The Double-Edged Sword of Implicit Bias: Generalization vs. Robustness in ReLU Networks

Spencer Frei*
UC Davis
sfrei@ucdavis.edu

Gal Vardi*
TTI-Chicago and Hebrew University
galvardi@ttic.edu

Peter L. Bartlett
UC Berkeley and Google DeepMind
peter@berkeley.edu

Nathan Srebro
TTI-Chicago
nati@ttic.edu

Abstract

In this work, we study the implications of the implicit bias of gradient flow on generalization and adversarial robustness in ReLU networks. We focus on a setting where the data consists of clusters and the correlations between cluster means are small, and show that in two-layer ReLU networks gradient flow is biased towards solutions that generalize well, but are vulnerable to adversarial examples. Our results hold even in cases where the network is highly overparameterized. Despite the potential for harmful overfitting in such settings, we prove that the implicit bias of gradient flow prevents it. However, the implicit bias also leads to non-robust solutions (susceptible to small adversarial ℓ_2-perturbations), even though robust networks that fit the data exist.

1 Introduction

A central question in the theory of deep learning is how neural networks can generalize even when trained without any explicit regularization, and when there are more learnable parameters than training examples. In such optimization problems there are many solutions that label the training data correctly, and gradient descent seems to prefer solutions that generalize well [Zha+17]. Thus, it is believed that gradient descent induces an *implicit bias* towards solutions which enjoy favorable properties [Ney+17]. Characterizing this bias in various settings has been a subject of extensive research in recent years, but it is still not well understood when the implicit bias provably implies generalization in non-linear neural networks.

An additional intriguing phenomenon in deep learning is the abundance of *adversarial examples* in trained neural networks. In a seminal paper, Szegedy et al. [Sze+14] observed that deep networks are extremely vulnerable to adversarial examples, namely, very small perturbations to the inputs can significantly change the predictions. This phenomenon has attracted considerable interest, and various attacks (e.g., [GSS15; CW17; Pap+17; ACW18; CW18; Wu+20]) and defenses (e.g., [Pap+16; KGB17; Mad+18; WK18; CH20; WRK20]) were developed. However, the fundamental principles underlying the existence of adversarial examples are still unclear, and it is believed that for most tasks where trained neural networks suffer from a vulnerability to adversarial attacks, there should exist other neural networks which can be robust to such attacks. This is suggestive of the possible role of the optimization algorithms used to train neural networks in the existence of adversarial examples.

In this work, we study the implications of the implicit bias for generalization and robustness in ReLU networks, in a setting where the data consists of clusters (i.e., Gaussian mixture model) and the

*Equal contribution.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

correlations between cluster means are small. We show that in two-layer ReLU networks trained with the logistic loss or the exponential loss, gradient flow is biased towards solutions that generalize well, albeit they are non-robust. Our results are independent of the network width, and hence they hold even where the network has significantly more parameters than training examples. In such an overparameterized setting, one might expect harmful overfitting to occur, but we prove that the implicit bias of gradient flow prevents it. On the flip side, in our setting the distances between clusters are large, and thus one might hope that gradient flow will converge to a robust network. However, we show that the implicit bias leads to non-robust solutions.

Our results rely on known properties of the implicit bias in two-layer ReLU networks trained with the logistic or the exponential loss, which were shown by Lyu and Li [LL20] and Ji and Telgarsky [JT20]. They proved that if gradient flow in homogeneous models (which include two-layer ReLU networks) with such losses reaches a small training loss, then it converges (in direction) to a KKT point of the maximum-margin problem in parameter space. We show that in clustered data distributions, with high probability over the training dataset, every network that satisfies the KKT conditions of the maximum-margin problem generalizes well but is non-robust. Thus, instead of analyzing the trajectory of gradient flow directly in the complex setting of training two-layer ReLU networks, we demonstrate that investigating the KKT points is a powerful tool for understanding generalization and robustness. We emphasize that our results hold in the *rich* (i.e., *feature learning*) regime, namely, the neural network training does not lie in the kernel regime, and thus we provide guarantees which go beyond the analysis achieved using NTK-based results.

In a bit more detail, our main contributions are the following:

- Suppose that the data distribution consists of k clusters, and the training dataset is of size $n \geq \tilde{\Omega}(k)$. We show that with high probability over the size-n dataset, if gradient flow achieves training loss smaller than $\frac{1}{n}$ at some time t_0, then it converges in direction to a network that generalizes well (i.e., has a small test error). Thus, gradient-flow-trained networks cannot harmfully overfit even if the network is highly overparameterized. The sample complexity $\tilde{\Omega}(k)$ in this result is optimal (up to log factors), since we cannot expect to perform well on unseen data using a training dataset that does not include at least one example from each cluster.
- In the same setting as above, we prove that gradient flow converges in direction to a non-robust network, even though there exist robust networks that classify the data correctly. Specifically, we consider data distributions on $\mathbb{R}^d$ such that the distance between every pair of clusters is $\Omega(\sqrt{d})$, and we show that there exists a two-layer ReLU network where flipping the output sign of a test example requires w.h.p. an ℓ_2-perturbation of size $\Omega(\sqrt{d})$, but gradient flow converges to a network where we can flip the output sign of a test example with an ℓ_2-perturbation of size much smaller than $\sqrt{d}$. Moreover, the adversarial perturbation depends only on the data distribution, and not on the specific test example or trained neural network. Thus, the perturbation is both *universal* [Moo+17; Zha+21] and *transferable* [Liu+17; AM18]. We argue that clustered data distributions are a natural setting for analyzing the tendency of gradient methods to converge to non-robust solutions. Indeed, if positive and negative examples are not well-separated (i.e., the distances between points with opposite labels are small), then robust solutions do not exist. Thus, in order to understand the role of the optimization algorithm, we need a setting with sufficient separation between positive and negative examples.

The remainder of this paper is structured as follows: Below we discuss related work. In Section 2 we provide necessary notations and background, and introduce our setting and assumptions. In Sections 3 and 4 we state our main results on generalization and robustness (respectively), and provide some proof ideas, with all formal proofs deferred to the appendix. We conclude with a short discussion (Section 5).

Related work

Implicit bias in neural networks. The literature on implicit bias in neural networks has rapidly expanded in recent years, and cannot be reasonably surveyed here (see Vardi [Var22] for a survey). In what follows, we discuss results that apply to two-layer ReLU or leaky-ReLU networks trained with gradient flow in classification settings.

By Lyu and Li [LL20] and Ji and Telgarsky [JT20], homogeneous neural networks (and specifically two-layer ReLU networks, which are the focus of this paper) trained with exponentially-tailed classification losses converge in direction to a KKT point of the maximum-margin problem. Our analysis of the implicit bias relies on this result. We note that the aforementioned KKT point may not be a global optimum of the maximum-margin problem [VSS22]. Recently, Kunin et al. [Kun+22] extended this result by showing bias towards margin maximization in a broader family of networks called *quasi-homogeneous*.

Lyu et al. [Lyu+21], Sarussi, Brutzkus, and Globerson [SBG21], and Frei et al. [Fre+23b] studied implicit bias in two-layer leaky-ReLU networks with linearly-separable data, and proved that under some additional assumptions gradient flow converges to a linear classifier. Chizat and Bach [CB20] studied the dynamics of gradient flow on infinite-width homogeneous two-layer networks with exponentially-tailed losses, and showed bias towards margin maximization w.r.t. a certain function norm known as the variation norm. Phuong and Lampert [PL20] studied the implicit bias in two-layer ReLU networks trained on *orthogonally separable data*.

Safran, Vardi, and Lee [SVL22] proved implicit bias towards minimizing the number of linear regions in univariate two-layer ReLU networks, and used this result to obtain generalization bounds. Similarly to our work, they used the KKT conditions of the maximum-margin problem in parameter space to prove generalization in overparameterized networks. However, our setting is significantly different. Implications of the bias towards KKT points of the maximum-margin problem were also studied in Haim et al. [Hai+22], where they showed that this implicit bias can be used for reconstructing training data from trained ReLU networks.

Theoretical explanations for non-robustness in neural networks. Despite much research, the reasons for the abundance of adversarial examples in trained networks are still unclear [GSS15; FFF18; Sha+19; Sch+18; KH18; Bub+19; AL21; Wan+20; Sha+20; SMB21; Sin+21; Wan+22; DB22]. Below we discuss several prior theoretical works on this question.

In one line of work, it has been shown that small adversarial perturbations can be found for any fixed input in certain neural networks with random weights (drawn from the Gaussian distribution) [DS20; Bub+21; BBC21; MW22]. These works differ in the assumptions about the width and depth of the networks as well as the activation functions considered. However, since trained networks are non-random, these works are unable to capture the existence of adversarial examples in trained networks.

The result closest to ours was shown in Vardi, Yehudai, and Shamir [VYS22]. Similarly to our result, they used the KKT conditions of the maximum-margin problem in parameter space, in order to prove that gradient flow converges to non-robust two-layer ReLU networks under certain assumptions. More precisely, they considered a setting where the training dataset $\mathcal{S}$ consists of nearly-orthogonal points, and proved that every KKT point is non-robust w.r.t. $\mathcal{S}$. Namely, for every two-layer network that satisfies the KKT conditions of the maximum-margin problem, and every point $\mathbf{x}_i$ from $\mathcal{S}$, it is possible to flip the output's sign with a small perturbation. Their result has two main limitations: (1) It considers robustness w.r.t. the training data, while the more common setting in the literature considers robustness w.r.t. test data, as it is often more crucial to avoid adversarial perturbations in test examples; (2) Since they assume near orthogonality of the training data, the size of the dataset $\mathcal{S}$ must be smaller than the input dimension.[2] Thus, they considered a *high dimensional* setting. We note that high-dimensional settings often have a different generalization behavior than low-dimensional settings (e.g., overfitting can be *benign* in the high-dimensional setting, but harmful in a low-dimensional setting [KYS23]). Our result does not suffer from these limitations, since we consider robustness w.r.t. test data, and the size of our training dataset might be very large. In our results, we essentially require near orthogonality of the cluster means, as opposed to near orthogonality of the training dataset in their result.

Finally, in Bubeck, Li, and Nagaraj [BLN21] and Bubeck and Sellke [BS21], the authors proved (under certain assumptions) that overparameterization is necessary if one wants to interpolate training data using a neural network with a small Lipschitz constant. Namely, neural networks with a small number of parameters are not expressive enough to interpolate the training data while having a small

[2] They also give a version of their result, where instead of assuming this upper bound on the size of the dataset, they assume an upper bound on the number of points that attain the margin in the trained network, but it is not clear a priori when this assumption is likely to hold.

Lipschitz constant. These results suggest that overparameterization might be necessary for robustness. In this work, we show that even if the network is highly overparameterized, the implicit bias of the optimization method can prevent convergence to robust solutions.

2 Preliminaries

We use bold-face letters to denote vectors, e.g., $\mathbf{x} = (x_1, \ldots, x_d)$. For $\mathbf{x} \in \mathbb{R}^d$ we denote by $\|\mathbf{x}\|$ the Euclidean norm. We denote by $\mathbb{1}[\cdot]$ the indicator function, for example $\mathbb{1}[t \geq 5]$ equals 1 if $t \geq 5$ and 0 otherwise. We denote $\mathrm{sign}(z) = 1$ if $z > 0$ and -1 otherwise. For an integer $d \geq 1$ we denote $[d] = \{1, \ldots, d\}$. For a set A we denote by $\mathcal{U}(A)$ the uniform distribution over A. We denote by $\mathsf{N}(\mu, \sigma^2)$ the normal distribution with mean $\mu \in \mathbb{R}$ and variance σ^2, and by $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$ the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The identity matrix of size d is denoted by I_d. We use standard asymptotic notation $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ to hide constant factors, and $\tilde{\mathcal{O}}(\cdot), \tilde{\Omega}(\cdot)$ to hide logarithmic factors. We use $\log$ for the logarithm with base 2 and $\ln$ for the natural logarithm.

In this work, we consider depth-2 ReLU neural networks. The ReLU activation function is defined by $\phi(z) = \max\{0, z\}$. Formally, a depth-2 network $\mathcal{N}_{\boldsymbol{\theta}}$ of width m is parameterized by $\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{b}, \mathbf{v}]$ where $\mathbf{w}_i \in \mathbb{R}^d$ for all $i \in [m]$ and $\mathbf{b}, \mathbf{v} \in \mathbb{R}^m$, and for every input $\mathbf{x} \in \mathbb{R}^d$ we have

$$\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) = \textstyle\sum_{j \in [m]} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) \ .$$

We sometimes view $\boldsymbol{\theta}$ as the vector obtained by concatenating the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{b}, \mathbf{v}$. Thus, $\|\boldsymbol{\theta}\|$ denotes the ℓ_2 norm of the vector $\boldsymbol{\theta}$. We note that in this work we train both layers of the ReLU network.

We denote $\Phi(\boldsymbol{\theta}; \mathbf{x}) := \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x})$. We say that a network is *homogeneous* if there exists $L > 0$ such that for every $\alpha > 0$ and $\boldsymbol{\theta}, \mathbf{x}$ we have $\Phi(\alpha\boldsymbol{\theta}; \mathbf{x}) = \alpha^L \Phi(\boldsymbol{\theta}; \mathbf{x})$. Note that depth-2 ReLU networks as defined above are homogeneous (with $L = 2$).

We next define gradient flow and remind the reader of some recent results on the implicit bias of gradient flow in two-layer ReLU networks. Let $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{-1, 1\}$ be a binary classification training dataset. Let $\Phi(\boldsymbol{\theta}; \cdot) : \mathbb{R}^d \to \mathbb{R}$ be a neural network parameterized by $\boldsymbol{\theta}$. For a loss function $\ell : \mathbb{R} \to \mathbb{R}$ the *empirical loss* of $\Phi(\boldsymbol{\theta}; \cdot)$ on the dataset $\mathcal{S}$ is

$$\mathcal{L}(\boldsymbol{\theta}) := \tfrac{1}{n} \textstyle\sum_{i=1}^n \ell(y_i \Phi(\boldsymbol{\theta}; \mathbf{x}_i)) \ . \tag{1}$$

We focus on the exponential loss $\ell(q) = e^{-q}$ and the logistic loss $\ell(q) = \log(1 + e^{-q})$.

We consider gradient flow on the objective given in Eq. (1). This setting captures the behavior of gradient descent with an infinitesimally small step size. Let $\boldsymbol{\theta}(t)$ be the trajectory of gradient flow. Starting from an initial point $\boldsymbol{\theta}(0)$, the dynamics of $\boldsymbol{\theta}(t)$ is given by the differential equation $\frac{d\boldsymbol{\theta}(t)}{dt} \in -\partial^\circ \mathcal{L}(\boldsymbol{\theta}(t))$. Here, ∂° denotes the *Clarke subdifferential* [Cla+08], which is a generalization of the derivative for non-differentiable functions.

We now remind the reader of a recent result concerning the implicit bias of gradient flow over the exponential and logistic losses for homogeneous neural networks. Note that since homogeneous networks satisfy $\mathrm{sign}(\Phi(\alpha\boldsymbol{\theta}; \mathbf{x})) = \mathrm{sign}(\Phi(\boldsymbol{\theta}; \mathbf{x}))$ for any $\alpha > 0$, the sign of the network output of homogeneous networks depends only on the direction of the parameters $\boldsymbol{\theta}$. The following theorem provides a characterization of the implicit bias of gradient flow by showing that the trajectory of the weights $\boldsymbol{\theta}(t)$ *converge in direction* to a first-order stationary point of a particular constrained optimization problem, where $\boldsymbol{\theta}$ *converges in direction* to $\tilde{\boldsymbol{\theta}}$ means $\lim_{t \to \infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|} = \frac{\tilde{\boldsymbol{\theta}}}{\|\tilde{\boldsymbol{\theta}}\|}$. Note that since ReLU networks are non-smooth, the first-order stationarity conditions (i.e., the Karush–Kuhn–Tucker conditions, or KKT conditions for short) are defined using the Clarke subdifferential (see Lyu and Li [LL20] and Dutta et al. [Dut+13] for more details on the KKT conditions in non-smooth optimization problems).

Theorem 2.1 (Paraphrased from Lyu and Li [LL20] and Ji and Telgarsky [JT20])**.** *Let $\Phi(\boldsymbol{\theta}; \cdot)$ be a homogeneous ReLU neural network parameterized by $\boldsymbol{\theta}$. Consider minimizing either the exponential or the logistic loss over a binary classification dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ using gradient flow. Assume that there exists time t_0 such that $\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{1}{n}$ (and thus $y_i \Phi(\boldsymbol{\theta}(t_0); \mathbf{x}_i) > 0$ for every $\mathbf{x}_i$). Then, gradient flow converges in direction to a first-order stationary point (KKT point) of the following* *maximum margin problem in parameter space:*

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{\theta}\|^2 \quad \textit{s.t.} \quad \forall i \in [n] \;\; y_i \Phi(\boldsymbol{\theta}; \mathbf{x}_i) \geq 1 \; . \tag{2}$$

Moreover, $\mathcal{L}(\boldsymbol{\theta}(t)) \to 0$ and $\|\boldsymbol{\theta}(t)\| \to \infty$ as $t \to \infty$.

Theorem 2.1 gives a characterization of the implicit bias of gradient flow with the exponential and the logistic loss for homogeneous ReLU networks. Note that the theorem makes no assumption on the initialization, training data, or number of parameters in the network; the only requirement is that the network is homogeneous and that at some time point in the gradient flow trajectory, the network is able to achieve small training loss. The theorem shows that although there are many ways to configure the network parameters to achieve small training loss (via overparameterization), gradient flow only converges (in direction) to networks which satisfy the KKT conditions of Problem (2). It is important to note that satisfaction of the KKT conditions is not sufficient for global optimality of the constrained optimization problem [VSS22]. We further note that if the training data are sampled i.i.d. from a distribution with label noise (e.g., a class-conditional Gaussian mixture model, or a distribution where labels y_i are flipped to $-y_i$ with some nonzero probability), networks which have parameters that are feasible w.r.t. the constraints of Problem (2) have overfit to noise, and understanding the generalization behavior of even globally optimal solutions to Problem (2) in this setting is the subject of significant research [Mon+19; CL21; Fre+23a].

