p[\cos(\boldsymbol{x}_k, w)]^{p-1}\,, \tag{51}$$

which essentially shows that when $\boldsymbol{w}_j$ is a *positive neuron* ($\operatorname{sign}(v_j(0)) > 0$), then gradient flow dynamics during alignment phase pushes $\boldsymbol{w}_j/\|\boldsymbol{w}_j\|$ toward the direction of $x^{(p)}(w)$.

Notably, $x^{(p)}(\boldsymbol{w}_j)$ critically depends on $p$. Roughly speaking, when $p = 1$, $x^{(p)}(\boldsymbol{w}_j)$ are more aligned with $\bar{\boldsymbol{\mu}}_+$ and $\bar{\boldsymbol{\mu}}_-$, while when $p > 3$, $x^{(p)}(\boldsymbol{w}_j)$ are more aligned with one of the subclass centers, thus by moving toward $x^{(p)}(\boldsymbol{w}_j)$ in direction, the neurons are likely to align with average class centers in the former case, and with subclass centers in the latter case. We elaborate this statement here with a toy example.

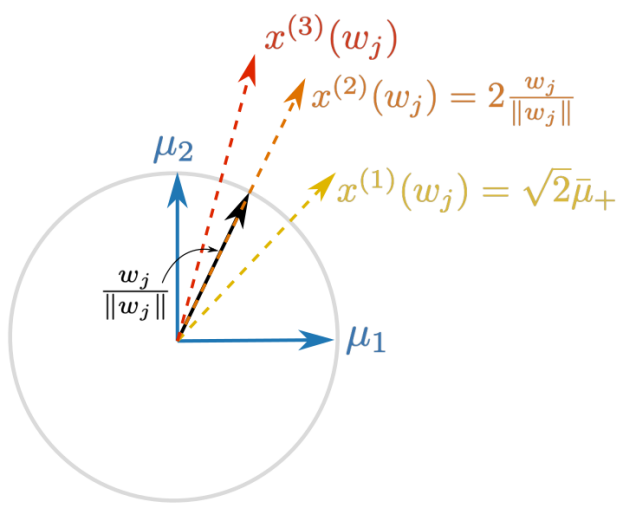


*Figure 7.* Alignment bias visualized. During alignment phase $\boldsymbol{w}_j$ is moving toward $x^{(p)}(\boldsymbol{w}_j)$ in direction. When $p = 1$, $x^{(1)}(\boldsymbol{w}_j)$ is aligned with average class center $\bar{\boldsymbol{\mu}}_+$; When $p = 3$, $x^{(p)}(\boldsymbol{w}_j)$ is more aligned with one of the subclass centers $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, depending on which one is closer to $\boldsymbol{w}_j$ in cosine distance.

Suppose the dataset ($K = 2, K_1 = 2$) only contains two orthogonal $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ in the 2-d plane and they both have positive labels. Given a positive neuron $\boldsymbol{w}_j$ that is activated by both $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$, as shown in Figure 7. During alignment phase $\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}$ is moving towards the direction of $x^{(p)}(\boldsymbol{w}_j)$, which is

- when $p = 1$,

$$x^{(1)}(\boldsymbol{w}_j) = \sum_{k=1}^{K} \gamma_k(\boldsymbol{w}_j) y_k \boldsymbol{x}_k = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 = \sqrt{2}\bar{\boldsymbol{\mu}}_+\,, \tag{52}$$

  exactly aligned with average class center $\bar{\boldsymbol{\mu}}_+$.

- when $p = 2$,

$$x^{(2)}(\boldsymbol{w}_j) = \sum_{k=1}^{K} \gamma_k(\boldsymbol{w}_j) y_k \boldsymbol{x}_k \cdot 2[\cos(\boldsymbol{x}_k, w)] = 2\left(\boldsymbol{\mu}_1 \cos\left(\boldsymbol{\mu}_1, \boldsymbol{w}_j\right) + \boldsymbol{\mu}_2 \cos\left(\boldsymbol{\mu}_2, \boldsymbol{w}_j\right)\right) = 2\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\,, \tag{53}$$

  exactly aligned with $\boldsymbol{w}_j$ itself.

- when $p = 3$,

$$x^{(3)}(w) = \sum_{k=1}^{K} \gamma_k(\boldsymbol{w}_j) y_k \boldsymbol{x}_k \cdot 3[\cos(\boldsymbol{x}_k, \boldsymbol{w}_j)]^2 = 3\left(\boldsymbol{\mu}_1 \cos\left(\boldsymbol{\mu}_1, \boldsymbol{w}_j\right)^2 + \boldsymbol{\mu}_2 \cos\left(\boldsymbol{\mu}_2, \boldsymbol{w}_j\right)^2\right), \tag{54}$$

  getting closer to either $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$, depending which one is closer to $\boldsymbol{w}_j$ in cosine distance.

Although this example is even more simplified than the one in Section 4, it is easy to visualize and keeps the core relationship between the alignment bias of the neurons and the pReLU activation. From this, we see how the alignment bias is altered under different choices of $p$.

### D.2. Auxiliary Lemma

We first prove the following, most analyses on gradient flow with small initialization (Boursier et al., 2022; Boursier & Flammarion, 2024; Min et al., 2024) have similar results, saying that the norm of the neurons stays close to zero during the alignment phase.