Finally, we introduce the distributional setting that we consider. We consider a distribution $\mathcal{D}_{\text{clusters}}$ on $\mathbb{R}^d \times \{-1, 1\}$ that consists of k clusters with means $\boldsymbol{\mu}^{(1)}, \ldots, \boldsymbol{\mu}^{(k)} \in \mathbb{R}^d$ and covariance $\sigma^2 I_d$ (i.e., a Gaussian mixture model), such that the examples in the j-th cluster are labeled by $y^{(j)} \in \{-1, 1\}$. More formally, $(\mathbf{x}, y) \sim \mathcal{D}_{\text{clusters}}$ is generated as follows: we draw $j \sim \mathcal{U}([k])$ and $\mathbf{x} \sim \mathsf{N}(\boldsymbol{\mu}^{(j)}, \sigma^2 I_d)$, and set $y = y^{(j)}$. We assume that there exist $i, j \in [k]$ with $y^{(i)} \neq y^{(j)}$. Moreover, we assume the following:

Assumption 2.2. *We have:*

- $\left\|\boldsymbol{\mu}^{(j)}\right\| = \sqrt{d}$ *for all* $j \in [k]$.
- $0 < \sigma \leq 1$.
- $k \left(\max_{i \neq j} |\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| + 4\sigma\sqrt{d}\ln(d) + 1 \right) \leq \frac{d - 4\sigma\sqrt{d}\ln(d) + 1}{10}$.

Example 1. *Below we provide simple examples of settings that satisfy the assumption:*

- *Suppose that the cluster means satisfy $|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| = \tilde{\mathcal{O}}(\sqrt{d})$ for every $i \neq j$. This condition holds, e.g., if we choose each cluster mean i.i.d. from the uniform distribution on the sphere $\sqrt{d} \cdot \mathbb{S}^{d-1}$ (see, e.g., Vardi, Yehudai, and Shamir [VYS22, Lemma 3.1]). Let $\sigma = 1$, namely, each cluster has a radius of roughly $\sqrt{d}$. Then, the assumption can be satisfied by choosing $k = \tilde{\mathcal{O}}(\sqrt{d})$.*
- *Suppose that the cluster means are exactly orthogonal (i.e., $\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle = 0$ for all $i \neq j$), and $\sigma = 1/\sqrt{d}$. Then, the assumption can be satisfied by choosing $k = \tilde{\mathcal{O}}(d)$.*
- *If the number of clusters is $k = \tilde{\mathcal{O}}(1)$, then the assumption may hold even where $\max_{i \neq j} |\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| = \tilde{\Theta}(d)$ (for any $0 < \sigma \leq 1$).*

A few remarks are in order. First, the assumption that $\left\|\boldsymbol{\mu}^{(j)}\right\|$ is exactly $\sqrt{d}$ is for convenience, and we note that it may be relaxed (to have all cluster means approximately of the same norm) without affecting our results significantly. Note that in the case where $\sigma = 1$, the radius of each cluster is roughly of the same magnitude as the cluster mean. Second, we assume for convenience that the noise (i.e., the deviation from the cluster's mean) is drawn from a Gaussian distribution with covariance matrix $\sigma^2 I_d$. However, we note that this assumption can be generalized to any distribution $\mathcal{D}_{\text{noise}}$ such that for every unit vector $\mathbf{e}$ the noise $\boldsymbol{\xi} \sim \mathcal{D}_{\text{noise}}$ satisfies w.h.p. that $\langle \boldsymbol{\xi}, \mathbf{e} \rangle = \tilde{\mathcal{O}}(1)$ and $\|\boldsymbol{\xi}\| = \tilde{\mathcal{O}}(\sqrt{d})$. This property holds, e.g., for a d-dimensional Gaussian distribution $\mathsf{N}(0, \Sigma)$, where $\text{tr}[\Sigma] = d$ and $\|\Sigma\|_2 = O(1)$ (see Frei et al. [Fre+23b, Lemma 3.3]), and more generally for a class of sub-Gaussian distributions (see Hu et al. [Hu+20, Claim 3.1]). Third, note that the third part of Assumption 2.2 essentially requires that the number of clusters k cannot be too large and the correlations between cluster means cannot be too large. Finally, we remark that when k is small, our results may be extended to the case where $\sigma > 1$. For example, if $k = \tilde{\mathcal{O}}(1)$ and $\max_{i \neq j} |\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| = \tilde{\mathcal{O}}(\sqrt{d})$, our generalization result (Theorem 3.1) can be extended to the case where $\sigma = \tilde{\mathcal{O}}(d^{1/8})$. We preferred to avoid handling $\sigma > 1$ in order to simplify the proofs.

Moreover, it is worth noting that Assumption 2.2 implies that the data is w.h.p. linearly separable (see Lemma 2.1 below, and a proof in Appendix B). However, in this work we consider learning using overparameterized ReLU networks, and it is not obvious a priori that gradient methods do not harmfully overfit in this case. Indeed, it has been shown that ReLU networks trained by gradient descent can interpolate training data and fail to generalize well in some distributional settings [Kou+23].

Lemma 2.1. *Let* $\mathbf{u} = \sum_{q \in [k]} y^{(q)} \boldsymbol{\mu}^{(q)}$. *Then, with probability at least* $1 - 2d^{1-\ln(d)/2} = 1 - o_d(1)$ *over* $(\mathbf{x}, y) \sim \mathcal{D}_{\text{clusters}}$, *we have* $y = \operatorname{sign}(\mathbf{u}^\top \mathbf{x})$.

3 Generalization

In this section, we show that under our assumptions on the distribution $\mathcal{D}_{\text{clusters}}$, gradient flow does not harmfully overfit. Namely, even if the learned network is highly overparameterized, the implicit bias of gradient flow guarantees convergence to a solution that generalizes well. Moreover, we show that the sample complexity is optimal. The main result of this section is stated in the following theorem:

Theorem 3.1. *Let* $\epsilon, \delta \in (0,1)$. *Let* $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{-1, 1\}$ *be a training set drawn i.i.d. from the distribution* $\mathcal{D}_{\text{clusters}}$, *where* $n \geq k \ln^2(d)$. *Let* $\mathcal{N}_{\boldsymbol{\theta}}$ *be a depth-2 ReLU network such that* $\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{b}, \mathbf{v}]$ *is a KKT point of Problem (2). Provided* d *is sufficiently large such that* $\delta^{-1} \leq \frac{1}{3} d^{\ln(d)-1}$ *and* $n \leq \min\left\{\sqrt{\frac{\delta}{3}} \cdot e^{d/32}, \frac{\sqrt{\delta}}{3} \cdot d^{\ln(d)/4}, \frac{\epsilon}{4} \cdot d^{\ln(d)/2}\right\}$, *then with probability at least* $1 - \delta$ *over* $\mathcal{S}$, *we have*

$$\Pr_{(\mathbf{x},y) \sim \mathcal{D}_{\text{clusters}}} \left[y \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) \leq 0 \right] \leq \epsilon .$$

The sample complexity requirement in Theorem 3.1 is $n = \tilde{\Omega}(k)$. Essentially, it requires that the dataset $\mathcal{S}$ will include at least one example from each cluster. Clearly, any learning algorithm cannot perform well on unseen clusters. Hence the sample complexity requirement in the theorem is tight (up to log factors).

The assumptions in Theorem 3.1 include upper bounds on δ^{-1} and n. Note that the expressions in these upper bounds are super-polynomial in d, and in particular if $n, \delta^{-1}, \epsilon^{-1} = \operatorname{poly}(d)$, then these assumptions hold for a sufficiently large d. Admittedly, enforcing an upper bound on the training dataset's size is uncommon in generalization results. However, if n is exponential in d, it is not hard to see that there will be clusters which have both positive and negative examples within radius σ of the cluster center, essentially introducing a form of label noise to the problem. Since KKT points of Problem (2) interpolate the training data, this would imply that the network has interpolated training data with label noise—in other words, it has 'overfit' to noise. Understanding the generalization behavior of interpolating neural networks in the presence of label noise is a very technically challenging problem for which much is unknown, especially if one seeks to understand this by only relying upon the properties of KKT conditions for margin maximization. It is noteworthy that all existing non-vacuous generalization bounds for interpolating nonlinear neural networks in the presence of label noise require $n < d$ [FCB22; Cao+22; XG23; Fre+23a; Kou+23].

Combining Theorem 3.1 with Theorem 2.1, we conclude that w.h.p. over a training dataset of size $n \geq k \ln^2(d)$ (and under some additional mild requirements), if gradient flow reaches empirical loss smaller than $\frac{1}{n}$, then it converges in direction to a neural network that generalizes well. This result is width-independent, thus, it holds irrespective of the network width. Specifically, even if the network is highly overparameterized, the implicit bias of gradient flow prevents harmful overfitting. Moreover, the result does not depend directly on the initialization of gradient flow. That is, it holds whenever gradient flow reaches small empirical loss after some finite time. Thus, by relying on the KKT conditions of the max-margin problem instead of analyzing the full gradient flow trajectory, we can prove generalization without the need to prove convergence.

3.1 Proof idea

The proof of Theorem 3.1 is given in Appendix A. Here we discuss the high-level approach. Let $\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{b}, \mathbf{v}]$ be a KKT point of Problem (2). Thus, we have $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j \in [m]} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j)$. Since $\boldsymbol{\theta}$ satisfies the KKT conditions of Problem (2), then there are $\lambda_1, \ldots, \lambda_n$ such that for every $j \in [m]$ we have

$$\mathbf{w}_j = \textstyle\sum_{i \in [n]} \lambda_i \nabla_{\mathbf{w}_j} \left(y_i \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) = \sum_{i \in [n]} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i \;, \tag{3}$$

where $\phi'_{i,j}$ is a subgradient of ϕ at $\mathbf{w}_j^\top \mathbf{x}_i + b_j$, i.e., if $\mathbf{w}_j^\top \mathbf{x}_i + b_j \neq 0$ then $\phi'_{i,j} = \mathbb{1}[\mathbf{w}_j^\top \mathbf{x}_i + b_j \geq 0]$, and otherwise $\phi'_{i,j}$ is some value in $[0, 1]$. Also we have $\lambda_i \geq 0$ for all i, and $\lambda_i = 0$ if $y_i \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq 1$. Likewise, we have

$$b_j = \textstyle\sum_{i \in [n]} \lambda_i \nabla_{b_j} \left(y_i \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) = \sum_{i \in [n]} \lambda_i y_i v_j \phi'_{i,j} \;. \tag{4}$$

In the proof, using a careful analysis of Eq. (3) and (4) we show that w.h.p. $\mathcal{N}_{\boldsymbol{\theta}}$ classifies correctly a fresh example. More precisely, the main argument can be described as follows. We denote $J := [m]$, $J_+ := \{j \in J : v_j > 0\}$, and $J_- := \{j \in J : v_j < 0\}$. Moreover, we denote $I := [n]$ and $Q := [k]$. For $q \in Q$ we denote $I^{(q)} = \{i \in I : \mathbf{x}_i \text{ is in cluster } q\}$. Consider the network's output for an input $\mathbf{x}$ from cluster $r \in Q$ with $y^{(r)} = 1$. Since $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j \in J_+} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) + \sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j)$ and $\phi(z) \geq z$, we have

$$\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) \geq \textstyle\sum_{j \in J_+} v_j (\mathbf{w}_j^\top \mathbf{x} + b_j) + \sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j). \tag{5}$$

This suggests the following possibility: if we can ensure that $\sum_{j \in J_+} v_j (\mathbf{w}_j^\top \mathbf{x} + b_j)$ is large and positive while $\sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j)$ is not too negative, then the network will accurately classify the example $\mathbf{x}$. Using Eq. (3) and (4) and that $y^{(r)} = 1$ (so $y_i = 1$ for $i \in I^{(r)}$), the first term in the above decomposition is equal to

$$\begin{aligned}
&\textstyle\sum_{j \in J_+} v_j (\mathbf{w}_j^\top \mathbf{x} + b_j) = \sum_{j \in J_+} v_j \left[\sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right] \\
&\quad = \textstyle\sum_{j \in J_+} \left[\left(\sum_{i \in I^{(r)}} \lambda_i v_j^2 \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right) + \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i y_i v_j^2 \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right] \\
&\quad \geq \left(\textstyle\sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right) - \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} |\mathbf{x}_i^\top \mathbf{x} + 1| \;.
\end{aligned}$$

Since $\mathbf{x}$ comes from cluster r and the clusters are nearly orthogonal, the pairwise correlations $\mathbf{x}_i^\top \mathbf{x}$ will be large and positive when $i \in I^{(r)}$ but will be small in magnitude when $i \in I^{(q)}$ for $q \neq r$. Thus, we can hope that this term will be large and positive if we can show that the quantity $\sum_{i \in I^{(r)}} \sum_{j \in J+} \lambda_i v_j^2 \phi'_{i,j}$ is not too small relative to the quantity $\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j}$. By similar arguments, in order to show the second term in Eq. (5) is not too negative, we need to understand how the quantity $\sum_{i \in I^{(q)}} \sum_{j \in J_-} \lambda_i v_j^2 \phi'_{i,j}$ varies across different clusters $q \in Q$. Hence, in the proof we analyze how the quantities $\sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j}$, $\sum_{i \in I^{(q)}} \sum_{j \in J_-} \lambda_i v_j^2 \phi'_{i,j}$ relate to each other for different clusters $q \in Q$, and show that these quantities are all of the same order. Then, we conclude that w.h.p. $\mathbf{x}$ is classified correctly.

4 Robustness

We begin by introducing the definition of $R(\cdot)$-robustness.

Definition 4.1. *Given some function $R(\cdot)$, we say that a neural network $\mathcal{N}_{\boldsymbol{\theta}}$ is* $R(d)$-robust *w.r.t. a distribution $\mathcal{D}_{\mathbf{x}}$ over $\mathbb{R}^d$ if for every $r = o(R(d))$, with probability $1 - o_d(1)$ over $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$, for every $\mathbf{x}' \in \mathbb{R}^d$ with $\|\mathbf{x} - \mathbf{x}'\| \leq r$ we have $\mathrm{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}')) = \mathrm{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}))$.*

Thus, a neural net $\mathcal{N}_{\boldsymbol{\theta}}$ is $R(d)$-robust if changing the label of an example cannot be done with a perturbation of size $o(R(d))$. Note that we consider here ℓ_2 perturbations.

For the distribution $\mathcal{D}_{\text{clusters}}$ under consideration, it is straightforward to show that classifiers cannot be $R(d)$-robust if $R(d) = \omega(\sqrt{d})$: since the distance between examples in different clusters is w.h.p.

$\mathcal{O}(\sqrt{d})$, it is clearly possible to flip the sign of an example with a perturbation of size $\mathcal{O}(\sqrt{d})$. In particular, the best we can hope for is $\sqrt{d}$-robustness. In the following theorem, we show that there exist two-layer ReLU networks which can both achieve small test error and the optimal level of $\sqrt{d}$-robustness.

Theorem 4.1. *For every $r \geq k$, there exists a depth-2 ReLU network $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}$ of width r such that for $(\mathbf{x}, y) \sim \mathcal{D}_{\text{clusters}}$, with probability at least $1 - d^{-\omega_d(1)}$ we have $y\mathcal{N}(\mathbf{x}) \geq 1$, and flipping the sign of the output requires a perturbation of size larger than $\frac{\sqrt{d}}{8}$ (for a sufficiently large d). Thus, $\mathcal{N}$ classifies the data correctly w.h.p., and it is $\sqrt{d}$-robust w.r.t. $\mathcal{D}_{\mathbf{x}}$.*

Thus, we see that $\sqrt{d}$-robust networks exist. In the following theorem, we show that the implicit bias of gradient flow constrains the level of robustness of *trained* networks whenever the number of clusters k is large.

Theorem 4.2. *Let $\epsilon, \delta \in (0, 1)$. Let $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{-1, 1\}$ be a training set drawn i.i.d. from the distribution $\mathcal{D}_{\text{clusters}}$, where $n \geq k \ln^2(d)$. We denote $Q_+ = \{q \in [k] : y^{(q)} = 1\}$ and $Q_- = \{q \in [k] : y^{(q)} = -1\}$, and assume that $\min\left\{\frac{|Q_+|}{k}, \frac{|Q_-|}{k}\right\} \geq c$ for some $c > 0$. Let $\mathcal{N}_{\boldsymbol{\theta}}$ be a depth-2 ReLU network such that $\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{b}, \mathbf{v}]$ is a KKT point of Problem (2). Provided d is sufficiently large such that $\delta^{-1} \leq \frac{1}{3} d^{\ln(d)-1}$ and $n \leq \min\left\{\sqrt{\frac{\delta}{3}} \cdot e^{d/32}, \frac{\sqrt{\delta}}{3} \cdot d^{\ln(d)/4}, \frac{\epsilon}{4} \cdot d^{\ln(d)/2}\right\}$, with probability at least $1 - \delta$ over $\mathcal{S}$, there is a vector $\mathbf{z} = \eta \cdot \sum_{j \in [k]} y^{(j)} \boldsymbol{\mu}^{(j)}$ with $\eta > 0$ and $\|\mathbf{z}\| \leq \mathcal{O}\left(\sqrt{d/c^2 k}\right)$, such that*

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clusters}}}\left[\text{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x})) \neq \text{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x} - y\mathbf{z}))\right] \geq 1 - \epsilon\ .$$

Note that the expressions in the upper bounds on n and δ^{-1} are super-polynomial in d, and hence these requirements are mild (e.g., they hold for a sufficiently large d when $n, \delta^{-1}, \epsilon^{-1} = \text{poly}(d)$). As we mentioned in the discussion following Theorem 3.1, we believe removing the requirement for an upper bound on n would be highly nontrivial.