**Lemma 4.** *Given some initialization in* (4)*, then for any* $\epsilon \leq \frac{1}{4\sqrt{h}M^2}$*, any solution to the gradient flow dynamics under the simplified training dataset satisfies*

$$\max_j \|\boldsymbol{w}_j(t)\|^2 \leq \frac{2\epsilon M^2}{\sqrt{h}}, \quad \max |f_p(\boldsymbol{x}_k; \boldsymbol{\theta}(t))| \leq 2\epsilon\sqrt{h}M^2\,, \tag{55}$$

$\forall t \leq \frac{1}{4K} \log \frac{1}{\sqrt{h}\epsilon}$.

and we need the following lemma

**Lemma 5.** *Given nonnegative* $z_1, \cdots, z_n$*, consider a function*

$$g_p(q; \{z_i\}_{i=1}^n) = \left(\sum_{i=1}^n z_i^q\right)\left(\sum_{i=1}^n z_i^{p+1-q}\right), \tag{56}$$

*then* $g_p(q; \{z_i\}_{i=1}^n)$ *is convex on* $\mathbb{R}$*. Moreover, as long as* $z_i \neq z_j$ *for some* $i, j$*, then* $g_p(q; \{z_i\}_{i=1}^n)$ *is strictly convex with minimum at* $q^* = \frac{p+1}{2}$.

we leave their proofs at the end of this section. Lastly, we use the following lemma from Min et al. (2024)

**Lemma 6.** *For* $\ell$ *being either exponential or logistic loss, we have*

$$|-\nabla_{\hat{y}}\ell(y, \hat{y}) - y| \leq 2|\hat{y}|, \forall y \in \{+1, -1\}, \quad \forall |\hat{y}| \leq 1\,. \tag{57}$$

### D.3. Proof for Lemma 1

**Lemma 1** (restated)**.** *Given some initialization from* (4)*, if* $\epsilon = \mathcal{O}(\frac{1}{\sqrt{h}})$*, then there exists* $T = \boldsymbol{\theta}(\frac{1}{K}\log\frac{1}{\sqrt{h}\epsilon})$ *such that the trajectory under gradient flow training with the simplified training dataset almost surely satisfies that* $\forall t \leq T$*,*

$$\max_j \left\| \frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \operatorname{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t)) \right\| = \mathcal{O}\left(\epsilon k\sqrt{h}\right),$$

*where* $\mathcal{P}^{\perp}_w = I - \frac{ww^\top}{\|w\|^2}$ *and*

$$x^{(p)}(w) = \sum_{k=1}^{K} \gamma_k(w) y_k \boldsymbol{x}_k \cdot p[\cos(\boldsymbol{x}_k, w)]^{p-1}, \tag{58}$$

*with* $\gamma_k(w)$ *being a subgradient of* $\sigma^p(z)$ *at* $z = \langle \boldsymbol{x}_k, w\rangle$.

*Proof.* For simplicity, we write $\boldsymbol{w}_j(t)$ as $\boldsymbol{w}_j$.

As we will show in the proof for Lemma 4, under balanced initialization,

$$\frac{d}{dt}\boldsymbol{w}_j = -\sum_{k=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f_p(\boldsymbol{x}_k;\boldsymbol{\theta}))\|\boldsymbol{w}_j\|\left(\frac{p[\sigma(\langle\boldsymbol{x}_k,\boldsymbol{w}_j\rangle)]^{p-1}}{\|\boldsymbol{w}_j\|^{p-1}}\boldsymbol{x}_k - (p-1)\frac{[\sigma(\langle\boldsymbol{x}_k,\boldsymbol{w}_j\rangle)]^{p}}{\|\boldsymbol{w}_j\|^{p+1}}\boldsymbol{w}_j\right). \tag{59}$$

Then for any $j \in [h]$,

$$\begin{aligned}\frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} &= \mathcal{P}^{\perp}_{\boldsymbol{w}_j}\cdot\frac{1}{\|\boldsymbol{w}_j\|}\cdot\frac{d}{dt}\boldsymbol{w}_j \\ &= -\text{sign}(v_j(0))\sum_{k=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f_p(\boldsymbol{x}_k;\boldsymbol{\theta}))\mathcal{P}^{\perp}_{\boldsymbol{w}_j}\left(p\cos(\boldsymbol{x}_k,\boldsymbol{w}_j)^{p-1}\boldsymbol{x}_k - (p-1)\frac{[\sigma(\langle\boldsymbol{x}_k,\boldsymbol{w}_j\rangle)]^{p}}{\|\boldsymbol{w}_j\|^{p+1}}\boldsymbol{w}_j\right) \\ &= -\text{sign}(v_j(0))\sum_{i=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f_p(\boldsymbol{x}_k;\boldsymbol{\theta}))\mathcal{P}^{\perp}_{\boldsymbol{w}_j}p\cos(\boldsymbol{x}_k,\boldsymbol{w}_j)^{p-1}\boldsymbol{x}_k\,,\end{aligned}$$

Then

$$\begin{aligned}&\max_j\left\|\frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \text{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)}x^{(p)}(\boldsymbol{w}_j(t))\right\| \\ =&\max_j\left\|\sum_{i=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}(-\ell(y_k, f_p(\boldsymbol{x}_k;\boldsymbol{\theta})) - y_k)\mathcal{P}^{\perp}_{\boldsymbol{w}_j}p\cos(\boldsymbol{x}_k,\boldsymbol{w}_j)^{p-1}\boldsymbol{x}_k\right\| \\ \le&\max_j\left\|\sum_{i=1}^{K}\gamma_k(\boldsymbol{w}_j)\left|\nabla_{\hat{y}}(-\ell(y_k, f_p(\boldsymbol{x}_k;\boldsymbol{\theta})) - y_k\right|\mathcal{P}^{\perp}_{\boldsymbol{w}_j}p\cos(\boldsymbol{x}_k,\boldsymbol{w}_j)^{p-1}\boldsymbol{x}_k\right\| \le 2Kp\max_k|f_p(\boldsymbol{x}_k;\boldsymbol{\theta}(t))|\,,\end{aligned}$$

by Lemma 6. Finally, by Lemma 4, we have for any $\epsilon \le \frac{1}{4\sqrt{h}M^2}$, $\forall t \le T = \frac{1}{4K}\log\frac{1}{\sqrt{h}\epsilon}$, we have

$$\max_j\left\|\frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \text{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)}x^{(p)}(\boldsymbol{w}_j(t))\right\| \le 2Kp\max|f_p(\boldsymbol{x}_k;\boldsymbol{\theta}(t))| \le 4\epsilon\sqrt{h}M^2Kp\,, \tag{60}$$

which finishes the proof. □

### D.4. Proof for Theorem 2

**Theorem 2** (Alignment bias of neurons, complete statement)**.** *Given some $0 < \underline{\delta} < \delta < 1$ and a fixed choice of $p \ge 1$, then $\exists\epsilon_0 := \epsilon_0(\delta, p) > 0$ such that for any solution of the gradient flow on $f_p(x;\boldsymbol{\theta})$ with the simplified training dataset, starting from some initialization from* (4) *with initialization scale $\epsilon < \epsilon_0$, almost surely we have that at any time $t \le T = \boldsymbol{\theta}(\frac{1}{n}\log\frac{1}{\sqrt{h}\epsilon})$ and*

- *$\forall j$ with $\text{sign}(v_j(0)) > 0$,*

$$\frac{d}{dt}\cos\big(\boldsymbol{w}_j(t), \bar{\boldsymbol{\mu}}_+\big)\Big|_{\cos(\boldsymbol{w}_j(t),\bar{\boldsymbol{\mu}}_+)=1-\delta}\begin{cases}>0, & \textit{when } p=1\\ <0, & \textit{when } p\ge 3\end{cases}, \tag{61}$$

*and*

$$\frac{d}{dt}\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_k)\Big|_{\cos(\boldsymbol{w}_j(t),\boldsymbol{\mu}_k)=1-\delta} > 0, \forall k \le K_1\,. \tag{62}$$

- *$\forall j$ with $\text{sign}(v_j(0)) < 0$,*

$$\frac{d}{dt}\cos\big(\boldsymbol{w}_j(t), \bar{\boldsymbol{\mu}}_-\big)\Big|_{\cos(\boldsymbol{w}_j(t),\bar{\boldsymbol{\mu}}_-)=1-\delta}\begin{cases}>0, & \textit{when } p=1\\ <0, & \textit{when } p\ge 3\end{cases}, \tag{63}$$

*and*

$$\frac{d}{dt}\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_k)\Big|_{\cos(\boldsymbol{w}_j(t),\boldsymbol{\mu}_k)=1-\delta} > 0, \forall k > K_1\,. \tag{64}$$

*Proof.* The proofs for positive neurons and for negative neurons are almost identical, we will prove it for positive neurons $\operatorname{sign}(v_j(0)) > 0$, i.e. $\operatorname{sign}(v_j(0)) = 1$. The first part concerns about $\frac{d}{dt}\cos\left(\boldsymbol{w}_j(t), \bar{\boldsymbol{\mu}}_+\right)\big|_{\cos(\boldsymbol{w}_j(t),\bar{\boldsymbol{\mu}}_+)=1-\delta}$.

$$\frac{d}{dt}\cos\left(\boldsymbol{w}_j(t), \bar{\boldsymbol{\mu}}_+\right) \tag{65}$$

$$= \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|}, \bar{\boldsymbol{\mu}}_+\right\rangle \tag{66}$$

$$= \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t)), \bar{\boldsymbol{\mu}}_+\right\rangle + \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \operatorname{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t)), \bar{\boldsymbol{\mu}}_+\right\rangle, \tag{67}$$

**When** $p = 1$, if we can show that for some choice of $\delta > 0$,

$$\inf_{\boldsymbol{w}_j\in\mathbb{S}^{D-1},\cos\left(\boldsymbol{w}_j,\bar{\boldsymbol{\mu}}_+\right)=1-\delta} \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(1)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+\right\rangle := \Delta_1(\delta) > 0\,, \tag{68}$$

then we pick $\epsilon \le \epsilon_0 = \frac{\Delta_1(\delta)}{8\sqrt{h}M^2Kp}$, and by Lemma 1, we have that for $\forall t \le T = \frac{1}{4K}\log\frac{1}{\sqrt{h}\epsilon}$,

$$\frac{d}{dt}\cos\left(\boldsymbol{w}_j(t), \bar{\boldsymbol{\mu}}_+\right)\Big|_{\cos(\boldsymbol{w}_j(t),\bar{\boldsymbol{\mu}}_+)=1-\delta} \ge \Delta_1(\delta) - \max_j\left\|\frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \operatorname{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t))\right\| \tag{69}$$

$$\ge \Delta_1(\delta) - 4\epsilon\sqrt{h}M^2Kp \ge \frac{\Delta_1(\delta)}{2} > 0\,, \tag{70}$$

which is what we stated in the theorem.