Theorem 4.2 implies that if $c^2 k = \omega_d(1)$, then w.h.p. over the training dataset, every KKT point of Problem (2) is not $\sqrt{d}$-robust. Specifically, if c is constant, namely, at least a constant fraction of the clusters have positive labels and a constant fraction of the clusters have negative labels, then the network is not $\sqrt{d}$-robust if $k = \omega_d(1)$. Recall that by Theorem 3.1, we also have w.h.p. that every KKT point generalizes well. Overall, combining Theorems 2.1, 3.1, 4.1, and 4.2, we conclude that for $c^2 k = \omega_d(1)$, w.h.p. over a training dataset of size $n \geq k \ln^2(d)$, if gradient flow reaches empirical loss smaller than $\frac{1}{n}$, then it converges in direction to a neural network that generalizes well but is not $\sqrt{d}$-robust, even though there exist $\sqrt{d}$-robust networks that generalize well. Thus, in our setting, there is bias towards solutions that generalize well but are non-robust.

Example 2. *Consider the setting from the first item of Example 1. Thus, the cluster means satisfy $|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| = \tilde{\mathcal{O}}(\sqrt{d})$ for every $i \neq j$, and we have $\sigma = 1$ and $k = \tilde{\Theta}(\sqrt{d})$. Suppose that $c = \Theta(1)$, namely, there is at least a constant fraction of clusters with each label $\{-1, 1\}$. Then, the adversarial perturbation $\mathbf{z}$ from Theorem 4.2 satisfies $\|\mathbf{z}\| = \mathcal{O}\left(\sqrt{d/k}\right) = \tilde{\mathcal{O}}\left(d^{1/4}\right) = o(\sqrt{d})$.*

Similarly to our discussion after Theorem 3.1, we note that Theorem 4.2 is width-independent, i.e., it holds irrespective of the network width. It implies that we cannot hope to obtain a robust solution by choosing an appropriate width for the trained network. As we discussed in the related work section, Bubeck, Li, and Nagaraj [BLN21] and Bubeck and Sellke [BS21] considered the expressive power of neural networks, and showed that overparameterization might be necessary for robustness. By Theorem 4.2, even when the network is overparameterized, the implicit bias of the optimization method can prevent convergence to robust solutions. Moreover, our result does not depend directly on the initialization of gradient flow. Recall that by Theorem 2.1 if gradient flow reaches small empirical loss then it converges in direction to a KKT point of Problem (2). Hence our result holds whenever gradient flow reaches a small empirical loss.

Note that in Theorem 4.2, the adversarial perturbation does not depend on the input (up to sign). It corresponds to the well-known empirical phenomenon of *universal adversarial perturbations*, where one can find a single perturbation that simultaneously flips the label of many inputs (cf. [Moo+17;

Zha+21]). Moreover, the same perturbation applies to all depth-2 networks to which gradient flow might converge (i.e., all KKT points). It corresponds to the well-known empirical phenomenon of *transferability* in adversarial examples, where one can find perturbations that simultaneously flip the labels of many different trained networks (cf. [Liu+17; AM18]).

It is worth noting that Theorems 3.1 and 4.2 demonstrate that trained neural networks exhibit different properties than the 1-nearest-neighbour learning rule, irrespective of the number of parameters in the network. For example, consider the case where $\sigma = \frac{1}{\sqrt{d}}$, namely, the examples of each cluster are concentrated within a ball of radius $O(1)$ around its mean. Then, the distance between every pair of points from the same cluster is $O(1)$, and the distance between points from different clusters is $\Omega(\sqrt{d})$. In this setting, both the 1-nearest-neighbour classifier and the trained neural network will classify a fresh example correctly w.h.p., but in the 1-nearest-neighbour classifier flipping the output's sign will require a perturbation of size $\Omega(\sqrt{d})$, while in the neural network a much smaller perturbation will suffice.

Finally, we remark that in the limit $\sigma \to 0$, we get a distribution supported on $\boldsymbol{\mu}^{(1)}, \ldots, \boldsymbol{\mu}^{(k)}$. Then, a training dataset of size $n \geq k \ln^2(d)$ will contain w.h.p. all examples in the support, and hence robustness w.r.t. test data is equivalent to robustness w.r.t. the training data. In this case, we recover the results of Vardi, Yehudai, and Shamir [VYS22] which characterized the non-robustness of KKT points of ReLU networks trained on nearly orthogonal training data. In particular, our Theorem 4.2 is a strict generalization of their Theorem 4.1.

4.1 Proof ideas

Here we discuss the main ideas in the proofs of Theorem 4.1 and 4.2 (see Appendices C and D for the formal proofs).

The proof of Theorem 4.1 follows by the following simple construction. The robust network includes k neurons, each corresponding to a single cluster. That is, we have $\mathcal{N}(\mathbf{x}) = \sum_{j=1}^{k} v_j \sigma(\mathbf{w}^\top \mathbf{x} + b_j)$, where $v_j = y^{(j)}$, $\mathbf{w}_j = \frac{4\mu^{(j)}}{d}$, and $b_j = -2$. Note that the j-th neuron points at the direction of the j-th cluster and has a negative bias term, such that the neuron is active on points from the j-th cluster, and inactive on points from the other clusters. Then, given a fresh example $(\mathbf{x}, y) \sim \mathcal{D}_{\text{clusters}}$, we show that the network classifies it correctly w.h.p. with margin at least 1. Also, there is w.h.p exactly one neuron that is active on $\mathbf{x}$, and hence the gradient of the network w.r.t. the input is affected only by this neuron and is of size $\mathcal{O}(1/\sqrt{d})$. Therefore, we need a perturbation of size $\Omega(\sqrt{d})$ in order to flip the output's sign.

The intuition for Theorem 4.2 can be described as follows. Recall that in our construction of a robust network above, an example $(\mathbf{x}, y) \sim \mathcal{D}_{\text{clusters}}$ is w.h.p. in an active region of exactly one neuron, and hence in the neighborhood of $\mathbf{x}$ the output of the network is sensitive only to perturbations in the direction of that neuron. Now, consider the linear model $\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$, where $\mathbf{w} = \sum_{q=1}^{k} \frac{1}{d} y^{(q)} \boldsymbol{\mu}^{(q)}$. It is not hard to verify that for $(\mathbf{x}, y) \sim \mathcal{D}_{\text{clusters}}$ we have w.h.p. that $0 < y\mathbf{w}^\top \mathbf{x} \leq \mathcal{O}(1)$. Moreover, the gradient of this linear predictor is of size $\|\mathbf{w}\| = \Omega(\sqrt{k/d})$. Hence, we can flip the output's sign with a perturbation of size $\mathcal{O}(\sqrt{d/k})$. Thus, the linear classifier is non-robust if $k = \omega_d(1)$. Intuitively, the difference between our robust ReLU network and the non-robust linear classifier is the fact that in the neighborhood of $\mathbf{x}$ the robust network is sensitive only to perturbations in the direction of one cluster, while the linear classifier is sensitive to perturbations in the directions of all k clusters. In the proof, we analyze ReLU networks which are KKT points of Problem (2), and show that although these ReLU networks are non-linear, they are still sensitive to perturbations in the directions of all k clusters, similarly to the above linear classifier. The formal proof follows by a careful analysis of the KKT conditions of Problem (2), given in Eq. (3) and (4).

We remark that in the proof of Theorem 4.2 we use some technical ideas from Vardi, Yehudai, and Shamir [VYS22]. However, there are significant differences between the two settings. For example, they assume that the training data are nearly orthogonal, which only holds when the dimension is large relative to the number of samples; thus, it is unclear whether the existence of small adversarial perturbations in their setting is due to the high-dimensionality of the data or if a similar phenomenon exists in the more common $n > d$ setting. At a more technical level, their proof relies on showing that in a KKT point all inputs must lie exactly on the margin, while in our setting they are not guaranteed to lie exactly on the margin.

5 Discussion

In this paper, we considered clustered data, and showed that gradient flow in two-layer ReLU networks does not harmfully overfit, but also hinders robustness. Our results follow by analyzing the KKT points of the max-margin problem in parameter space. In our distributional setting, the clusters are well-separated, and hence there exist robust classifiers, which allows us to consider the effect of the implicit bias of gradient flow on both generalization and robustness. Understanding generalization and robustness in additional data distributions and neural network architectures is a challenging but important question. As a possible next step, it would be interesting to study whether the approach used in this paper can be extended to the following data distributions:

First, our assumption on the data distribution (Assumption 2.2) implies that the number of clusters cannot be too large, and as a result the data is linearly separable (Lemma 2.1). We conjecture that our results hold even for a significantly larger number of clusters, such that the data is not linearly separable.

Second, it would be interesting to understand whether our generalization result holds for linearly separable data distributions that are not clustered. That is, given a distribution that is linearly separable with some margin $\gamma > 0$ and a training dataset that is large enough to allow learning with a max-margin linear classifier, are there KKT points of the max-margin problem for two-layer ReLU networks that do not generalize well? In other words, do ReLU networks that satisfy the KKT conditions generalize at least as well as max-margin linear classifiers?

Acknowledgments and Disclosure of Funding

SF, GV, PB, and NS acknowledge the support of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639, and of the NSF through grant DMS-2023505.

References

[AM18] Naveed Akhtar and Ajmal Mian. “Threat of adversarial attacks on deep learning in computer vision: A survey”. In: *Ieee Access* 6 (2018), pp. 14410–14430 (Cited on pages 2, 9).

[AL21] Zeyuan Allen-Zhu and Yuanzhi Li. “Feature purification: How adversarial training performs robust deep learning”. In: *IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. 2021 (Cited on page 3).

[ACW18] Anish Athalye, Nicholas Carlini, and David Wagner. “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples”. In: *International conference on machine learning (ICML)*. 2018 (Cited on page 1).

[BBC21] Peter Bartlett, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. “Adversarial examples in multi-layer random relu networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021) (Cited on page 3).

[Bub+21] Sébastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, and Rémi Tachet des Combes. “A single gradient step finds adversarial examples on random two-layers neural networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021) (Cited on page 3).

[Bub+19] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. “Adversarial examples from computational constraints”. In: *International Conference on Machine Learning (ICML)*. 2019 (Cited on page 3).

[BLN21] Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. “A law of robustness for two-layers neural networks”. In: *Conference on Learning Theory (COLT)*. 2021 (Cited on pages 3, 8).

[BS21] Sébastien Bubeck and Mark Sellke. "A universal law of robustness via isoperimetry". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021) (Cited on pages 3, 8).

[Cao+22] Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. "Benign Overfitting in Two-layer Convolutional Neural Networks". In: *Preprint, arXiv:2202.06526* (2022) (Cited on page 6).

[CW17] Nicholas Carlini and David Wagner. "Adversarial examples are not easily detected: Bypassing ten detection methods". In: *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 2017 (Cited on page 1).

[CW18] Nicholas Carlini and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text". In: *IEEE Security and Privacy Workshops (SPW)*. 2018 (Cited on page 1).

[CL21] Niladri S. Chatterji and Philip M. Long. "Finite-sample analysis of interpolating linear classifiers in the overparameterized regime". In: *Journal of Machine Learning Research* 22.129 (2021), pp. 1–30 (Cited on page 5).

[CB20] Lenaic Chizat and Francis Bach. "Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss". In: *Conference on Learning Theory (COLT)*. 2020 (Cited on page 3).

[Cla+08] Francis H Clarke, Yuri S Ledyaev, Ronald J Stern, and Peter R Wolenski. *Nonsmooth analysis and control theory*. Vol. 178. Springer Science & Business Media, 2008 (Cited on page 4).

[CH20] Francesco Croce and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks". In: *International conference on machine learning (ICML)*. 2020 (Cited on page 1).

[DS20] Amit Daniely and Hadas Shacham. "Most ReLU Networks Suffer from ℓ_2 Adversarial Perturbations". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2020) (Cited on page 3).

[DB22] Elvis Dohmatob and Alberto Bietti. "On the (non-) robustness of two-layer neural networks in different learning regimes". In: *arXiv preprint arXiv:2203.11864* (2022) (Cited on page 3).

[Dut+13] Joydeep Dutta, Kalyanmoy Deb, Rupesh Tulshyan, and Ramnik Arora. "Approximate KKT points and a proximity measure for termination". In: *Journal of Global Optimization* 56.4 (2013), pp. 1463–1499 (Cited on page 4).

[FFF18] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. "Adversarial vulnerability for any classifier". In: *Advances in neural information processing systems (NeurIPS)* (2018) (Cited on page 3).

[FCB22] Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. "Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data". In: *Conference on Learning Theory (COLT)*. 2022 (Cited on page 6).

[Fre+23a] Spencer Frei, Gal Vardi, Peter L. Bartlett, and Nathan Srebro. "Benign Overfitting in Linear Classifiers and Leaky ReLU Networks from KKT Conditions for Margin Maximization". In: *Preprint, arXiv:2303.01462* (2023) (Cited on pages 5, 6).

[Fre+23b] Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. "Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data". In: *International Conference on Learning Representations (ICLR)*. 2023 (Cited on pages 3, 5).

[GSS15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *International Conference on Learning Representations (ICLR)*. 2015 (Cited on pages 1, 3).

[Hai+22] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. "Reconstructing training data from trained neural networks". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022) (Cited on page 3).

[Hu+20] Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. "The surprising simplicity of the early-time learning dynamics of neural networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020 (Cited on page 5).

[JT20] Ziwei Ji and Matus Telgarsky. "Directional convergence and alignment in deep learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020 (Cited on pages 2–4).

[KH18] Marc Khoury and Dylan Hadfield-Menell. "On the geometry of adversarial examples". In: *Preprint, arXiv:1811.00525* (2018) (Cited on page 3).

[KYS23] Guy Kornowski, Gilad Yehudai, and Ohad Shamir. "From Tempered to Benign Overfitting in ReLU Neural Networks". In: *arXiv preprint arXiv:2305.15141* (2023) (Cited on page 3).

[Kou+23] Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. "Benign Overfitting for Two-layer ReLU Networks". In: *International Conference on Machine Learning (ICML)*. 2023 (Cited on page 6).

[Kun+22] Daniel Kunin, Atsushi Yamamura, Chao Ma, and Surya Ganguli. "The Asymmetric Maximum Margin Bias of Quasi-Homogeneous Neural Networks". In: *arXiv preprint arXiv:2210.03820* (2022) (Cited on page 3).

[KGB17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale". In: *International Conference on Learning Representations (ICLR)*. 2017 (Cited on page 1).

[LM00] Beatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection". In: *Annals of Statistics* (2000), pp. 1302–1338 (Cited on page 15).

[Liu+17] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. "Delving into transferable adversarial examples and black-box attacks". In: *International Conference on Learning Representations (ICLR)*. 2017 (Cited on pages 2, 9).

[LL20] Kaifeng Lyu and Jian Li. "Gradient descent maximizes the margin of homogeneous neural networks". In: *International Conference on Learning Representations (ICLR)*. 2020 (Cited on pages 2–4).

[Lyu+21] Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. "Gradient descent on two-layer nets: Margin maximization and simplicity bias". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021 (Cited on page 3).

[Mad+18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *International Conference on Learning Representations (ICLR)*. 2018 (Cited on page 1).

[Mon+19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. "The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime". In: *Preprint, arXiv:1911.01544* (2019) (Cited on page 5).

[MW22] Andrea Montanari and Yuchen Wu. "Adversarial Examples in Random Neural Networks with General Activations". In: *arXiv preprint arXiv:2203.17209* (2022) (Cited on page 3).

[Moo+17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. "Universal adversarial perturbations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2017 (Cited on pages 2, 8).

[Ney+17] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. "Exploring generalization in deep learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 (Cited on page 1).

[Pap+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. "Practical black-box attacks against machine learning". In: *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017 (Cited on page 1).

[Pap+16] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. "Distillation as a defense to adversarial perturbations against deep neural networks". In: *IEEE symposium on security and privacy (SP)*. 2016 (Cited on page 1).

[PL20] Mary Phuong and Christoph H Lampert. "The inductive bias of ReLU networks on orthogonally separable data". In: *International Conference on Learning Representations (ICLR)*. 2020 (Cited on page 3).

[SVL22] Itay Safran, Gal Vardi, and Jason D Lee. "On the Effective Number of Linear Regions in Shallow Univariate ReLU Networks: Convergence Guarantees and Implicit Bias". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022) (Cited on page 3).

[SBG21] Roei Sarussi, Alon Brutzkus, and Amir Globerson. "Towards understanding learning in neural networks with linear teachers". In: *International Conference on Machine Learning (ICML)*. 2021 (Cited on page 3).

[Sch+18] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. "Adversarially robust generalization requires more data". In: *Advances in neural information processing systems (NeurIPS)* (2018) (Cited on page 3).

[Sha+19] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. "Are adversarial examples inevitable?" In: *International Conference on Learning Representations (ICLR)*. 2019 (Cited on page 3).

[Sha+20] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. "The pitfalls of simplicity bias in neural networks". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2020) (Cited on page 3).

[SMB21] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. "The dimpled manifold model of adversarial examples in machine learning". In: *Preprint, arXiv:2106.10151* (2021) (Cited on page 3).

[Sin+21] Vasu Singla, Songwei Ge, Basri Ronen, and David Jacobs. "Shift invariance can reduce adversarial robustness". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021) (Cited on page 3).

[Sze+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks". In: *International Conference on Learning Representations (ICLR)*. 2014 (Cited on page 1).

[Var22] Gal Vardi. "On the Implicit Bias in Deep-Learning Algorithms". In: *Preprint, arXiv:2208.12591* (2022) (Cited on page 2).

[VSS22] Gal Vardi, Ohad Shamir, and Nathan Srebro. "On Margin Maximization in Linear and ReLU Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022 (Cited on pages 3, 5).

[VYS22] Gal Vardi, Gilad Yehudai, and Ohad Shamir. "Gradient Methods Provably Converge to Non-Robust Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022 (Cited on pages 3, 5, 9).

[Wan+20] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. "High-frequency component helps explain the generalization of convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 (Cited on page 3).

[Wan+22] Yunjuan Wang, Enayat Ullah, Poorya Mianjy, and Raman Arora. "Adversarial Robustness is at Odds with Lazy Training". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022) (Cited on page 3).

[WK18] Eric Wong and Zico Kolter. "Provable defenses against adversarial examples via the convex outer adversarial polytope". In: *International Conference on Machine Learning (ICML)*. 2018 (Cited on page 1).

[WRK20] Eric Wong, Leslie Rice, and J Zico Kolter. "Fast is better than free: Revisiting adversarial training". In: *International Conference on Learning Representations (ICLR)*. 2020 (Cited on page 1).