Similarly, **When** $p > 3$, if we can show that for some choice of $\delta > 0$,

$$\inf_{\boldsymbol{w}_j\in\mathbb{S}^{D-1},\cos\left(\boldsymbol{w}_j,\bar{\boldsymbol{\mu}}_+\right)=1-\delta} \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(p)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+\right\rangle := \Delta_p(\delta) < 0\,, \tag{71}$$

then we pick $\epsilon \le \epsilon_0 = \frac{\Delta_p(\delta)}{8\sqrt{h}M^2Kp}$, and by Lemma 1, we have that for $\forall t \le T = \frac{1}{4K}\log\frac{1}{\sqrt{h}\epsilon}$,

$$\frac{d}{dt}\cos\left(\boldsymbol{w}_j(t), \bar{\boldsymbol{\mu}}_+\right)\Big|_{\cos(\boldsymbol{w}_j(t),\bar{\boldsymbol{\mu}}_+)=1-\delta} \le \Delta_p(\delta) + \max_j\left\|\frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \operatorname{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t))\right\| \tag{72}$$

$$\le \Delta_p(\delta) + 4\epsilon\sqrt{h}M^2Kp \le \frac{\Delta_p(\delta)}{2} < 0\,. \tag{73}$$

Therefore, for the first part, it suffices to show

$$\inf_{\boldsymbol{w}_j\in\mathbb{S}^{D-1},\cos\left(\boldsymbol{w}_j,\bar{\boldsymbol{\mu}}_+\right)=1-\delta} \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(p)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+\right\rangle := \Delta_p(\delta)\begin{cases} > 0, & \text{for } p = 1 \\ > 0, & \text{for } p \ge 3 \end{cases} \tag{74}$$

Now there exists $1 > \bar{\delta}_1 > 0$ such that when $\delta > \bar{\delta}_1$, and $\cos\left(\boldsymbol{w}_j, \bar{\boldsymbol{\mu}}_+\right) = \sqrt{1-\delta}$, we have $\gamma_k(\boldsymbol{w}_j) = 1, \forall k \le K_1$ and $\gamma_k(\boldsymbol{w}_j) = 0, \forall k > K_1$, i.e., $\boldsymbol{w}_j$ is activated by all $\boldsymbol{x}_k, k \ge K_1$ with positive label and is not activated by any of the $\boldsymbol{x}_k, k > K_1$ with negative label. Moreover, there exists $z_k, k \le K_1$, such that $\boldsymbol{w}_j = \|\boldsymbol{w}_j\|\sum_{k\le K_1} z_k\boldsymbol{x}_k$ and $z_k = \left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle$, i.e., $\boldsymbol{w}_j$ lies completely within the span of $\boldsymbol{x}_k, k \le K_1$.

With this, we have

$$
\begin{aligned}
&\frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(p)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+\right\rangle \\
&= \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \left(I - \frac{\boldsymbol{w}_j\boldsymbol{w}_j^T}{\|\boldsymbol{w}_j\|^2}\right)\sum_k^K \gamma_k(\boldsymbol{w}_j) y_k \boldsymbol{x}_k \cdot p[\cos(\boldsymbol{x}_k, \boldsymbol{w}_j)]^{p-1}, \bar{\boldsymbol{\mu}}_+\right\rangle \\
&= \frac{1}{K}\left\langle \left(I - \frac{\boldsymbol{w}_j\boldsymbol{w}_j^T}{\|\boldsymbol{w}_j\|^2}\right)\sum_{k\le K_1} \boldsymbol{x}_k \cdot p[\cos(\boldsymbol{x}_k, \boldsymbol{w}_j)]^{p-1}, \sum_{i\le K_1}\boldsymbol{x}_k\right\rangle \\
&= \left\langle \sum_{k\le K_1}\boldsymbol{x}_k, \sum_{k\le K_1}\boldsymbol{x}_k\cdot p[\cos(\boldsymbol{x}_k, \boldsymbol{w}_j)]^{p-1}\right\rangle - \left\langle \sum_{k\le K_1}\boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle \sum_{k\le K_1}\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle \cdot p[\cos(\boldsymbol{x}_k, \boldsymbol{w}_j)]^{p-1} \\
&= \left\langle \sum_{k\le K_1}\boldsymbol{x}_k, \sum_{k\le K_1}\boldsymbol{x}_k\cdot p\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle^{p-1}\right\rangle - \left\langle \sum_{k\le K_1}\boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle \sum_{k\le K_1}\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle \cdot p\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle^{p-1} \\
&= \sum_{k\le K_1} p\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle^{p-1} - \left(\sum_{k\le K_1}\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle\right)\left(p\sum_{k\le K_1}\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle^{p}\right) \\
&= p\sum_{k\le K_1}\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle^{p-1} - p\left(\sum_{k\le K_1}\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle\right)\left(\sum_{k\le K_1}\left\langle \boldsymbol{x}_k, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}\right\rangle^{p}\right) \\
&= p\left(\sum_{k\le K_1} z_k^{p-1} - \left(\sum_{k\le K_1} z_k\right)\left(\sum_{k\le K_1} z_k^p\right)\right),
\end{aligned}
$$

Since $\boldsymbol{w}_j$ lies completely within the span of $\boldsymbol{x}_k, k\le K_1$, we have $\sum_{k\le K_1} z_k^2 = 1$, then

$$
\begin{aligned}
&p\left(\sum_{k\le K_1} z_k^{p-1} - \left(\sum_{k\le K_1} z_k\right)\left(\sum_{k\le K_1} z_k^p\right)\right) \\
=&p\left(\left(\sum_{k\le K_1} z_k^{p-1}\right)\left(\sum_{k\le K_1} z_k^2\right) - \left(\sum_{k\le K_1} z_k\right)\left(\sum_{k\le K_1} z_k^p\right)\right) = p\left[g_p(2;\{z_k\}_{k\le K_1}) - g_p(1;\{z_k\}_{k\le K_1})\right],
\end{aligned}
$$

where $g_p(\cdot;\{z_k\}_{k\le K_1})$ is defined in Lemma 5.

By Lemma 5, when $p=1$, $g_1(\cdot;\{z_k\}_{k\le K_1})$ is strictly convex and takes minimum at $q^* = \frac{1+p}{2} = 1$, thus

$$
g_1(2;\{z_k\}_{k\le K_1}) - g_1(1;\{z_k\}_{k\le K_1}) > 0\,, \tag{75}
$$

then we know that

$$
\inf_{\boldsymbol{w}_j\in\mathbb{S}^{D-1}, \cos\left(\boldsymbol{w}_j, \bar{\boldsymbol{\mu}}_+\right)=1-\delta} \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(1)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+\right\rangle := \Delta_1(\delta) \ge 0\,. \tag{76}
$$

However, $\Delta_1(\delta)$ can not be zero: If this is the case, since the set $\{\boldsymbol{w}_j : \boldsymbol{w}_j\in\mathbb{S}^{D-1}, \cos\left(\boldsymbol{w}_j, \bar{\boldsymbol{\mu}}_+\right) = 1-\delta\}$ is compact and $\frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(1)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+\right\rangle$ is continuous on this set. It attains minimum 0 at some $\boldsymbol{w}_j$, which implies the non-strong convexity of $g_1(\cdot;\{z_k\}_{k\le K_1})$ that, by Lemma 5, requires all $z_k, k\le K_1$ to be equal to each other (This can only happen if $\cos\left(\boldsymbol{w}_j, \bar{\boldsymbol{\mu}}_+\right) = 1$ ). Contradiction. Then one must have

$$
\inf_{\boldsymbol{w}_j\in\mathbb{S}^{D-1}, \cos\left(\boldsymbol{w}_j, \bar{\boldsymbol{\mu}}_+\right)=1-\delta} \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|}\left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(1)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+\right\rangle := \Delta_1(\delta) > 0\,. \tag{77}
$$

Similarly, by Lemma 5, when $p\ge 3$, $g_1(\cdot;\{z_k\}_{k\le K_1})$ is strictly convex and takes minimum at $q^* = \frac{1+p}{2} \ge 2$, thus

$$
g_1(2;\{z_k\}_{k\le K_1}) - g_1(1;\{z_k\}_{k\le K_1}) < 0\,, \tag{78}
$$

then we know that

$$\inf_{\boldsymbol{w}_j \in \mathbb{S}^{D-1}, \cos\left(\boldsymbol{w}_j, \bar{\boldsymbol{\mu}}_+\right)=1-\delta} \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|} \left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(1)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+ \right\rangle := \Delta_p(\delta) \leq 0\,. \tag{79}$$

Using the same argument, we eliminate the case of this infimum being zero. Then one must have

$$\inf_{\boldsymbol{w}_j \in \mathbb{S}^{D-1}, \cos\left(\boldsymbol{w}_j, \bar{\boldsymbol{\mu}}_+\right)=1-\delta} \frac{1}{\|\bar{\boldsymbol{\mu}}_+\|} \left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(1)}(\boldsymbol{w}_j), \bar{\boldsymbol{\mu}}_+ \right\rangle := \Delta_p(\delta) < 0\,. \tag{80}$$

The second part concerns about $\frac{d}{dt}\cos\left(\boldsymbol{w}_j(t), \boldsymbol{\mu}_k\right)\big|_{\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_k)=1-\delta}$, for some $k \leq K_1$. Without loss of generality, we let $k=1$. Thus we intend to show $\frac{d}{dt}\cos\left(\boldsymbol{w}_j(t), \boldsymbol{\mu}_1\right)\big|_{\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_1)=1-\delta}$ is positive.

We also let $\zeta := 1-(1-\delta)^2$, so that the condition $\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_1) = 1-\delta$ becomes $\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_1) = \sqrt{1-\zeta}$.

Since $\sum_{k=1}^{K} \left|\left\langle \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}, \boldsymbol{\mu}_k \right\rangle\right|^2 \leq 1$, $\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_1) = \sqrt{1-\zeta}$ implies $\sum_{l=2}^{K} \left|\left\langle \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}, \boldsymbol{\mu}_l \right\rangle\right|^2 \leq \zeta$.