[Wu+20] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. "Making an invisibility cloak: Real world adversarial attacks on object detectors". In: *European Conference on Computer Vision (ECCV)*. 2020 (Cited on page 1).

[XG23] Xingyu Xu and Yuantao Gu. "Benign overfitting of non-smooth neural networks beyond lazy training". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. 2023 (Cited on page 6).

[Zha+21] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. "A survey on universal adversarial attack". In: *International Joint Conference on Artificial Intelligence (IJCAI)* (2021) (Cited on pages 2, 9).

[Zha+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Delving into transferable adversarial examples and black-box attacks". In: *International Conference on Learning Representations (ICLR)*. 2017 (Cited on page 1).

A Proof of Theorem 3.1

We will prove the following theorem:

Theorem A.1. *Let $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{-1, 1\}$ be a training set drawn i.i.d. from the distribution $\mathcal{D}_{clusters}$, where $n \geq k \ln^2(d)$. Let $\mathcal{N}_{\boldsymbol{\theta}}$ be a depth-2 ReLU network such that $\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{b}, \mathbf{v}]$ is a KKT point of Problem (2). Then, with probability at least*

$$1 - \left(3n^2 d^{-\frac{\ln(d)}{2}} + n^2 e^{-d/16} + d^{1-\ln(d)}\right)$$

over $\mathcal{S}$, we have

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{clusters}} [y\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x})) \leq 0] \leq 4nd^{-\frac{\ln(d)}{2}} .$$

It is easy to verify that Theorem A.1 implies Theorem 3.1. Indeed, if $\frac{1}{\delta} \leq \frac{1}{3} d^{\ln(d)-1}$ and

$$n \leq \min\left\{\sqrt{\frac{\delta}{3}} \cdot e^{d/32}, \frac{\sqrt{\delta}}{3} \cdot d^{\ln(d)/4}, \frac{\epsilon}{4} \cdot d^{\ln(d)/2}\right\} ,$$

then we have:

1.
$$3n^2 d^{-\frac{\ln(d)}{2}} \leq 3\left(\frac{\sqrt{\delta}}{3} d^{\ln(d)/4}\right)^2 d^{-\ln(d)/2} = 3 \cdot \frac{\delta}{9} \cdot d^{\ln(d)/2} d^{-\ln(d)/2} = \frac{\delta}{3} .$$

2.
$$n^2 e^{-d/16} \leq \left(\sqrt{\frac{\delta}{3}} \cdot e^{d/32}\right)^2 e^{-d/16} = \frac{\delta}{3} .$$

3.
$$d^{1-\ln(d)} \leq \frac{\delta}{3} .$$

4.
$$4nd^{-\frac{\ln(d)}{2}} \leq 4 \cdot \frac{\epsilon}{4} \cdot d^{\ln(d)/2} \cdot d^{-\ln(d)/2} = \epsilon .$$

Hence, under the above assumptions on δ and n, Theorem A.1 implies that with probability at least $1 - \delta$ over $\mathcal{S}$, we have $\Pr_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clusters}}} [y\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x})) \leq 0] \leq \epsilon$.

We now turn to prove Theorem A.1. The high-level idea for the proof is as follows. First, we show that the training dataset is sufficiently "nice" with high probability, in the sense that samples within each cluster are highly correlated while samples in orthogonal clusters are nearly orthogonal (see properties (P1) through (P6) below). This analysis appears in Section A.1. We then show that datasets with "nice" properties impose a number of structural constraints on the properties of KKT points of the margin maximization problem for ReLU nets; this appears in Section A.2. We conclude in Section A.3 by showing how these structural conditions allow for generalization on fresh test data.

A.1 Training dataset properties

We denote $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j\in[m]} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j)$. Thus, $\mathcal{N}_{\boldsymbol{\theta}}$ is a network of width m, where the weights in the first layer are $\mathbf{w}_1, \ldots, \mathbf{w}_m$, the bias terms are $b_1, \ldots, b_m$, and the weights in the second layer are $v_1, \ldots, v_m$. We denote $J := [m]$, $J_+ := \{j \in J : v_j > 0\}$, and $J_- := \{j \in J : v_j < 0\}$. Moreover, we denote $I := [n]$, $I_+ := \{i \in I : y_i = 1\}$, and $I_- := \{i \in I : y_i = -1\}$. Finally, we denote $Q := [k]$, $Q_+ := \{q \in Q : y^{(q)} = 1\}$, and $Q_- := \{q \in Q : y^{(q)} = -1\}$.

We denote $p := \max_{q\neq q'} |\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\mu}^{(q')}\rangle|$. The distribution $\mathcal{D}_{\text{clusters}}$ is such that each example $(\mathbf{x}_i, y_i)$ in $\mathcal{S}$ is generated as follows: we draw $q_i \sim \mathcal{U}(Q)$ and $\boldsymbol{\xi}_i \sim \mathsf{N}(\mathbf{0}, \sigma^2 I_d)$ and set $\mathbf{x}_i = \boldsymbol{\mu}^{(q_i)} + \boldsymbol{\xi}_i$ and $y_i = y^{(q_i)}$. We denote $\text{cluster}(i) = q_i$. For $q \in Q$ we denote $I^{(q)} = \{i \in I : \text{cluster}(i) = q\}$. We also denote $\Delta = 4\sigma\sqrt{d}\ln(d)$.

Our goal in this section will be to show that with high probability, the dataset $\mathcal{S}$ satisfies the following properties.

(P1) For every $i \in I$ we have $\|\boldsymbol{\xi}_i\| \leq \sigma\sqrt{2d}$.

(P2) For every $i \neq i'$ in I we have $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \leq \sigma^2 \sqrt{2d} \ln(d)$.

(P3) For every $i \in I$ and $q \in Q$ we have $|\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi}_i \rangle| \leq \sigma \sqrt{d} \ln(d)$.

(P4) For every $i, i' \in I$ with $\text{cluster}(i) \neq \text{cluster}(i')$ we have $|\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle| \leq p + \Delta$.

(P5) For every $i, i' \in I$ with $\text{cluster}(i) = \text{cluster}(i')$ we have $d - \Delta \leq \langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle \leq 3d + \Delta$.

(P6) For every $q \in Q$ there exists $i \in I$ with $\text{cluster}(i) = q$ (i.e., $I^{(q)} \neq \emptyset$).

More formally, in the remainder of this section we shall show the following proposition.

Proposition A.1. *With probability at least* $1 - \left(3n^2 d^{-\frac{\ln(d)}{2}} + n^2 e^{-d/16} + d^{1-\ln(d)}\right)$*, the dataset* $\mathcal{S}$ *satisfies the properties (P1) through (P6).*

We start with some auxiliary lemmas. The first bounds the norm of $\boldsymbol{\xi}$.

Lemma A.1. *Let* $\boldsymbol{\xi} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I_d)$*. Then,*

$$\Pr\left[\|\boldsymbol{\xi}\| \geq \sigma\sqrt{2d}\right] \leq e^{-d/16}\ .$$

Proof. Note that $\left\|\frac{\boldsymbol{\xi}}{\sigma}\right\|^2$ has the Chi-squared distribution. A concentration bound by Laurent and Massart [LM00, Lemma 1] implies that for all $t > 0$ we have

$$\Pr\left[\left\|\frac{\boldsymbol{\xi}}{\sigma}\right\|^2 - d \geq 2\sqrt{dt} + 2t\right] \leq e^{-t}\ .$$

Plugging-in $t = \frac{d}{16}$, we get

$$\Pr\left[\left\|\frac{\boldsymbol{\xi}}{\sigma}\right\|^2 \geq 2d\right] \leq \Pr\left[\left\|\frac{\boldsymbol{\xi}}{\sigma}\right\|^2 - d \geq d/2 + d/8\right] \leq e^{-d/16}\ .$$

Thus, we have

$$\Pr\left[\|\boldsymbol{\xi}\| \geq \sigma\sqrt{2d}\right] \leq e^{-d/16}\ .$$

□

Our next lemma bounds the projection of a Gaussian $\boldsymbol{\xi}'$ onto a fixed vector $\boldsymbol{\xi}$.

Lemma A.2. *Let* $\boldsymbol{\xi} \in \mathbb{R}^d$ *and let* $\boldsymbol{\xi}' \sim \mathsf{N}(\mathbf{0}, \sigma^2 I_d)$*. Then,*

$$\Pr\left[|\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle| \geq \|\boldsymbol{\xi}\|\, \sigma \ln(d)\right] \leq 2d^{-\frac{\ln(d)}{2}}\ .$$

Proof. Note that $\langle \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|}, \boldsymbol{\xi}' \rangle$ has the distribution $\mathsf{N}(0, \sigma^2)$. By a standard tail bound, we have for every $t \geq 0$ that $\Pr\left[\left|\langle \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|}, \boldsymbol{\xi}' \rangle\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$. Hence,

$$\Pr\left[\left|\left\langle \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|}, \boldsymbol{\xi}' \right\rangle\right| \geq \sigma \ln(d)\right] \leq 2\exp\left(-\frac{\sigma^2 \ln^2(d)}{2\sigma^2}\right) = 2d^{-\frac{\ln(d)}{2}}\ .$$

The lemma now follows immediately. □

We can utilize the two preceding lemmas to show that the pairwise correlations between independent Gaussians is small relative to the norms of the Gaussians.

Lemma A.3. *Let* $\boldsymbol{\xi}, \boldsymbol{\xi}'$ *drawn i.i.d. from* $\mathsf{N}(\mathbf{0}, \sigma^2 I_d)$*. Then,*

$$\Pr\left[|\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle| \geq \sqrt{2d} \ln(d) \sigma^2\right] \leq e^{-d/16} + 2d^{-\frac{\ln(d)}{2}}\ .$$

Proof. Note that if $|\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle| \geq \sqrt{2d}\ln(d)\sigma^2$ then we have at least one of the following: (1) $\|\boldsymbol{\xi}\| \geq \sigma\sqrt{2d}$; (2) $|\langle \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|}, \boldsymbol{\xi}' \rangle| \geq \sigma \ln(d)$. We will bound the probability of these events.

First, by Lemma A.1 we have

$$\Pr\left[\|\boldsymbol{\xi}\| \geq \sigma\sqrt{2d}\right] \leq e^{-d/16} .$$

Next, by Lemma A.2 we have

$$\Pr\left[\left|\left\langle \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|}, \boldsymbol{\xi}' \right\rangle\right| \geq \sigma \ln(d)\right] \leq 2d^{-\frac{\ln(d)}{2}} .$$

Combining the above displayed equations we conclude that

$$\Pr\left[|\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle| \geq \sqrt{2d}\ln(d)\sigma^2\right] \leq e^{-d/16} + 2d^{-\frac{\ln(d)}{2}} .$$

□

Our next lemma bounds the projection of the noise vectors onto any cluster mean.

Lemma A.4. *Let $i \in [k]$ and let $\boldsymbol{\xi} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I_d)$. Then*

$$\Pr\left[|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi} \rangle| \geq \sigma\sqrt{d}\ln(d)\right] \leq 2d^{-\frac{\ln(d)}{2}} .$$

Proof. Follows immediately from Lemma A.2. □

The following lemmas use bounds on the pairwise interactions between noise vectors and cluster means to bound the correlations between the sums of noises and cluster means.

Lemma A.5. *Let $i \neq j$ be indices in $[k]$. Let $\boldsymbol{\xi}, \boldsymbol{\xi}'$ such that the following hold:*

- $|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi}' \rangle| \leq \sigma\sqrt{d}\ln(d)$.
- $|\langle \boldsymbol{\mu}^{(j)}, \boldsymbol{\xi} \rangle| \leq \sigma\sqrt{d}\ln(d)$.
- $|\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle| \leq \sigma\sqrt{2d}\ln(d)$.

Then,

$$|\langle \boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}, \boldsymbol{\mu}^{(j)} + \boldsymbol{\xi}' \rangle| \leq 4\sigma\sqrt{d}\ln(d) + |\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| .$$

Proof. We have

$$\begin{aligned} |\langle \boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}, \boldsymbol{\mu}^{(j)} + \boldsymbol{\xi}' \rangle| &\leq |\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| + |\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi}' \rangle| + |\langle \boldsymbol{\xi}, \boldsymbol{\mu}^{(j)} \rangle| + |\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle| \\ &\leq |\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| + \sigma\sqrt{d}\ln(d) + \sigma\sqrt{d}\ln(d) + \sigma\sqrt{2d}\ln(d) \\ &\leq |\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(j)} \rangle| + 4\sigma\sqrt{d}\ln(d) . \end{aligned}$$

□

Lemma A.6. *Let $i \in [k]$, and let $\boldsymbol{\xi}, \boldsymbol{\xi}'$ such that the following hold:*

- $|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi} \rangle| \leq \sigma\sqrt{d}\ln(d)$.
- $|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi}' \rangle| \leq \sigma\sqrt{d}\ln(d)$.
- $|\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle| \leq \sigma\sqrt{2d}\ln(d)$.

Then,

$$\left|\langle \boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}, \boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}' \rangle - d\right| \leq 4\sigma\sqrt{d}\ln(d) .$$

Proof. We have

$$\begin{aligned}\left|\langle \boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}, \boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}' \rangle - d\right| &= \left|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi}' \rangle + \langle \boldsymbol{\xi}, \boldsymbol{\mu}^{(i)} \rangle + \langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle\right| \\ &\leq \sigma\sqrt{d}\ln(d) + \sigma\sqrt{d}\ln(d) + \sigma\sqrt{2d}\ln(d) \\ &\leq 4\sigma\sqrt{d}\ln(d)\ .\end{aligned}$$

□

The next lemma uses bounds on the projection of the noise vector onto cluster means and the norms of the cluster means to derive bounds on the norm of the sum $\boldsymbol{\mu}^{(i)} + \xi$.

Lemma A.7. *Let $i \in [k]$, and let $\boldsymbol{\xi}$ such that the following hold:*

- $|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi} \rangle| \leq \sigma\sqrt{d}\ln(d)$.
- $\|\boldsymbol{\xi}\|^2 \leq 2\sigma^2 d$.

Then,

$$d - 2\sigma\sqrt{d}\ln(d) \leq \left\|\boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}\right\|^2 \leq 3d + 2\sigma\sqrt{d}\ln(d)\ .$$

Proof. We have

$$\left\|\boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}\right\|^2 = \left\|\boldsymbol{\mu}^{(i)}\right\|^2 + \|\boldsymbol{\xi}\|^2 + 2\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi} \rangle \geq \left\|\boldsymbol{\mu}^{(i)}\right\|^2 - 2\left|\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi} \rangle\right| \geq d - 2\sigma\sqrt{d}\ln(d)\ ,$$

and

$$\left\|\boldsymbol{\mu}^{(i)} + \boldsymbol{\xi}\right\|^2 = \left\|\boldsymbol{\mu}^{(i)}\right\|^2 + \|\boldsymbol{\xi}\|^2 + 2\langle \boldsymbol{\mu}^{(i)}, \boldsymbol{\xi} \rangle \leq d + 2\sigma^2 d + 2\sigma\sqrt{d}\ln(d) \leq 3d + 2\sigma\sqrt{d}\ln(d)\ .$$

□

Our final lemma in this section shows that each cluster contains some examples with high probability.

Lemma A.8. *With probability at least $1 - d^{1-\ln(d)}$ the dataset $\mathcal{S}$ contains at least one example from each cluster in $[k]$.*

Proof. Note that this problem corresponds to the "coupons collector's problem". The probability that $\mathcal{S}$ does not contain points from cluster j is at most

$$\left(1 - \frac{1}{k}\right)^n \leq \exp\left(-\frac{n}{k}\right) \leq \exp\left(-\ln^2(d)\right) = d^{-\ln(d)}\ ,$$

where in the second inequality we used $n \geq k\ln^2(d)$. By the union bound, the probability that there is a cluster that does not appear in $\mathcal{S}$ is at most $k \cdot d^{-\ln(d)}$. Since $k \leq d$ this probability is at most $d^{1-\ln(d)}$. □

The proof of Proposition A.1 now follows by putting together Lemmas A.1, A.3, A.4, A.5, A.6, A.7,and A.8, and using $\sigma \leq 1$.

A.2 Structural implications of the KKT conditions

In this section we show that if the dataset $\mathcal{S}$ satisfies Properties (P1) through (P6), then the KKT conditions impose a number of constraints on the behavior of the neural network. We shall show that these constraints imply that the network will generalize well to unseen test data. The reader may find it useful to refer back to the beginning of Section A.1 before proceeding.

We first outline what types of structural conditions on the KKT points would be useful for understanding generalization. Suppose that $\mathbf{x}$ is a test example coming from cluster $r \in Q_+$. Our hope is that $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) > 0$ for such an example. Recall that since $\boldsymbol{\theta}$ satisfies the KKT conditions of Problem (2), then there are $\lambda_1, \dots, \lambda_n$ such that for every $j \in J$ we have

$$\mathbf{w}_j = \sum_{i \in I} \lambda_i \nabla_{\mathbf{w}_j} \left(y_i \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) = \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i \,, \tag{6}$$

where $\phi'_{i,j}$ is a subgradient of ϕ at $\mathbf{w}_j^\top \mathbf{x}_i + b_j$, i.e., if $\mathbf{w}_j^\top \mathbf{x}_i + b_j \neq 0$ then $\phi'_{i,j} = \mathbb{1}[\mathbf{w}_j^\top \mathbf{x}_i + b_j \geq 0]$, and otherwise $\phi'_{i,j}$ is some value in $[0,1]$. Also we have $\lambda_i \geq 0$ for all i, and $\lambda_i = 0$ if $y_i \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq 1$. Likewise, we have

$$b_j = \sum_{i \in I} \lambda_i \nabla_{b_j} \left(y_i \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) = \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \,. \tag{7}$$

Now, consider the network output for an input $\mathbf{x}$,

$$\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j \in J_+} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) + \sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) \geq \sum_{j \in J_+} v_j (\mathbf{w}_j^\top \mathbf{x} + b_j) + \sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j), \tag{8}$$

where we have used $\phi(z) \geq z$ for all $z \in \mathbb{R}$. This suggests the following possibility: if we can ensure that $\sum_{j \in J_+} v_j (\mathbf{w}_j^\top \mathbf{x} + b_j)$ is large and positive while $\sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j)$ is not too negative, then the network will accurately classify the example $\mathbf{x}$. Using the KKT conditions and that $r \in Q_+$ (so $y_i = 1$ for $i \in I^{(r)}$), the first term in the above decomposition is equal to

$$\begin{aligned}
\sum_{j \in J_+} v_j (\mathbf{w}_j^\top \mathbf{x} + b_j) &= \sum_{j \in J_+} v_j \left[\sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right] \\
&= \sum_{j \in J_+} \left[\left(\sum_{i \in I^{(r)}} \lambda_i v_j^2 \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right) + \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i y_i v_j^2 \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right] \\
&\geq \left(\sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right) - \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} |\mathbf{x}_i^\top \mathbf{x} + 1| \,.
\end{aligned}$$

Since $\mathbf{x}$ comes from cluster r and the clusters are nearly orthogonal, the pairwise correlations $\mathbf{x}_i^\top \mathbf{x}$ will be large and positive when $i \in I^{(r)}$ but will be small in magnitude when $i \in I^{(q)}$ for $q \neq r$. Thus, we can hope that this term will be large and positive if we can show that the quantity $\sum_{i \in I^{(r)}} \sum_{j \in J+} \lambda_i v_j^2 \phi'_{i,j}$ is not too small relative to the quantity $\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j}$. By similar arguments, in order to show the second term in Eq. (8) is not too negative, we need to understand how the quantity $\sum_{i \in I^{(q)}} \sum_{j \in J_-} \lambda_i v_j^2 \phi'_{i,j}$ varies across different clusters $q \in Q$.

The above sketch motivates a characterization of how the quantities

$$\sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j}, \quad \sum_{i \in I^{(q)}} \sum_{j \in J_-} \lambda_i v_j^2 \phi'_{i,j}$$

relate to each other for different clusters $q \in Q$. We will obtain upper and lower bounds for these quantities in Lemmas A.10 and A.11 below. We now proceed with the proof.

Recall that $\Delta := 4\sigma\sqrt{d}\ln(d)$. Since by Assumption 2.2 we have $k \leq \frac{d-\Delta+1}{10(p+\Delta+1)}$, we let $c' \leq \frac{1}{10}$ be such that $k = c' \cdot \frac{d-\Delta+1}{p+\Delta+1}$. Note that $d > \Delta$, and more precisely, the following holds

Lemma A.9. *We have* $\Delta \leq \frac{d}{21}$.

Proof. Recall that $k\,(p + \Delta + 1) \leq \frac{d-\Delta+1}{10}$. Since $k \geq 2$ and $p \geq 0$ it implies that $2(\Delta + 1) \leq \frac{d-\Delta+1}{10}$. Hence, $\Delta \leq \frac{d-19}{21} \leq \frac{d}{21}$. □

We now show that the sums of the form $\sum_{i \in I^{(q)}} \sum_{j \in J_\circ} v_j^2 \lambda_i \phi'_{i,j}$, for $\circ \in \{+, -\}$, are never too large.

Lemma A.10. *If $\mathcal{S}$ satisfies the properties (P1) through (P6), then for all $q \in Q$ we have*

$$\max\left\{\sum_{i\in I^{(q)}}\sum_{j\in J_+} v_j^2\lambda_i\phi'_{i,j}, \sum_{i\in I^{(q)}}\sum_{j\in J_-} v_j^2\lambda_i\phi'_{i,j}\right\} \leq \frac{1}{(1-2c')(d-\Delta+1)} .$$

Proof. Let $\alpha_+ = \max_{q\in Q}\left(\sum_{i\in I^{(q)}}\sum_{j\in J_+} v_j^2\lambda_i\phi'_{i,j}\right)$, and let $\alpha_- = \max_{q\in Q}\left(\sum_{i\in I^{(q)}}\sum_{j\in J_-} v_j^2\lambda_i\phi'_{i,j}\right)$. Assume w.l.o.g. that $\alpha_+ \geq \alpha_-$ (the proof for the case $\alpha_+ < \alpha_-$ is similar). Let $\alpha = \alpha_+$ and $r \in \operatorname{argmax}_{q\in Q}\left(\sum_{i\in I^{(q)}}\sum_{j\in J_+} v_j^2\lambda_i\phi'_{i,j}\right)$. Assume towards contradiction that $\alpha > \frac{1}{(1-2c')(d-\Delta+1)}$. Note that we have $\sum_{i\in I^{(r)}}\lambda_i > 0$, since otherwise $\alpha = 0$. Hence, there exists $i' \in I^{(r)}$ with $\lambda_{i'} > 0$, and thus $y_{i'}\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_{i'}) = 1$.
By Eq. (6) and (7) for every $j \in J$ we have

$$\begin{aligned}\mathbf{w}_j^\top\mathbf{x}_{i'} + b_j &= \sum_{i\in I}\lambda_i y_i v_j\phi'_{i,j}\mathbf{x}_i^\top\mathbf{x}_{i'} + \sum_{i\in I}\lambda_i y_i v_j\phi'_{i,j} \\ &= \sum_{i\in I}\lambda_i y_i v_j\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}_{i'} + 1) \\ &= \left(\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i y_i v_j\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}_{i'} + 1)\right) + \sum_{i\in I^{(r)}}\lambda_i y_i v_j\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}_{i'} + 1) . \qquad (9)\end{aligned}$$

We consider two cases:

Case 1: Assume that $r \in Q_+$. We have

$$\begin{aligned}1 &= y_{i'}\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_{i'}) \\ &= 1\cdot\sum_{j\in J} v_j\phi(\mathbf{w}_j^\top\mathbf{x}_{i'} + b_j) \\ &\geq \sum_{j\in J_+} v_j(\mathbf{w}_j^\top\mathbf{x}_{i'} + b_j) + \sum_{j\in J_-} v_j\phi(\mathbf{w}_j^\top\mathbf{x}_{i'} + b_j) . \qquad (10)\end{aligned}$$

By the case assumption $r \in Q_+$, Eq. (9) and our assumptions on the dataset $\mathcal{S}$, we have

$$\begin{aligned}\sum_{j\in J_+} v_j(\mathbf{w}_j^\top\mathbf{x}_{i'} + b_j) &= \sum_{j\in J_+}\left[\left(\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i y_i v_j^2\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}_{i'} + 1)\right) + \sum_{i\in I^{(r)}}\lambda_i y_i v_j^2\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}_{i'} + 1)\right] \\ &\geq \sum_{j\in J_+}\left[\left(-\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i v_j^2\phi'_{i,j}(p+\Delta+1)\right) + \sum_{i\in I^{(r)}}\lambda_i v_j^2\phi'_{i,j}(d-\Delta+1)\right] \\ &= \left(-(p+\Delta+1)\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\sum_{j\in J_+}\lambda_i v_j^2\phi'_{i,j}\right) + (d-\Delta+1)\sum_{i\in I^{(r)}}\sum_{j\in J_+}\lambda_i v_j^2\phi'_{i,j} \\ &\geq -(p+\Delta+1)k\alpha + (d-\Delta+1)\alpha . \qquad (11)\end{aligned}$$

In the last line we have used the definition $\alpha = \max_{q \in Q} \left(\sum_{i \in I^{(q)}} \sum_{j \in J_+} v_j^2 \lambda_i \phi'_{i,j} \right)$ and that the cluster with index r achieves this maximum. Moreover, using Eq. (9) again we have

$$\begin{aligned}
\sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x}_{i'} + b_j) &= \sum_{j \in J_-} v_j \phi \left(\left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x}_{i'} + 1) \right) + \sum_{i \in I^{(r)}} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x}_{i'} + 1) \right) \\
&\geq \sum_{j \in J_-} v_j \phi \left(\left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i |v_j| \phi'_{i,j} (p + \Delta + 1) \right) + \sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} (d - \Delta + 1) \right) \\
&\geq \sum_{j \in J_-} v_j \phi \left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i |v_j| \phi'_{i,j} (p + \Delta + 1) \right) \\
&= \sum_{j \in J_-} v_j \left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i |v_j| \phi'_{i,j} (p + \Delta + 1) \right) \\
&= -(p + \Delta + 1) \left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_-} \lambda_i v_j^2 \phi'_{i,j} \right) \\
&\geq -(p + \Delta + 1) k \alpha \; . \qquad (12)
\end{aligned}$$

The first inequality above uses the properties of the dataset $\mathcal{S}$, that $j \in J_-$ and ϕ is non-decreasing, as well as the case assumption that $r \in Q_+$. The last inequality uses the definition of α. Combining Eq. (10), (11), and (12) we get

$$\begin{aligned}
1 &\geq -(p + \Delta + 1) k \alpha + (d - \Delta + 1) \alpha - (p + \Delta + 1) k \alpha \\
&= \alpha \left((d - \Delta + 1) - 2k (p + \Delta + 1) \right) \\
&= \alpha \left((d - \Delta + 1) - 2 \cdot \frac{c'(d - \Delta + 1)}{p + \Delta + 1} (p + \Delta + 1) \right) \\
&= \alpha \left((d - \Delta + 1) - 2c'(d - \Delta + 1) \right) \\
&= \alpha (d - \Delta + 1)(1 - 2c') \\
&> \frac{1}{(1 - 2c')(d - \Delta + 1)} (d - \Delta + 1)(1 - 2c') \\
&= 1 \; ,
\end{aligned}$$

where in the last inequality we used our assumption on α. We have thus reached a contradiction following our assumption on α in the case where $r \in Q_+$.

Case 2: Assume that $r \in Q_-$. Fix some $j \in J_+$. If $\phi'_{i,j} = 0$ for every $i \in I^{(r)}$ then

$$\sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} (d - \Delta + 1) = 0 \leq \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} (p + \Delta + 1) \; . \qquad (13)$$

Otherwise, i.e., if there is $s \in I^{(r)}$ such that $\phi'_{s,j} > 0$, then by the definition of $\phi'_{s,j}$ we have $\mathbf{w}_j^\top \mathbf{x}_s + b_j \geq 0$, and hence by Eq. (9) we have

$$\begin{aligned}
0 &\leq \mathbf{w}_j^\top \mathbf{x}_s + b_j \\
&= \left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x}_s + 1) \right) + \sum_{i \in I^{(r)}} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x}_s + 1) \\
&\leq \left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} (p + \Delta + 1) \right) - \sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} (d - \Delta + 1)
\end{aligned}$$

Hence, we get again an expression similar to Eq. (13). Thus for any $j \in J_+$ we have,

$$\sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} \leq \frac{p + \Delta + 1}{d - \Delta + 1} \left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} \right) \; .$$

Since this holds for every $j \in J_+$, we get

$$\begin{aligned}
\sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} &= \sum_{j \in J_+} v_j \sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} \\
&\leq \sum_{j \in J_+} v_j \cdot \frac{p+\Delta+1}{d-\Delta+1} \left(\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} \right) \\
&= \frac{p+\Delta+1}{d-\Delta+1} \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \\
&\leq \frac{p+\Delta+1}{d-\Delta+1} \cdot k \cdot \max_{q \in Q} \left(\sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right) \\
&= \frac{p+\Delta+1}{d-\Delta+1} \cdot \frac{c'(d-\Delta+1)}{p+\Delta+1} \cdot \max_{q \in Q} \left(\sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right) \\
&< \max_{q \in Q} \left(\sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right) .
\end{aligned}$$

Since $r \in \operatorname{argmax}_{q \in Q} \left(\sum_{i \in I^{(q)}} \sum_{j \in J_+} v_j^2 \lambda_i \phi'_{i,j} \right)$, we have reached a contradiction following our assumption on α for the case $r \in Q_-$. This completes the proof that we must have $\alpha \leq \frac{1}{(1-2c')(d-\Delta+1)}$. ☐

We next show that the relevant sums of the form $\sum_{i \in I^{(q)}} \sum_{j \in J_\circ} v_j^2 \lambda_i \phi'_{i,j}$, for $\circ \in \{+,-\}$, are never too small.

Lemma A.11. *If $\mathcal{S}$ satisfies the properties (P1) through (P6), then for all $q \in Q_+$ we have*

$$\sum_{i \in I^{(q)}} \sum_{j \in J_+} v_j^2 \lambda_i \phi'_{i,j} \geq \left(1 - \frac{c'}{1-2c'}\right) \frac{1}{3d+\Delta+1} ,$$

and for all $q \in Q_-$ we have

$$\sum_{i \in I^{(q)}} \sum_{j \in J_-} v_j^2 \lambda_i \phi'_{i,j} \geq \left(1 - \frac{c'}{1-2c'}\right) \frac{1}{3d+\Delta+1} .$$

Proof. Let $r \in Q_+$ and let $s \in I^{(r)}$. We have

$$1 \leq \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}_s) = \sum_{j \in J} v_j \phi(\mathbf{w}_j^\top \mathbf{x}_s + b_j) \leq \sum_{j \in J_+} v_j \phi(\mathbf{w}_j^\top \mathbf{x}_s + b_j) \leq \sum_{j \in J_+} v_j \left| \mathbf{w}_j^\top \mathbf{x}_s + b_j \right| .$$

By Eq. (6) and (7), since $r \in Q_+$ the above equals

$$\begin{aligned}
\sum_{j \in J_+} v_j & \left| \sum_{i \in I} \left(\lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top \mathbf{x}_s + \lambda_i y_i v_j \phi'_{i,j} \right) \right| \leq \sum_{j \in J_+} v_j \sum_{q \in Q} \sum_{i \in I^{(q)}} \left| \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x}_s + 1) \right| \\
&= \sum_{j \in J_+} v_j \left[\left(\sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} \left| \mathbf{x}_i^\top \mathbf{x}_s + 1 \right| \right) + \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} \left| \mathbf{x}_i^\top \mathbf{x}_s + 1 \right| \right] \\
&= \left(\sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \left| \mathbf{x}_i^\top \mathbf{x}_s + 1 \right| \right) + \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \left| \mathbf{x}_i^\top \mathbf{x}_s + 1 \right| \\
&\leq \left((3d+\Delta+1) \sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right) + (p+\Delta+1) \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j}
\end{aligned}$$

The final inequality uses the properties of the dataset $\mathcal{S}$. Combining the above with Lemma A.10 we get

$$\begin{aligned}1 &\leq \left((3d+\Delta+1) \sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right) + (p+\Delta+1)k \cdot \frac{1}{(1-2c')(d-\Delta+1)} \\ &= \left((3d+\Delta+1) \sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right) + (p+\Delta+1) \cdot \frac{c'(d-\Delta+1)}{p+\Delta+1} \cdot \frac{1}{(1-2c')(d-\Delta+1)} \\ &= \left((3d+\Delta+1) \sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right) + \frac{c'}{1-2c'} \,.\end{aligned}$$

Therefore,

$$\sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \geq \left(1 - \frac{c'}{1-2c'}\right) \frac{1}{3d+\Delta+1} \,.$$

By similar arguments with $r \in Q_-$ we also get

$$\sum_{i \in I^{(r)}} \sum_{j \in J_-} \lambda_i v_j^2 \phi'_{i,j} \geq \left(1 - \frac{c'}{1-2c'}\right) \frac{1}{3d+\Delta+1} \,.$$

□

Recall that test examples will come from one of the k nearly-orthogonal clusters. Since the clusters are nearly-orthogonal, the pairwise correlations between the test example and training data from the same cluster will be much larger than the pairwise correlations between the test example and training data from the other (nearly-orthogonal) clusters. To characterize the decision boundary of the neural network on test data it therefore suffices to characterize the decision boundary for an example $\mathbf{x}$ that is (1) highly correlated to examples from a given cluster and (2) nearly-orthogonal to samples from other clusters. The next lemma leverages the structural conditions provided in Lemmas A.10 and A.11 to show exactly this.

Lemma A.12. *Suppose $\mathcal{S}$ satisfies the properties (P1) through (P6). Let $\mathbf{x} \in \mathbb{R}^d$ and $r \in Q$ be such that for all $i \in I^{(r)}$ we have $\langle \mathbf{x}, \mathbf{x}_i \rangle \in [d-\Delta, d+\Delta]$, and for all $q \in Q \setminus \{r\}$ and $i \in I^{(q)}$ we have $|\langle \mathbf{x}, \mathbf{x}_i \rangle| \leq p + \Delta$. Then,* $\operatorname{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x})) = y^{(r)}$.