Similar to the first part of the proof, we have

$$\frac{d}{dt}\cos\left(\boldsymbol{w}_j(t), \boldsymbol{\mu}_1\right) \tag{81}$$

$$= \left\langle \frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|}, \boldsymbol{\mu}_1 \right\rangle \tag{82}$$

$$= \left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t)), \boldsymbol{\mu}_1 \right\rangle + \left\langle \frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \operatorname{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t)), \boldsymbol{\mu}_1 \right\rangle\,, \tag{83}$$

if we can show that for some choice of $\delta > 0$ (or equivalently some $\zeta > 0$),

$$\inf_{\boldsymbol{w}_j \in \mathbb{S}^{D-1}, \cos(\boldsymbol{w}_j, \boldsymbol{\mu}_1)=1-\delta} \left\langle \mathcal{P}^{\perp}_{\boldsymbol{w}_j} x^{(1)}(\boldsymbol{w}_j), \boldsymbol{\mu}_1 \right\rangle := \Lambda_p(\delta) > 0\,, \tag{84}$$

then we pick $\epsilon \leq \epsilon_0 = \frac{\Lambda_p(\delta)}{8\sqrt{h}M^2Kp}$, and by Lemma 1, we have that for $\forall t \leq T = \frac{1}{4K}\log\frac{1}{\sqrt{h}\epsilon}$,

$$\frac{d}{dt}\cos\left(\boldsymbol{w}_j(t), \boldsymbol{\mu}_1\right)\Big|_{\cos(\boldsymbol{w}_j(t), \boldsymbol{\mu}_1)=1-\delta} \geq \Lambda_p(\delta) - \max_j \left\| \frac{d}{dt}\frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} - \operatorname{sign}(v_j(0))\mathcal{P}^{\perp}_{\boldsymbol{w}_j(t)} x^{(p)}(\boldsymbol{w}_j(t)) \right\| \tag{85}$$

$$\geq \Lambda_p(\delta) - 4\epsilon\sqrt{h}M^2Kp \geq \frac{\Lambda_p(\delta)}{2} > 0\,, \tag{86}$$

which is what we stated in the theorem. The remaining proof is to find a lower bound on $\Lambda_p(\delta)$, which is given by

$$
\begin{aligned}
&\left\langle \mathcal{P}_{\boldsymbol{w}_j}^{\perp} x^{(p)}(\boldsymbol{w}_j), \boldsymbol{\mu}_1 \right\rangle \\
&= \left\langle \left( I - \frac{\boldsymbol{w}_j \boldsymbol{w}_j^T}{\|\boldsymbol{w}_j\|^2} \right) \sum_k^K \gamma_k(\boldsymbol{w}_j) y_k \boldsymbol{x}_k \cdot p[\cos(\boldsymbol{x}_k, \boldsymbol{w}_j)]^{p-1}, \boldsymbol{\mu}_1 \right\rangle \\
&= \left\langle \left( I - \frac{\boldsymbol{w}_j \boldsymbol{w}_j^T}{\|\boldsymbol{w}_j\|^2} \right) \sum_{k \le K_1} \gamma_k(\boldsymbol{w}_j) y_k \boldsymbol{x}_k \cdot p[\cos(\boldsymbol{x}_k, \boldsymbol{w}_j)]^{p-1}, \boldsymbol{\mu}_1 \right\rangle \\
&= p \cos^{p-1}(\boldsymbol{\mu}_1, \boldsymbol{w}_j) \left(1 - \cos^2(\boldsymbol{\mu}_1, \boldsymbol{w}_j)\right) + \sum_{l=2}^K \gamma_l(\boldsymbol{w}_j) y_l p \cos^p(\boldsymbol{\mu}_l, \boldsymbol{w}_j) \cos(\boldsymbol{w}_j, \boldsymbol{\mu}_1) \\
&= \sqrt{1-\zeta} \left( p(1-\zeta)^{\frac{p-2}{2}} \zeta + \sum_{l=2}^K \gamma_l(\boldsymbol{w}_j) y_l p \cos^p(\boldsymbol{\mu}_l, \boldsymbol{w}_j) \right) \\
&\ge \sqrt{1-\zeta} p \left( (1-\zeta)^{\frac{p-2}{2}} \zeta - \sum_{l=2}^K \left| \left\langle \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}, \boldsymbol{\mu}_l \right\rangle \right|^p \right) \\
&\ge \sqrt{1-\zeta} p \left( (1-\zeta)^{\frac{p-2}{2}} \zeta - \left( \sum_{l=2}^K \left| \left\langle \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|}, \boldsymbol{\mu}_l \right\rangle \right|^2 \right)^{\frac{p}{2}} \right) \\
&\ge \sqrt{1-\zeta} p \left( (1-\zeta)^{\frac{p-2}{2}} \zeta - \zeta^{\frac{p}{2}} \right) \ge \sqrt{1-\zeta} p \left( \left( 1 - \frac{p-2}{2} \zeta \right) \zeta - \zeta^{\frac{p}{2}} \right) = p\zeta + o(\zeta) \,.
\end{aligned}
$$

Therefore, as long as $\zeta > 0$ is small enough, which can be achieved by picking some $\delta < \bar{\delta}_2 < 1$, then $\inf_{\boldsymbol{w}_j \in \mathbb{S}^{D-1}, \cos(\boldsymbol{w}_j, \boldsymbol{\mu}_1) = 1-\delta} \left\langle \mathcal{P}_{\boldsymbol{w}_j}^{\perp} x^{(1)}(\boldsymbol{w}_j), \boldsymbol{\mu}_1 \right\rangle = \Lambda_p(\delta)$ is positive. □