Proof. We prove the claim for $r \in Q_+$. The proof for $r \in Q_-$ is similar. By Eq. (6) and (7), for every $j \in J$ we have

$$\begin{aligned}\mathbf{w}_j^\top \mathbf{x} + b_j &= \left(\sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top \mathbf{x} \right) + \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \\ &= \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \\ &= \left(\sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right) + \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \,. \end{aligned} \tag{14}$$

Now,

$$\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j \in J} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) \geq \sum_{j \in J_+} v_j (\mathbf{w}_j^\top \mathbf{x} + b_j) + \sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) \,. \tag{15}$$

By Eq. (14) we have

$$\begin{aligned}\sum_{j\in J_+} v_j(\mathbf{w}_j^\top \mathbf{x}+b_j) &= \sum_{j\in J_+}\left[\left(\sum_{i\in I^{(r)}}\lambda_i v_j^2\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}+1)\right)+\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i y_i v_j^2\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}+1)\right]\\
&\geq \sum_{j\in J_+}\left[\left(\sum_{i\in I^{(r)}}\lambda_i v_j^2\phi'_{i,j}(d-\Delta+1)\right)-\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i v_j^2\phi'_{i,j}(p+\Delta+1)\right]\\
&= \left((d-\Delta+1)\sum_{i\in I^{(r)}}\sum_{j\in J_+}\lambda_i v_j^2\phi'_{i,j}\right)\\
&\quad -\left((p+\Delta+1)\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\sum_{j\in J_+}\lambda_i v_j^2\phi'_{i,j}\right)\,.\end{aligned}$$

By Lemma A.10 and Lemma A.11 the above is at least

$$\begin{aligned}&(d-\Delta+1)\left(1-\frac{c'}{1-2c'}\right)\frac{1}{3d+\Delta+1}-(p+\Delta+1)k\cdot\frac{1}{(1-2c')(d-\Delta+1)}\\
&\quad=\left(1-\frac{c'}{1-2c'}\right)\frac{d-\Delta+1}{3d+\Delta+1}-(p+\Delta+1)c'\cdot\frac{d-\Delta+1}{p+\Delta+1}\cdot\frac{1}{(1-2c')(d-\Delta+1)}\\
&\quad=\left(1-\frac{c'}{1-2c'}\right)\frac{d-\Delta+1}{3d+\Delta+1}-\frac{c'}{1-2c'}\,. \qquad (16)\end{aligned}$$

Likewise, we have

$$\begin{aligned}\sum_{j\in J_-} v_j\phi(\mathbf{w}_j^\top \mathbf{x}+b_j) &= \sum_{j\in J_-} v_j\phi\left(\left(\sum_{i\in I^{(r)}}\lambda_i v_j\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}+1)\right)+\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i y_i v_j\phi'_{i,j}(\mathbf{x}_i^\top\mathbf{x}+1)\right)\\
&\geq \sum_{j\in J_-} v_j\phi\left(\left(\sum_{i\in I^{(r)}}\lambda_i v_j\phi'_{i,j}(d-\Delta+1)\right)+\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i |v_j|\phi'_{i,j}(p+\Delta+1)\right)\\
&\geq \sum_{j\in J_-} v_j\phi\left(\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i |v_j|\phi'_{i,j}(p+\Delta+1)\right)\\
&= -\sum_{j\in J_-}\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\lambda_i v_j^2\phi'_{i,j}(p+\Delta+1)\\
&= -(p+\Delta+1)\sum_{q\in Q\setminus\{r\}}\sum_{i\in I^{(q)}}\sum_{j\in J_-}\lambda_i v_j^2\phi'_{i,j}\,.\end{aligned}$$

By Lemma A.10 the above is at least

$$\begin{aligned}-(p+\Delta+1)k\cdot\frac{1}{(1-2c')(d-\Delta+1)} &= -(p+\Delta+1)c'\cdot\frac{d-\Delta+1}{p+\Delta+1}\cdot\frac{1}{(1-2c')(d-\Delta+1)}\\
&= -\frac{c'}{1-2c'}\,. \qquad (17)\end{aligned}$$

Combining Eq. (15), (16), and (17), we get

$$\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) \geq \left(1-\frac{c'}{1-2c'}\right)\frac{d-\Delta+1}{3d+\Delta+1}-\frac{c'}{1-2c'}-\frac{c'}{1-2c'}\,.$$

Using $c' \leq \frac{1}{10}$ and $\Delta \leq d$ (which holds by Lemma A.9), the above is at least

$$\frac{7}{8}\cdot\frac{d-\Delta+1}{3d+\Delta+1}-\frac{2}{8} \geq \frac{7}{8}\cdot\frac{d-\Delta}{3d+\Delta}-\frac{2}{8}\,.$$

By Lemma A.9, the displayed equation is at least

$$\frac{7}{8} \cdot \frac{d - d/21}{3d + d/21} - \frac{2}{8} = \frac{7}{8} \cdot \frac{5}{16} - \frac{2}{8} > 0 \,.$$

□

A.3 Generalization from KKT conditions

Lemma A.12 shows that in order to show generalization, it suffices to show that with high probability, a test example is highly correlated to one cluster and nearly-orthogonal to all other clusters. In this section we prove that this is the case. We shall re-apply many of the concentration bounds provided in Section A.1 to do so.

Lemma A.13. *Suppose $\mathcal{S}$ satisfies Properties (P1) through (P6). Let $r \in Q$ and let $\mathbf{x} = \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I_d)$. With probability at least $1 - 4nd^{-\frac{\ln(d)}{2}}$ over $\mathbf{x}$ the following hold for all $i \in I$:*

- $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \leq \Delta$.
- $|\langle \mathbf{x}_i, \boldsymbol{\xi} \rangle| \leq \Delta$.
- *If $i \notin I^{(r)}$ then $|\langle \mathbf{x}_i, \mathbf{x} \rangle| \leq p + \Delta$.*
- *If $i \in I^{(r)}$ then $|\langle \mathbf{x}_i, \mathbf{x} \rangle - d| \leq \Delta$.*

Proof. By our assumption on the dataset $\mathcal{S}$, for all $i \in I$ and $q \in Q$ we have $\|\boldsymbol{\xi}_i\| \leq \sigma\sqrt{2d}$ and $\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi}_i \rangle \leq \sigma\sqrt{d}\ln(d)$. By Lemma A.2 and since $\sigma \leq 1$, for $i \in I$ we have

$$\Pr\left[|\langle \boldsymbol{\xi}, \boldsymbol{\xi}_i \rangle| \geq 2\sigma\sqrt{d}\ln(d)\right] \leq \Pr\left[|\langle \boldsymbol{\xi}, \boldsymbol{\xi}_i \rangle| \geq \sigma\sqrt{2d} \cdot \sigma\ln(d)\right] \leq \Pr\left[|\langle \boldsymbol{\xi}, \boldsymbol{\xi}_i \rangle| \geq \|\boldsymbol{\xi}_i\| \, \sigma\ln(d)\right] \leq 2d^{-\frac{\ln(d)}{2}} \,,$$

and by Lemma A.4 for $q \in Q$ we have

$$\Pr\left[|\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi} \rangle| \geq \sigma\sqrt{d}\ln(d)\right] \leq 2d^{-\frac{\ln(d)}{2}} \,.$$

Fix some $q \in Q$ and $i \in I^{(q)}$. With probability at least $1 - 4d^{-\frac{\ln(d)}{2}}$ over $\boldsymbol{\xi}$, we have $|\langle \boldsymbol{\xi}, \boldsymbol{\xi}_i \rangle| \leq 2\sigma\sqrt{d}\ln(d)$ and $|\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi} \rangle| \leq \sigma\sqrt{d}\ln(d)$.

Then, the following hold:

- $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \leq \Delta$.
- We have
$$|\langle \mathbf{x}_i, \boldsymbol{\xi} \rangle| \leq |\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi} \rangle| + |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \leq \sigma\sqrt{d}\ln(d) + 2\sigma\sqrt{d}\ln(d) \leq \Delta \,.$$
- If $q \neq r$, then
$$\begin{aligned} |\langle \mathbf{x}, \mathbf{x}_i \rangle| &= |\langle \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}, \boldsymbol{\mu}^{(q)} + \boldsymbol{\xi}_i \rangle| \\ &\leq |\langle \boldsymbol{\mu}^{(r)}, \boldsymbol{\mu}^{(q)} \rangle| + |\langle \boldsymbol{\mu}^{(r)}, \boldsymbol{\xi}_i \rangle| + |\langle \boldsymbol{\xi}, \boldsymbol{\mu}^{(q)} \rangle| + |\langle \boldsymbol{\xi}, \boldsymbol{\xi}_i \rangle| \\ &\leq p + \sigma\sqrt{d}\ln(d) + \sigma\sqrt{d}\ln(d) + 2\sigma\sqrt{d}\ln(d) \\ &= p + \Delta \,. \end{aligned}$$
- If $q = r$, then
$$\begin{aligned} |\langle \mathbf{x}, \mathbf{x}_i \rangle - d| &= \left| \langle \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}, \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}_i \rangle - d \right| \\ &= \left| \langle \boldsymbol{\mu}^{(r)}, \boldsymbol{\xi}_i \rangle + \langle \boldsymbol{\xi}, \boldsymbol{\mu}^{(r)} \rangle + \langle \boldsymbol{\xi}, \boldsymbol{\xi}_i \rangle \right| \\ &\leq |\langle \boldsymbol{\mu}^{(r)}, \boldsymbol{\xi}_i \rangle| + |\langle \boldsymbol{\xi}, \boldsymbol{\mu}^{(q)} \rangle| + |\langle \boldsymbol{\xi}, \boldsymbol{\xi}_i \rangle| \\ &\leq \sigma\sqrt{d}\ln(d) + \sigma\sqrt{d}\ln(d) + 2\sigma\sqrt{d}\ln(d) \\ &= \Delta \,. \end{aligned}$$

Overall, by the union bound, with probability at least $1 - 4nd^{-\frac{\ln(d)}{2}}$ the requirements hold for all $i \in I$. $\square$

Theorem A.1 now follows immediately from Proposition A.1 and Lemmas A.12 and A.13.

B Proof of Lemma 2.1

Let $\mathbf{x} = \boldsymbol{\mu}^{(j)} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I_d)$, and $y = y^{(j)}$. Then,

$$y\mathbf{u}^\top \mathbf{x} = y^{(j)} \sum_{q \in [k]} y^{(q)} (\boldsymbol{\mu}^{(q)})^\top (\boldsymbol{\mu}^{(j)} + \boldsymbol{\xi}) \geq \left\|\boldsymbol{\mu}^{(j)}\right\|^2 - k \left(\max_{q \neq j} |(\boldsymbol{\mu}^{(q)})^\top \boldsymbol{\mu}^{(j)}| + \max_{q \in [k]} |(\boldsymbol{\mu}^{(q)})^\top \boldsymbol{\xi}| \right) .$$

By Lemma A.4 we have $\Pr\left[|\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi} \rangle| \geq \sigma\sqrt{d} \ln(d) \right] \leq 2d^{-\frac{\ln(d)}{2}}$. Hence, by the union bound with probability at least $1 - 2kd^{-\ln(d)/2} \geq 1 - 2d^{1 - \ln(d)/2}$ we have $\max_{q \in [k]} |(\boldsymbol{\mu}^{(q)})^\top \boldsymbol{\xi}| \leq \sigma\sqrt{d} \ln(d)$. Note that by Assumption 2.2 we must have $k \leq d$. Using Assumption 2.2 again, we conclude that with probability at least $1 - 2d^{1-\ln(d)/2}$ we have

$$y\mathbf{u}^\top \mathbf{x} \geq d - k \left(\max_{q \neq j} |(\boldsymbol{\mu}^{(q)})^\top \boldsymbol{\mu}^{(j)}| + \sigma\sqrt{d}\ln(d) \right) \geq d - \frac{d - 4\sigma\sqrt{d}\ln(d) + 1}{10} \geq \frac{9d - 1}{10} > 0 .$$

C Proof of Theorem 4.1

We prove the theorem for $r = k$. The proof for $r > k$ follows immediately by adding zero-weight neurons. Consider the network $\mathcal{N}(\mathbf{x}) = \sum_{j=1}^k v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j)$ such that for every $j \in [k]$ we have $v_j = y^{(j)}$, $\mathbf{w}_j = \frac{4\boldsymbol{\mu}^{(j)}}{d}$, and $b_j = -2$. Let $q \sim \mathcal{U}([k])$, let $\mathbf{x} = \boldsymbol{\mu}^{(q)} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I_d)$, and let $y = y^{(q)}$. Thus $(\mathbf{x}, y)$ is drawn from $\mathcal{D}_{\text{clusters}}$. We have,

$$\mathbf{w}_q^\top \mathbf{x} + b_q = \frac{4(\boldsymbol{\mu}^{(q)})^\top (\boldsymbol{\mu}^{(q)} + \boldsymbol{\xi})}{d} - 2 = \frac{4(d + \langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi} \rangle)}{d} - 2 .$$

By Lemma A.4, we have $\Pr\left[|\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi} \rangle| \geq \sigma\sqrt{d}\ln(d) \right] \leq 2d^{-\frac{\ln(d)}{2}}$. Hence, with probability at least $1 - 2d^{-\frac{\ln(d)}{2}}$ over $\boldsymbol{\xi}$, for large enough d we have

$$\mathbf{w}_q^\top \mathbf{x} + b_q \geq \frac{4(d - \sigma\sqrt{d}\ln(d))}{d} - 2 \geq \frac{4(d - \sqrt{d}\ln(d))}{d} - 2 = 2 - \frac{4\ln(d)}{\sqrt{d}} \geq 1 . \qquad (18)$$

For $j \neq q$ we have

$$\mathbf{w}_j^\top \mathbf{x} + b_j = \frac{4(\boldsymbol{\mu}^{(j)})^\top (\boldsymbol{\mu}^{(q)} + \boldsymbol{\xi})}{d} - 2 \leq \frac{4\max_{j \neq q} |\langle \boldsymbol{\mu}^{(j)}, \boldsymbol{\mu}^{(q)} \rangle|}{d} + \frac{4(\boldsymbol{\mu}^{(j)})^\top \boldsymbol{\xi}}{d} - 2 ,$$

and thus with probability at least $1 - 2d^{-\frac{\ln(d)}{2}}$ over $\boldsymbol{\xi}$ we have

$$\mathbf{w}_j^\top \mathbf{x} + b_j \leq -2 + \frac{4}{d} \left(\max_{j \neq q} |\langle \boldsymbol{\mu}^{(j)}, \boldsymbol{\mu}^{(q)} \rangle| + \sigma\sqrt{d}\ln(d) \right) .$$

Since by Assumption 2.2 we have $k\left(\max_{j \neq q} |\langle \boldsymbol{\mu}^{(j)}, \boldsymbol{\mu}^{(q)} \rangle| + 4\sigma\sqrt{d}\ln(d) + 1 \right) \leq \frac{d - 4\sigma\sqrt{d}\ln(d) + 1}{10}$, then the displayed equation is at most

$$-2 + \frac{4}{d} \cdot \frac{d+1}{10k} \leq -2 + \frac{4}{d} \cdot \frac{2d}{10} \leq -1 . \qquad (19)$$

Overall, by the union bound, with probability at least $1 - 2kd^{-\frac{\ln(d)}{2}} \geq 1 - 2d^{1-\frac{\ln(d)}{2}} = 1 - o_d(1)$ we have $\mathcal{N}(\mathbf{x}) = \sum_{j=1}^k v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) = v_q(\mathbf{w}_q^\top \mathbf{x} + b_q) + 0$, and

$$\operatorname{sign}(\mathcal{N}(\mathbf{x})) = \operatorname{sign}(v_q) = y^{(q)} = y .$$

We now prove that $\mathcal{N}$ is $\sqrt{d}$-robust w.r.t. $\mathcal{D}_{\mathbf{x}}$. Thus, we show that with probability at least $1 - o_d(1)$ over $\mathbf{x}$, for every $\mathbf{x}' \in \mathbb{R}^d$ such that $\|\mathbf{x} - \mathbf{x}'\| \leq \frac{\sqrt{d}}{8}$ we have $\text{sign}(\mathcal{N}(\mathbf{x})) = \text{sign}(\mathcal{N}(\mathbf{x}'))$. Note that with probability $1 - o_d(1)$, Eq. (18) holds, and Eq. (19) holds for all $j \neq q$, and hence

$$\mathbf{w}_q^\top \mathbf{x}' + b_q = \mathbf{w}_q^\top (\mathbf{x}' - \mathbf{x}) + \left(\mathbf{w}_q^\top \mathbf{x} + b_q\right) \geq -\|\mathbf{w}_q\| \cdot \|\mathbf{x}' - \mathbf{x}\| + 1 \geq -\frac{4}{\sqrt{d}} \cdot \frac{\sqrt{d}}{8} + 1 = \frac{1}{2} ,$$

and for all $j \neq q$ we have

$$\mathbf{w}_j^\top \mathbf{x}' + b_j = \mathbf{w}_j^\top (\mathbf{x}' - \mathbf{x}) + \left(\mathbf{w}_j^\top \mathbf{x} + b_j\right) \leq \|\mathbf{w}_j\| \cdot \|\mathbf{x}' - \mathbf{x}\| - 1 \leq \frac{4}{\sqrt{d}} \cdot \frac{\sqrt{d}}{8} - 1 = -\frac{1}{2} .$$

Therefore, $\text{sign}(\mathcal{N}(\mathbf{x}')) = \text{sign}(v_q) = y^{(q)} = \text{sign}(\mathcal{N}(\mathbf{x}))$.

D Proof of Theorem 4.2

We will prove the following theorem:

Theorem D.1. *Let $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{-1, 1\}$ be a training set drawn i.i.d. from the distribution $\mathcal{D}_{\text{clusters}}$, where $n \geq k \ln^2(d)$. We denote $Q_+ = \{q \in [k] : y^{(q)} = 1\}$ and $Q_- = \{q \in [k] : y^{(q)} = -1\}$, and assume that $\min\left\{\frac{|Q_+|}{k}, \frac{|Q_-|}{k}\right\} \geq c$ for some $c > 0$. Let $\mathcal{N}_{\boldsymbol{\theta}}$ be a depth-2 ReLU network such that $\boldsymbol{\theta} = [\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{b}, \mathbf{v}]$ is a KKT point of Problem (2). Then, with probability at least*

$$1 - \left(3n^2 d^{-\frac{\ln(d)}{2}} + n^2 e^{-d/16} + d^{1-\ln(d)}\right)$$

over $\mathcal{S}$, there is a vector $\mathbf{z} = \eta \cdot \sum_{j \in [k]} y^{(j)} \boldsymbol{\mu}^{(j)}$ with $\eta > 0$ and $\|\mathbf{z}\| \leq \mathcal{O}\left(\sqrt{\frac{d}{c^2 k}}\right)$, such that

$$\Pr_{(\mathbf{x},y) \sim \mathcal{D}_{\text{clusters}}} [\text{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x})) \neq \text{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x} - y\mathbf{z}))] \geq 1 - 4nd^{-\frac{\ln(d)}{2}} .$$

It is easy to verify that Theorem D.1 implies Theorem 4.2. Indeed, if $\frac{1}{\delta} \leq \frac{1}{3} d^{\ln(d)-1}$ and

$$n \leq \min\left\{\sqrt{\frac{\delta}{3}} \cdot e^{d/32}, \frac{\sqrt{\delta}}{3} \cdot d^{\ln(d)/4}, \frac{\epsilon}{4} \cdot d^{\ln(d)/2}\right\} ,$$

then we showed in Appendix A that

- $3n^2 d^{-\frac{\ln(d)}{2}} \leq \frac{\delta}{3}$.
- $n^2 e^{-d/16} \leq \frac{\delta}{3}$.
- $d^{1-\ln(d)} \leq \frac{\delta}{3}$.
- $4nd^{-\frac{\ln(d)}{2}} \leq \epsilon$.