### D.5. Proof for Auxiliary Lemmas

**Balancedness**: Under GF, balancedness (Du et al., 2018) is preserved: $v_j^2(t) - \|\boldsymbol{w}_j(t)\|^2 = 0, \forall t \ge 0, \forall j \in [h]$, from the fact that:

$$
\begin{aligned}
\frac{d}{dt} \|\boldsymbol{w}_j\|^2 &= \langle \boldsymbol{w}_j, \dot{w}_j \rangle \\
&= -2 \sum_{k=1}^K \gamma_k(\boldsymbol{w}_j) \nabla_{\hat{y}} \ell(y_k, f_p(\boldsymbol{x}_k; W, v)) v_j \left( \frac{p[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j \rangle)]^{p-1}}{\|\boldsymbol{w}_j\|^{p-1}} \langle \boldsymbol{w}_j, \boldsymbol{x}_k \rangle - (p-1) \frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j \rangle)]^p}{\|\boldsymbol{w}_j\|^{p+1}} \|\boldsymbol{w}_j\|^2 \right) \\
&= -2 \sum_{k=1}^K \gamma_k(\boldsymbol{w}_j) \nabla_{\hat{y}} \ell(y_k, f_p(\boldsymbol{x}_k; W, v)) v_j \left( \frac{p[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j \rangle)]^{p-1}}{\|\boldsymbol{w}_j\|^{p}} - (p-1) \frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j \rangle)]^p}{\|\boldsymbol{w}_j\|^{p-1}} \right) \\
&= -2 \sum_{k=1}^K \gamma_k(\boldsymbol{w}_j) \nabla_{\hat{y}} \ell(y_k, f_p(\boldsymbol{x}_k; W, v)) v_j \frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j \rangle)]^p}{\|\boldsymbol{w}_j\|^{p-1}} \\
&= \frac{d}{dt} v_j^2
\end{aligned}
$$

In addition, $\operatorname{sign}(v_j(t)) = \operatorname{sign}(v_j(0)), \forall t \ge 0, \forall j \in [h]$, and the dynamical behaviors of neurons will be divided into two types, depending on $\operatorname{sign}(v_j(0))$. Therefore, throughout the gradient flow trajectory, we have $v_j = \operatorname{sign}(v_j(0)) \|\boldsymbol{w}_j\|$. This fact will be used in the subsequent proof.

*Proof for Lemma 4.* Under gradient flow, we have

$$
\frac{d}{dt} \boldsymbol{w}_j = - \sum_{k=1}^K \gamma_k(\boldsymbol{w}_j) \nabla_{\hat{y}} \ell(y_k, f_p(\boldsymbol{x}_k; W, v)) v_j \left( \frac{p[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j \rangle)]^{p-1}}{\|\boldsymbol{w}_j\|^{p-1}} \boldsymbol{x}_k - (p-1) \frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j \rangle)]^p}{\|\boldsymbol{w}_j\|^{p+1}} \boldsymbol{w}_j \right) . \tag{87}
$$

and for $\|\boldsymbol{w}_j\|$,

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{w}_j\|^2 =& \langle \boldsymbol{w}_j, \dot{w}_j\rangle \\
=& -2\sum_{k=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f_p(\boldsymbol{x}_k; W, v))v_j\left(\frac{p[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j\rangle)]^{p-1}}{\|\boldsymbol{w}_j\|^{p-1}}\langle \boldsymbol{w}_j, \boldsymbol{x}_k\rangle - (p-1)\frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j\rangle)]^p}{\|\boldsymbol{w}_j\|^{p+1}}\|\boldsymbol{w}_j\|^2\right) \\
=& -2\sum_{k=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f_p(\boldsymbol{x}_k; W, v))v_j\left(\frac{p[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j\rangle)]^{p-1}}{\|\boldsymbol{w}_j\|^{p}} - (p-1)\frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j\rangle)]^p}{\|\boldsymbol{w}_j\|^{p-1}}\right) \\
=& -2\sum_{k=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f_p(\boldsymbol{x}_k; W, v))v_j\frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j\rangle)]^p}{\|\boldsymbol{w}_j\|^{p-1}}
\end{aligned}$$

Balanced initialization enforces $v_j = \text{sign}(v_j(0))\|\boldsymbol{w}_j\|$, hence

$$\frac{d}{dt}\|\boldsymbol{w}_j\|^2 = -2\sum_{k=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f_p(\boldsymbol{x}_k; W, v))\text{sign}(v_j(0))\|\boldsymbol{w}_j\|^2\frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j\rangle)]^p}{\|\boldsymbol{w}_j\|^p}\,. \tag{88}$$

Let $T := \inf\{t:\ \max_i |f(\boldsymbol{x}_k; W(t), v(t))| > 2\epsilon\sqrt{h}M^2\}$, then $\forall t \le T, j \in [h]$, we have

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{w}_j\|^2 =& -2\sum_{k=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f_p(\boldsymbol{x}_k; W, v))\text{sign}(v_j(0))\|\boldsymbol{w}_j\|^2\frac{[\sigma(\langle \boldsymbol{x}_k, \boldsymbol{w}_j\rangle)]^p}{\|\boldsymbol{w}_j\|^p} \\
=& -2\sum_{k=1}^{K}\gamma_k(\boldsymbol{w}_j)\nabla_{\hat{y}}\ell(y_k, f(\boldsymbol{x}_k; W, v))\text{sign}(v_j(0))\|\boldsymbol{w}_j\|^2\frac{(\langle \boldsymbol{x}_k, \boldsymbol{w}_j\rangle)^p}{\|\boldsymbol{w}_j\|^p} \\
\le& \ 2\sum_{k=1}^{K}|\nabla_{\hat{y}}\ell(y_k, f(\boldsymbol{x}_k; W, v))|\,\|\boldsymbol{w}_j\|^2 \\
\le& \ 2\sum_{k=1}^{K}(|y_k| + 2|f(\boldsymbol{x}_k; W, v)|)\|\boldsymbol{w}_j\|^2 \\
\le& \ 2\sum_{k=1}^{K}(1 + 4\epsilon\sqrt{h}M^2)\|\boldsymbol{w}_j\| \\
\le& \ 2n(+4\epsilon\sqrt{h}M^2)\|\boldsymbol{w}_j\|^2\,.
\end{aligned} \tag{89}$$