Hence, under the above assumptions on δ and n, Theorem D.1 implies that with probability at least $1 - \delta$ over $\mathcal{S}$, we have $\Pr_{(\mathbf{x},y) \sim \mathcal{D}_{\text{clusters}}} [\text{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x})) \neq \text{sign}(\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x} - y\mathbf{z}))] \geq 1 - \epsilon$.

We now turn to prove Theorem D.1. The reader may find it useful to refer back to the notations from the proof of Theorem 3.1 in Section A. We shall show that when the dataset $\mathcal{S}$ satisfies the "nice" properties outlined in Properties (P1) through (P6), then every KKT point of Problem (2) is non-robust in the sense stated in the theorem. By Proposition A.1, the dataset $\mathcal{S}$ satisfies these "nice" properties with probability at least $1 - \left(3n^2 d^{-\frac{\ln(d)}{2}} + n^2 e^{-d/16} + d^{1-\ln(d)}\right)$.

In the following lemma, we state several "nice" properties that are satisfied w.h.p. in a test example.

Lemma D.1. *Suppose $\mathcal{S}$ satisfies Properties (P1) through (P6). With probability at least $1 - 4nd^{-\frac{\ln(d)}{2}}$ over $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$ there exists $r \in Q$ such that the following hold: $\mathbf{x} = \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}$ where for all $i \in I$ we have $|\langle \mathbf{x}_i, \boldsymbol{\xi} \rangle| \leq \Delta$ and $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \leq \Delta$. Also, $|\langle \mathbf{x}_i, \mathbf{x} \rangle - d| \leq \Delta$ for all $i \in I^{(r)}$ and $|\langle \mathbf{x}_i, \mathbf{x} \rangle| \leq p + \Delta$ for all $i \notin I^{(r)}$.*

Proof. The lemma follows immediately from Lemma A.13. □

We next show that if the training data and test data are "nice", then KKT points have network outputs that are not too far from the margin.

Lemma D.2. *Suppose $\mathcal{S}$ satisfies Properties (P1) through (P6). Let $\mathbf{x} \in \mathbb{R}^d$ and $r \in Q$ such that for all $i \in I^{(r)}$ we have $\langle \mathbf{x}, \mathbf{x}_i \rangle \in [d - \Delta, d + \Delta]$, and for all $q \in Q \setminus \{r\}$ and $i \in I^{(q)}$ we have $|\langle \mathbf{x}, \mathbf{x}_i \rangle| \leq p + \Delta$. Then, $|\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x})| \leq 2$.*

Proof. We prove the claim for $r \in Q_+$. The proof for $r \in Q_-$ is similar. By Lemma A.12 we have $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) > 0$. We now show that $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) \leq 2$.

By Eq. (6) and (7), for every $j \in J$ we have

$$\begin{aligned}
\mathbf{w}_j^\top \mathbf{x} + b_j &= \left(\sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top \mathbf{x} \right) + \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \\
&= \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \\
&= \left(\sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right) + \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1)\ . \qquad (20)
\end{aligned}$$

Now,

$$\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j \in J} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) \leq \sum_{j \in J_+} v_j \left| \mathbf{w}_j^\top \mathbf{x} + b_j \right|\ .$$

By Eq. (20) the above is at most

$$\begin{aligned}
&\sum_{j \in J_+} v_j \left[\left(\sum_{i \in I^{(r)}} \lambda_i v_j \phi'_{i,j} (d + \Delta + 1) \right) + \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} (p + \Delta + 1) \right] \\
&= \sum_{j \in J_+} \left[\left(\sum_{i \in I^{(r)}} \lambda_i v_j^2 \phi'_{i,j} (d + \Delta + 1) \right) + \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i v_j^2 \phi'_{i,j} (p + \Delta + 1) \right] \\
&= \left[(d + \Delta + 1) \sum_{i \in I^{(r)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right] + \left[(p + \Delta + 1) \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \sum_{j \in J_+} \lambda_i v_j^2 \phi'_{i,j} \right] \\
&\leq \left[(d + \Delta + 1) \cdot \frac{1}{(1 - 2c')(d - \Delta + 1)} \right] + \left[(p + \Delta + 1) k \cdot \frac{1}{(1 - 2c')(d - \Delta + 1)} \right]\ ,
\end{aligned}$$

where in the last inequality we used Lemma A.10. Plugging in $k = c' \cdot \frac{d - \Delta + 1}{p + \Delta + 1}$, the above equals

$$\begin{aligned}
&\left[\frac{d + \Delta + 1}{(1 - 2c')(d - \Delta + 1)} \right] + \left[c' \cdot \frac{d - \Delta + 1}{p + \Delta + 1} \cdot \frac{p + \Delta + 1}{(1 - 2c')(d - \Delta + 1)} \right] \\
&= \frac{d + \Delta + 1}{(1 - 2c')(d - \Delta + 1)} + \frac{c'}{(1 - 2c')}\ .
\end{aligned}$$

Using $c' \leq \frac{1}{10}$, the above is at most

$$\frac{5(d + \Delta + 1)}{4(d - \Delta + 1)} + \frac{1}{8} \leq \frac{5(d + \Delta)}{4(d - \Delta)} + \frac{1}{8}\ .$$

By Lemma A.9, the displayed equation is at most

$$\frac{5(22d/21)}{4(20d/21)} + \frac{1}{8} = \frac{5 \cdot 22}{4 \cdot 20} + \frac{1}{8} = \frac{3}{2} \leq 2\ .$$

□

Next, we show that the inputs to the neurons are not too negative for nice test examples.

Lemma D.3. *Suppose $\mathcal{S}$ satisfies Properties (P1) through (P6). Let $r \in Q$ and let $\mathbf{x} = \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}$ such that for all $i \in I$ we have $|\langle \mathbf{x}_i, \boldsymbol{\xi} \rangle| \leq \Delta$ and $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \leq \Delta$. Also, assume that $\langle \mathbf{x}_i, \mathbf{x} \rangle \in [d - \Delta, d + \Delta]$ for all $i \in I^{(r)}$ and $|\langle \mathbf{x}_i, \mathbf{x} \rangle| \leq p + \Delta$ for all $i \notin I^{(r)}$. Then, for all $j \in J$ we have*

$$\mathbf{w}_j^\top \mathbf{x} + b_j \geq -\sum_{i \in I} \lambda_i |v_j| \phi'_{i,j} (2\Delta + p + 1) \ .$$

Proof. Suppose towards contradiction that there exists $j \in J$ such that

$$\mathbf{w}_j^\top \mathbf{x} + b_j < -\sum_{i \in I} \lambda_i |v_j| \phi'_{i,j} (2\Delta + p + 1) \leq -\sum_{i \in I} 2\lambda_i |v_j| \phi'_{i,j} \Delta \ . \tag{21}$$

Suppose first that $y^{(r)} v_j < 0$. By Eq. (6) for all $i' \in I^{(r)}$ we have

$$\begin{aligned}
\mathbf{w}_j^\top \mathbf{x}_{i'} + b_j &= \mathbf{w}_j^\top \mathbf{x} + b_j + \mathbf{w}_j^\top (\mathbf{x}_{i'} - \mathbf{x}) \\
&= \mathbf{w}_j^\top \mathbf{x} + b_j + \mathbf{w}_j^\top (\boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}_{i'} - \boldsymbol{\mu}^{(r)} - \boldsymbol{\xi}) \\
&< -\sum_{i \in I} 2\lambda_i |v_j| \phi'_{i,j} \Delta + \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top (\boldsymbol{\xi}_{i'} - \boldsymbol{\xi}) \\
&= -\sum_{i \in I} 2\lambda_i |v_j| \phi'_{i,j} \Delta - \left(\sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top \boldsymbol{\xi} \right) + \left(\sum_{i \in I \setminus \{i'\}} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top \boldsymbol{\xi}_{i'} \right) + \lambda_{i'} y_{i'} v_j \phi'_{i',j} \mathbf{x}_{i'}^\top \boldsymbol{\xi}_{i'} \\
&\leq -\sum_{i \in I} 2\lambda_i |v_j| \phi'_{i,j} \Delta + \left(\sum_{i \in I} \lambda_i |v_j| \phi'_{i,j} \Delta \right) + \left(\sum_{i \in I \setminus \{i'\}} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top \boldsymbol{\xi}_{i'} \right) + \lambda_{i'} y^{(r)} v_j \phi'_{i',j} \mathbf{x}_{i'}^\top \boldsymbol{\xi}_{i'} \ .
\end{aligned} \tag{22}$$

Recall that by our assumption on $\mathcal{S}$, for every $i \in I$ and $q \in Q$ we have $|\langle \boldsymbol{\mu}^{(q)}, \boldsymbol{\xi}_i \rangle| \leq \sigma \sqrt{d} \ln(d)$, and for every $s \neq s'$ in I we have $|\langle \boldsymbol{\xi}_s, \boldsymbol{\xi}_{s'} \rangle| \leq \sigma^2 \sqrt{2d} \ln(d)$. Thus, for $q \in Q$ and $i \in I^{(q)}$ such that $i \neq i'$ we have

$$|\mathbf{x}_i^\top \boldsymbol{\xi}_{i'}| \leq |(\boldsymbol{\mu}^{(q)})^\top \boldsymbol{\xi}_{i'}| + |\boldsymbol{\xi}_i^\top \boldsymbol{\xi}_{i'}| \leq \sigma \sqrt{d} \ln(d) + \sigma^2 \sqrt{2d} \ln(d) \leq \Delta \ .$$

Moreover,

$$\mathbf{x}_{i'}^\top \boldsymbol{\xi}_{i'} = (\boldsymbol{\mu}^{(r)})^\top \boldsymbol{\xi}_{i'} + \boldsymbol{\xi}_{i'}^\top \boldsymbol{\xi}_{i'} \geq (\boldsymbol{\mu}^{(r)})^\top \boldsymbol{\xi}_{i'} \geq -\sigma \sqrt{d} \ln(d) \geq -\Delta \ .$$

Using the above displayed equations and the assumption $y^{(r)} v_j < 0$, the RHS in Eq. (22) is at most

$$\begin{aligned}
-\sum_{i \in I} & 2\lambda_i |v_j| \phi'_{i,j} \Delta + \left(\sum_{i \in I} \lambda_i |v_j| \phi'_{i,j} \Delta \right) + \left(\sum_{i \in I \setminus \{i'\}} \lambda_i |v_j| \phi'_{i,j} \Delta \right) + \lambda_{i'} |v_j| \phi'_{i',j} \Delta \\
&= -\sum_{i \in I} 2\lambda_i |v_j| \phi'_{i,j} \Delta + 2 \left(\sum_{i \in I} \lambda_i |v_j| \phi'_{i,j} \Delta \right) \\
&= 0 \ .
\end{aligned}$$

Hence, $y^{(r)} v_j < 0$ implies that $\phi'_{i',j} = \mathbb{1}[\mathbf{w}_j^\top \mathbf{x}_{i'} + b_j \geq 0] = 0$ for all $i' \in I^{(r)}$.

By Eq. (6) and (7) we have

$$\begin{aligned}
\mathbf{w}_j^\top \mathbf{x} + b_j &= \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top \mathbf{x} + \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \\
&= \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \\
&= \left[\sum_{i \in I^{(r)}} \lambda_i y^{(r)} v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right] + \left[\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \right] \ .
\end{aligned} \tag{23}$$

We now analyze both terms in the above RHS.

Note that if $y^{(r)} v_j \geq 0$ then

$$\sum_{i \in I^{(r)}} \lambda_i y^{(r)} v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) \geq \sum_{i \in I^{(r)}} \lambda_i y^{(r)} v_j \phi'_{i,j} (d - \Delta + 1) \geq 0 ,$$

and if $y^{(r)} v_j < 0$ then we have

$$\sum_{i \in I^{(r)}} \lambda_i y^{(r)} v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) = 0$$

since $\phi'_{i',j} = 0$ for all $i' \in I^{(r)}$. Moreover,

$$\begin{aligned}\sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i y_i v_j \phi'_{i,j} (\mathbf{x}_i^\top \mathbf{x} + 1) &\geq - \sum_{q \in Q \setminus \{r\}} \sum_{i \in I^{(q)}} \lambda_i |v_j| \phi'_{i,j} (p + \Delta + 1) \\ &\geq - \sum_{i \in I} \lambda_i |v_j| \phi'_{i,j} (p + \Delta + 1) .\end{aligned}$$

Plugging the above equations into Eq. (23) we get

$$\mathbf{w}_j^\top \mathbf{x} + b_j \geq - \sum_{i \in I} \lambda_i |v_j| \phi'_{i,j} (p + \Delta + 1) \geq - \sum_{i \in I} \lambda_i |v_j| \phi'_{i,j} (p + 2\Delta + 1) ,$$

in contradiction to Eq. (21). □

We now obtain a lower bound for the rate that perturbations in the direction $\mathbf{u} = \sum_{r \in Q} y^{(r)} \boldsymbol{\mu}^{(r)}$ change the inputs to the neurons.

Lemma D.4. *Suppose $\mathcal{S}$ satisfies Properties (P1) through (P6). Let $\mathbf{u} = \sum_{r \in Q} y^{(r)} \boldsymbol{\mu}^{(r)}$. For every $j \in J_+$ we have*

$$\mathbf{w}_j^\top \mathbf{u} \geq \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) .$$

For every $j \in J_-$ we have

$$\mathbf{w}_j^\top \mathbf{u} \leq \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) .$$

Proof. Let $j \in J$. Using Eq. (6) we have

$$\begin{aligned}\mathbf{w}_j^\top \sum_{r \in Q} y^{(r)} \boldsymbol{\mu}^{(r)} &= \sum_{i \in I} \lambda_i y_i v_j \phi'_{i,j} \mathbf{x}_i^\top \sum_{r \in Q} y^{(r)} \boldsymbol{\mu}^{(r)} \\ &= \sum_{q \in Q} \sum_{i \in I^{(q)}} \lambda_i y^{(q)} v_j \phi'_{i,j} \mathbf{x}_i^\top \left(y^{(q)} \boldsymbol{\mu}^{(q)} + \sum_{r \in Q \setminus \{q\}} y^{(r)} \boldsymbol{\mu}^{(r)} \right) \\ &= \sum_{q \in Q} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} \left((y^{(q)})^2 \mathbf{x}_i^\top \boldsymbol{\mu}^{(q)} + \sum_{r \in Q \setminus \{q\}} y^{(q)} y^{(r)} \mathbf{x}_i^\top \boldsymbol{\mu}^{(r)} \right) \\ &= \sum_{q \in Q} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} \left[(\boldsymbol{\mu}^{(q)})^\top \boldsymbol{\mu}^{(q)} + \boldsymbol{\xi}_i^\top \boldsymbol{\mu}^{(q)} + \sum_{r \in Q \setminus \{q\}} \left(y^{(q)} y^{(r)} (\boldsymbol{\mu}^{(q)})^\top \boldsymbol{\mu}^{(r)} + y^{(q)} y^{(r)} \boldsymbol{\xi}_i^\top \boldsymbol{\mu}^{(r)} \right) \right] .\end{aligned} \tag{24}$$

For $j \in J_+$ the above is at least

$$\sum_{q \in Q} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - \sum_{r \in Q \setminus \{q\}} (p + \Delta) \right) \geq \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - k(p + \Delta) \right) .$$

Similarly, for $j \in J_-$, Eq. (24) is at most

$$\sum_{q \in Q} \sum_{i \in I^{(q)}} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - \sum_{r \in Q \setminus \{q\}} (p + \Delta) \right) \leq \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - k(p + \Delta) \right) \ .$$

□

We next show that the quantity $d - \Delta - kp - k\Delta$ appearing in the lemma above is strictly positive.

Lemma D.5. *We have $d - \Delta - k(p + \Delta) > 0$.*

Proof. By Assumption 2.2 and Lemma A.9, we have

$$\begin{aligned} k(p+\Delta) < k(p+\Delta+1) &\leq \frac{1}{10} \cdot (d - \Delta + 1) = (d - \Delta) - \left[\frac{9}{10}(d - \Delta) - \frac{1}{10} \right] \\ &\leq (d - \Delta) - \left[\frac{9}{10} \cdot \frac{20d}{21} - \frac{1}{10} \right] \leq (d - \Delta) - \left[\frac{18 \cdot 1}{21} - \frac{1}{10} \right] < d - \Delta \ . \end{aligned}$$

□

Using the above lemmas, we show that a small perturbation to nice inputs suffices for obtaining positive inputs to all hidden neurons.

Lemma D.6. *Suppose $\mathcal{S}$ satisfies Properties (P1) through (P6). Let $\mathbf{z} = \eta \sum_{q \in Q} y^{(q)} \boldsymbol{\mu}^{(q)}$ for $\eta \geq \frac{2\Delta + p + 1}{d - \Delta - kp - k\Delta}$. Let $r \in Q$ and let $\mathbf{x} = \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}$ such that for all $i \in I$ we have $|\langle \mathbf{x}_i, \boldsymbol{\xi} \rangle| \leq \Delta$ and $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \leq \Delta$. Also, assume that $\langle \mathbf{x}_i, \mathbf{x} \rangle \in [d - \Delta, d + \Delta]$ for all $i \in I^{(r)}$ and $|\langle \mathbf{x}_i, \mathbf{x} \rangle| \leq p + \Delta$ for all $i \notin I^{(r)}$. Then, we have for all $j \in J_-$ that $\mathbf{w}_j^\top (\mathbf{x} - \mathbf{z}) + b_j \geq 0$, and for all $j \in J_+$ that $\mathbf{w}_j^\top (\mathbf{x} + \mathbf{z}) + b_j \geq 0$.*

Proof. Let $j \in J_-$. By Lemma D.3, we have

$$\mathbf{w}_j^\top \mathbf{x} + b_j \geq \sum_{i \in I} \lambda_i v_j \phi'_{i,j} (2\Delta + p + 1) \ .$$

By Lemma D.4, we have

$$-\mathbf{w}_j^\top \mathbf{z} \geq -\eta \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) \ .$$

Combining the last two displayed equations, we get

$$\begin{aligned} \mathbf{w}_j^\top (\mathbf{x} - \mathbf{z}) + b_j &= \mathbf{w}_j^\top \mathbf{x} + b_j - \mathbf{w}_j^\top \mathbf{z} \\ &\geq \sum_{i \in I} \lambda_i v_j \phi'_{i,j} (2\Delta + p + 1) - \eta \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) \\ &= \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(2\Delta + p + 1 - \eta (d - \Delta - kp - k\Delta) \right) \ . \end{aligned}$$

Note that by Lemma D.5 we have $d - \Delta - kp - k\Delta > 0$. Hence, for $\eta \geq \frac{2\Delta + p + 1}{d - \Delta - kp - k\Delta}$ we have $\mathbf{w}_j^\top (\mathbf{x} - \mathbf{z}) + b_j \geq 0$.