Let $\tau_j := \inf\{t : \|\boldsymbol{w}_j(t)\|^2 > \frac{2\epsilon M^2}{\sqrt{h}}\}$, and let $j^* := \arg\min_j \tau_j$, then $\tau_{j^*} = \min_j \tau_j \le T$ due to the fact that

$$|f(x_i; W, v)| = \left|\sum_{j\in[h]}\mathbb{1}_{\langle \boldsymbol{w}_j, \boldsymbol{x}_k\rangle > 0}v_j\frac{(\langle \boldsymbol{w}_j, \boldsymbol{x}_k\rangle)^p}{\|\boldsymbol{w}_j\|^p}\right| \le \sum_{j\in[h]}\|\boldsymbol{w}_j\|^2 \le h\max_{j\in[h]}\|\boldsymbol{w}_j\|^2\,,$$

which implies "$|f(\boldsymbol{x}_k; W(t), v(t))| > 2\epsilon\sqrt{h}M^2 \Rightarrow \exists j, s.t. \|\boldsymbol{w}_j(t)\|^2 > \frac{2\epsilon M^2}{\sqrt{h}}$".

Then for $t \le \tau_{j^*}$, we have

$$\frac{d}{dt}\|w_{j^*}\|^2 \le 2n(+4\epsilon\sqrt{h}M^2)\|w_{j^*}\|^2\,. \tag{90}$$

By Grönwall's inequality, we have $\forall t \le \tau_{j^*}$

$$\begin{aligned}
\|w_{j^*}(t)\|^2 \le& \ \exp\left(2n(+4\epsilon\sqrt{h}M^2)t\right)\|w_{j^*}(0)\|^2\,, \\
=& \ \exp\left(2n(+4\epsilon\sqrt{h}M^2)t\right)\epsilon^2\|[W_0]_{:,j^*}\|^2 \\
\le& \ \exp\left(2n(+4\epsilon\sqrt{h}M^2)t\right)\epsilon^2 M^2\,.
\end{aligned}$$

Suppose $\tau_{j^*} < \frac{1}{4n}\log\left(\frac{1}{\sqrt{h}\epsilon}\right)$, then by the continuity of $\|w_{j^*}(t)\|^2$, we have

$$\begin{aligned}\frac{2\epsilon M^2}{\sqrt{h}} \le \|w_{j^*}(\tau_{j^*})\|^2 &\le \exp\left(2n(+4\epsilon\sqrt{h}M^2)\tau_{j^*}\right)\epsilon^2 M^2 \\ &\le \exp\left(2n(+4\epsilon\sqrt{h}M^2)\frac{1}{4n}\log\left(\frac{1}{\sqrt{h}\epsilon}\right)\right)\epsilon^2 M^2 \\ &\le \exp\left(\frac{1+4\epsilon\sqrt{h}M^2}{2}\log\left(\frac{1}{\sqrt{h}\epsilon}\right)\right)\epsilon^2 M^2 \\ &\le \exp\left(\log\left(\frac{1}{\sqrt{h}\epsilon}\right)\right)\epsilon^2 M^2 = \frac{\epsilon M^2}{\sqrt{h}}\,,\end{aligned}$$

which leads to a contradiction $2\epsilon \le \epsilon$. Therefore, one must have $T \ge \tau_{j^*} \ge \frac{1}{4n}\log\left(\frac{1}{\sqrt{h}\epsilon}\right)$. This finishes the proof. □

*Proof of Lemma 5.* Since

$$g_p'(q;\{z_i\}_{i=1}^n) = \left(\sum_{k=1}^K z_i^q \log z_i\right)\left(\sum_{k=1}^K z_i^{p+1-q}\right) - \left(\sum_{k=1}^K z_i^q\right)\left(\sum_{k=1}^K z_i^{p+1-q}\log z_i\right)\,, \tag{91}$$

we immediately find $g_p'(q^*;\{z_i\}_{i=1}^n) = 0$. Now we compute the second-order derivative

$$\begin{aligned}g_p''(q;\{z_i\}_{i=1}^n) &= -2\left(\sum_{k=1}^K z_i^q \log z_i\right)\left(\sum_{k=1}^K z_i^{p+1-q}\log z_i\right) \\ &\quad + \left(\sum_{k=1}^K z_i^q\right)\left(\sum_{k=1}^K z_i^{p+1-q}\log^2 z_i\right) + \left(\sum_{k=1}^K z_i^q \log^2 z_i\right)\left(\sum_{k=1}^K z_i^{p+1-q}\right) \\ &= \sum_{1\le i,j\le n} z_i^q z_j^{p+1-q}(-2\log z_i \log z_j) \\ &\quad + \sum_{1\le i,j\le n} z_i^q z_j^{p+1-q}\log^2 z_j + \sum_{1\le i,j\le n} z_i^q z_j^{p+1-q}\log^2 z_i \\ &= \sum_{1\le i,j\le n} z_i^q z_j^{p+1-q}(\log z_i - \log z_j)^2 \ge 0\,,\end{aligned}$$

and the equality holds only when $z_1 = \cdots = z_n$. The desired results follow. □