The proof for $j \in J_+$ is similar. Namely, by Lemmas D.3 and D.4, we have

$$\begin{aligned} \mathbf{w}_j^\top (\mathbf{x} + \mathbf{z}) + b_j &= \mathbf{w}_j^\top \mathbf{x} + b_j + \mathbf{w}_j^\top \mathbf{z} \\ &\geq - \sum_{i \in I} \lambda_i v_j \phi'_{i,j} (2\Delta + p + 1) + \eta \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) \\ &= \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(-2\Delta - p - 1 + \eta (d - \Delta - kp - k\Delta) \right) \ , \end{aligned}$$

and hence for $\eta \geq \frac{2\Delta + p + 1}{d - \Delta - kp - k\Delta}$ we get $\mathbf{w}_j^\top (\mathbf{x} + \mathbf{z}) + b_j \geq 0$. □

Let now $\eta_1 = \frac{2\Delta+p+1}{d-\Delta-kp-k\Delta}$ and $\eta_2 = \frac{3(3d+\Delta+1)(1-2c')}{(d-\Delta-kp-k\Delta)(1-3c')ck}$. Note that by Lemma D.5 both η_1 and η_2 are positive. We denote $\mathbf{z} = (\eta_1 + \eta_2) \sum_{q \in Q} y^{(q)} \boldsymbol{\mu}^{(q)}$.

In the next lemma, we show that perturbing nice test examples from positive clusters with the vector $-\mathbf{z}$ changes the sign of the network output.

Lemma D.7. *Suppose $\mathcal{S}$ satisfies Properties (P1) through (P6). Let $r \in Q_+$ and let $\mathbf{x} = \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}$ such that for all $i \in I$ we have $|\langle \mathbf{x}_i, \boldsymbol{\xi} \rangle| \leq \Delta$ and $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi} \rangle| \leq \Delta$. Also, assume that $\langle \mathbf{x}_i, \mathbf{x} \rangle \in [d - \Delta, d + \Delta]$ for all $i \in I^{(r)}$ and $|\langle \mathbf{x}_i, \mathbf{x} \rangle| \leq p + \Delta$ for all $i \notin I^{(r)}$. Then, $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x} - \mathbf{z}) \leq -1$.*

Proof. We denote $\mathbf{x}' = \mathbf{x} - \mathbf{z}$. By Lemma D.4, for every $j \in J_+$ we have

$$\begin{aligned}
\mathbf{w}_j^\top \mathbf{x}' + b_j &= \mathbf{w}_j^\top \mathbf{x} + b_j - \mathbf{w}_j^\top (\eta_1 + \eta_2) \sum_{q \in Q} y^{(q)} \boldsymbol{\mu}^{(q)} \\
&\leq \mathbf{w}_j^\top \mathbf{x} + b_j - (\eta_1 + \eta_2) \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) \ .
\end{aligned}$$

By Lemma D.5 we get

$$\mathbf{w}_j^\top \mathbf{x}' + b_j \leq \mathbf{w}_j^\top \mathbf{x} + b_j. \tag{25}$$

Thus, in the neurons J_+ the input does not increase when moving from $\mathbf{x}$ to $\mathbf{x}'$.

Consider now $j \in J_-$. Let $\tilde{\mathbf{x}} = \mathbf{x} - \eta_1 \sum_{q \in Q} y^{(q)} \boldsymbol{\mu}^{(q)}$. By Lemma D.6, we have $\mathbf{w}_j^\top \tilde{\mathbf{x}} + b_j \geq 0$. Also, by Lemma D.4, we have

$$\begin{aligned}
\mathbf{w}_j^\top \tilde{\mathbf{x}} + b_j &= \mathbf{w}_j^\top \mathbf{x} + b_j - \mathbf{w}_j^\top \eta_1 \sum_{q \in Q} y^{(q)} \boldsymbol{\mu}^{(q)} \\
&\geq \mathbf{w}_j^\top \mathbf{x} + b_j - \eta_1 \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) \ ,
\end{aligned}$$

and by Lemma D.5 the above is at least $\mathbf{w}_j^\top \mathbf{x} + b_j$. Thus, when moving from $\mathbf{x}$ to $\tilde{\mathbf{x}}$ the input to the neurons J_- can only increase, and at $\tilde{\mathbf{x}}$ it is non-negative.

Next, we move from $\tilde{\mathbf{x}}$ to $\mathbf{x}'$. We have

$$\mathbf{w}_j^\top \mathbf{x}' + b_j = \mathbf{w}_j^\top \tilde{\mathbf{x}} + b_j - \eta_2 \mathbf{w}_j^\top \sum_{q \in Q} y^{(q)} \boldsymbol{\mu}^{(q)} \geq \max\left\{0, \mathbf{w}_j^\top \mathbf{x} + b_j\right\} - \eta_2 \mathbf{w}_j^\top \sum_{q \in Q} y^{(q)} \boldsymbol{\mu}^{(q)} \ .$$

By Lemma D.4, the above is at least

$$\max\left\{0, \mathbf{w}_j^\top \mathbf{x} + b_j\right\} - \eta_2 \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) \geq 0 \ , \tag{26}$$

where in the last inequality we use Lemma D.5.

Overall, we have

$$\begin{aligned}
\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}') &= \left[\sum_{j \in J_+} v_j \phi(\mathbf{w}_j^\top \mathbf{x}' + b_j) \right] + \left[\sum_{j \in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x}' + b_j) \right] \\
&\stackrel{(i)}{=} \left[\sum_{j \in J_+} v_j \phi(\mathbf{w}_j^\top \mathbf{x}' + b_j) \right] + \left[\sum_{j \in J_-} v_j (\mathbf{w}_j^\top \mathbf{x}' + b_j) \right] \\
&\stackrel{(ii)}{\leq} \left[\sum_{j \in J_+} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j) \right] + \\
&\quad \left[\sum_{j \in J_-} v_j \left(\max\left\{0, \mathbf{w}_j^\top \mathbf{x} + b_j\right\} - \eta_2 \sum_{i \in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta \right) \right) \right] \ ,
\end{aligned}$$

where in (i) we used Eq. (26), and in (ii) we used both Eq. (25) and Eq. (26). Now, the above equals

$$\left[\sum_{j\in J} v_j\phi(\mathbf{w}_j^\top \mathbf{x} + b_j)\right] - \left[\sum_{j\in J_-} v_j\eta_2 \sum_{i\in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta\right)\right]$$
$$= \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) - \eta_2 \left(d - \Delta - kp - k\Delta\right) \left[\sum_{q'\in Q} \sum_{i\in I^{(q')}} \sum_{j\in J_-} \lambda_i v_j^2 \phi'_{i,j}\right] .$$

Combining the above with Lemma D.2, Lemma A.11, and Lemma D.5, we get

$$\begin{aligned}\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}') &\leq 2 - \eta_2 \left(d - \Delta - kp - k\Delta\right) \left[\sum_{q'\in Q_-} \sum_{i\in I^{(q')}} \sum_{j\in J_-} \lambda_i v_j^2 \phi'_{i,j}\right] \\ &\leq 2 - \eta_2 \left(d - \Delta - kp - k\Delta\right) |Q_-| \left(1 - \frac{c'}{1-2c'}\right) \frac{1}{3d + \Delta + 1} \\ &\leq 2 - \eta_2 \left(d - \Delta - kp - k\Delta\right) ck \left(\frac{1-3c'}{1-2c'}\right) \frac{1}{3d + \Delta + 1} .\end{aligned}$$

For

$$\eta_2 = \frac{3(3d + \Delta + 1)(1 - 2c')}{(d - \Delta - kp - k\Delta)(1 - 3c')ck}$$

we conclude that $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}')$ is at most -1. □

Next, we show that perturbing nice test examples from negative clusters with the vector $\mathbf{z}$ changes the sign of the network output.

Lemma D.8. *Suppose $\mathcal{S}$ satisfies Properties (P1) through (P6). Let $r \in Q_-$ and let $\mathbf{x} = \boldsymbol{\mu}^{(r)} + \boldsymbol{\xi}$ such that for all $i \in I$ we have $|\langle \mathbf{x}_i, \boldsymbol{\xi}\rangle| \leq \Delta$ and $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}\rangle| \leq \Delta$. Also, assume that $\langle \mathbf{x}_i, \mathbf{x}\rangle \in [d - \Delta, d + \Delta]$ for all $i \in I^{(r)}$ and $|\langle \mathbf{x}_i, \mathbf{x}\rangle| \leq p + \Delta$ for all $i \notin I^{(r)}$. Then, $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x} + \mathbf{z}) \geq 1$.*

Proof. The proof follows similar arguments to the proof of Lemma D.7. We provide it here for completeness.

We denote $\mathbf{x}' = \mathbf{x} + \mathbf{z}$. By Lemma D.4, for every $j \in J_-$ we have

$$\begin{aligned}\mathbf{w}_j^\top \mathbf{x}' + b_j &= \mathbf{w}_j^\top \mathbf{x} + b_j + \mathbf{w}_j^\top (\eta_1 + \eta_2) \sum_{q\in Q} y^{(q)} \boldsymbol{\mu}^{(q)} \\ &\leq \mathbf{w}_j^\top \mathbf{x} + b_j + (\eta_1 + \eta_2) \sum_{i\in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta\right) .\end{aligned}$$

By Lemma D.5 we get

$$\mathbf{w}_j^\top \mathbf{x}' + b_j \leq \mathbf{w}_j^\top \mathbf{x} + b_j. \tag{27}$$

Thus, in the neurons J_- the input does not increase when moving from $\mathbf{x}$ to $\mathbf{x}'$.

Consider now $j \in J_+$. Let $\tilde{\mathbf{x}} = \mathbf{x} + \eta_1 \sum_{q\in Q} y^{(q)} \boldsymbol{\mu}^{(q)}$. By Lemma D.6, we have $\mathbf{w}_j^\top \tilde{\mathbf{x}} + b_j \geq 0$. Also, by Lemma D.4, we have

$$\begin{aligned}\mathbf{w}_j^\top \tilde{\mathbf{x}} + b_j &= \mathbf{w}_j^\top \mathbf{x} + b_j + \mathbf{w}_j^\top \eta_1 \sum_{q\in Q} y^{(q)} \boldsymbol{\mu}^{(q)} \\ &\geq \mathbf{w}_j^\top \mathbf{x} + b_j + \eta_1 \sum_{i\in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta\right) ,\end{aligned}$$

and by Lemma D.5 the above is at least $\mathbf{w}_j^\top \mathbf{x} + b_j$. Thus, when moving from $\mathbf{x}$ to $\tilde{\mathbf{x}}$ the input to the neurons J_+ can only increase, and at $\tilde{\mathbf{x}}$ it is non-negative.

Next, we move from $\tilde{\mathbf{x}}$ to $\mathbf{x}'$. We have

$$\mathbf{w}_j^\top \mathbf{x}' + b_j = \mathbf{w}_j^\top \tilde{\mathbf{x}} + b_j + \eta_2 \mathbf{w}_j^\top \sum_{q\in Q} y^{(q)} \boldsymbol{\mu}^{(q)} \geq \max\left\{0, \mathbf{w}_j^\top \mathbf{x} + b_j\right\} + \eta_2 \mathbf{w}_j^\top \sum_{q\in Q} y^{(q)} \boldsymbol{\mu}^{(q)} .$$

By Lemma D.4, the above is at least

$$\max\left\{0, \mathbf{w}_j^\top \mathbf{x} + b_j\right\} + \eta_2 \sum_{i\in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta\right) \geq 0 \,, \tag{28}$$

where in the last inequality we use Lemma D.5.

Overall, we have

$$\begin{aligned}
\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}') &= \left[\sum_{j\in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x}' + b_j)\right] + \left[\sum_{j\in J_+} v_j \phi(\mathbf{w}_j^\top \mathbf{x}' + b_j)\right] \\
&\overset{(i)}{=} \left[\sum_{j\in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x}' + b_j)\right] + \left[\sum_{j\in J_+} v_j (\mathbf{w}_j^\top \mathbf{x}' + b_j)\right] \\
&\overset{(ii)}{\geq} \left[\sum_{j\in J_-} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j)\right] + \\
&\quad \left[\sum_{j\in J_+} v_j \left(\max\left\{0, \mathbf{w}_j^\top \mathbf{x} + b_j\right\} + \eta_2 \sum_{i\in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta\right)\right)\right] \,,
\end{aligned}$$

where in (i) we used Eq. (28), and in (ii) we used both Eq. (27) and Eq. (28). Now, the above equals

$$\begin{aligned}
&\left[\sum_{j\in J} v_j \phi(\mathbf{w}_j^\top \mathbf{x} + b_j)\right] + \left[\sum_{j\in J_+} v_j \eta_2 \sum_{i\in I} \lambda_i v_j \phi'_{i,j} \left(d - \Delta - kp - k\Delta\right)\right] \\
&= \mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}) + \eta_2 \left(d - \Delta - kp - k\Delta\right) \left[\sum_{q'\in Q} \sum_{i\in I^{(q')}} \sum_{j\in J_+} \lambda_i v_j^2 \phi'_{i,j}\right] \,.
\end{aligned}$$

Combining the above with Lemma D.2, Lemma A.11, and Lemma D.5, we get

$$\begin{aligned}
\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}') &\geq -2 + \eta_2 \left(d - \Delta - kp - k\Delta\right) \left[\sum_{q'\in Q_+} \sum_{i\in I^{(q')}} \sum_{j\in J_+} \lambda_i v_j^2 \phi'_{i,j}\right] \\
&\geq -2 + \eta_2 \left(d - \Delta - kp - k\Delta\right) |Q_+| \left(1 - \frac{c'}{1-2c'}\right) \frac{1}{3d + \Delta + 1} \\
&\geq -2 + \eta_2 \left(d - \Delta - kp - k\Delta\right) ck \left(\frac{1-3c'}{1-2c'}\right) \frac{1}{3d + \Delta + 1} \,.
\end{aligned}$$

Plugging-in η_2, we conclude that $\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{x}')$ is at least 1. □

Finally, we show that the scale of the perturbation $\mathbf{z}$ is small when k is large.

Lemma D.9. *We have* $\|\mathbf{z}\| = \mathcal{O}\left(\sqrt{\frac{d}{c^2 k}}\right)$.

Proof. We have

$$\|\mathbf{z}\|^2 = (\eta_1 + \eta_2)^2 \left\|\sum_{q\in Q} y^{(q)} \boldsymbol{\mu}^{(q)}\right\|^2$$

Now,

$$
\begin{aligned}
(\eta_1+\eta_2)^2 &= \left(\frac{2\Delta+p+1}{d-\Delta-kp-k\Delta} + \frac{3(3d+\Delta+1)(1-2c')}{(d-\Delta-kp-k\Delta)(1-3c')ck}\right)^2 \\
&\leq \left(\frac{2(p+\Delta+1)}{d-\Delta-k(p+\Delta+1)} + \frac{3(3d+\Delta+1)(1-2c')}{(d-\Delta-k(p+\Delta+1))(1-3c')ck}\right)^2 \\
&\overset{(i)}{=} \left(\frac{2c'(d-\Delta+1)/k}{d-\Delta-c'(d-\Delta+1)} + \frac{3(3d+\Delta+1)(1-2c')}{(d-\Delta-c'(d-\Delta+1))(1-3c')ck}\right)^2 \\
&\overset{(ii)}{\leq} \left(\frac{1}{k}\cdot\frac{\frac{2}{10}(d+1)}{d-\frac{d}{21}-\frac{1}{10}(d+1)} + \frac{1}{ck}\cdot\frac{3(3d+\frac{d}{21}+1)(1-\frac{2}{10})}{(d-\frac{d}{21}-\frac{1}{10}(d+1))(1-\frac{3}{10})}\right)^2 \\
&\leq \left(\frac{1}{k}\cdot\frac{\mathcal{O}(d)}{\Omega(d)} + \frac{1}{ck}\cdot\frac{\mathcal{O}(d)}{\Omega(d)}\right)^2 \\
&\leq \mathcal{O}\left(\frac{1}{c^2k^2}\right) ,
\end{aligned}
$$

where in (i) we used $k = c' \cdot \frac{d-\Delta+1}{p+\Delta+1}$, and in (ii) we used Lemma A.9.

Moreover,

$$
\begin{aligned}
\left\|\sum_{q\in Q} y^{(q)}\boldsymbol{\mu}^{(q)}\right\|^2 &= \sum_{r\in Q}\sum_{q\in Q} y^{(r)}y^{(q)}\langle\boldsymbol{\mu}^{(r)},\boldsymbol{\mu}^{(q)}\rangle \\
&= \sum_{r\in Q}\left[\left\|\boldsymbol{\mu}^{(r)}\right\|^2 + \sum_{q\neq r} y^{(r)}y^{(q)}\langle\boldsymbol{\mu}^{(r)},\boldsymbol{\mu}^{(q)}\rangle\right] \\
&\leq kd + k^2p \\
&= kd + k^2\left(\frac{c'(d-\Delta+1)}{k} - \Delta - 1\right) \\
&\leq kd + kc'(d-\Delta+1) \\
&\leq \mathcal{O}(kd) .
\end{aligned}
$$

Overall, we get

$$
\|\mathbf{z}\|^2 \leq \mathcal{O}\left(\frac{d}{c^2k}\right) .
$$

□

The theorem now follows immediately from Lemmas D.1, D.7, D.8, and D.9.