Proceedings of Machine Learning Research vol 195:1–56, 2023 36th Annual Conference on Learning Theory

Benign Overfitting in Linear Classifiers and Leaky ReLU Networks from KKT Conditions for Margin Maximization

Spencer Frei* FREI@BERKELEY.EDU
UC Berkeley

Gal Vardi* GALVARDI@TTIC.EDU
TTI-Chicago and Hebrew University

Peter L. Bartlett PETER@BERKELEY.EDU
UC Berkeley and Google DeepMind

Nathan Srebro NATI@TTIC.EDU
TTI-Chicago

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Linear classifiers and leaky ReLU networks trained by gradient flow on the logistic loss have an implicit bias towards solutions which satisfy the Karush–Kuhn–Tucker (KKT) conditions for margin maximization. In this work we establish a number of settings where the satisfaction of these KKT conditions implies benign overfitting in linear classifiers and in two-layer leaky ReLU networks: the estimators interpolate noisy training data and simultaneously generalize well to test data. The settings include variants of the noisy class-conditional Gaussians considered in previous work as well as new distributional settings where benign overfitting has not been previously observed. The key ingredient to our proof is the observation that when the training data is nearly-orthogonal, both linear classifiers and leaky ReLU networks satisfying the KKT conditions for their respective margin maximization problems behave like a weighted average of the training examples.

Keywords: Benign overfitting, Linear classifiers, Leaky ReLU networks, Implicit bias.

1. Introduction

The phenomenon of 'benign overfitting'—referring to settings where a model achieves a perfect fit to noisy training data and still generalizes well to unseen data—has attracted significant attention in recent years. Following the initial experiments of Zhang et al. (2017), researchers have sought to understand how this phenomenon can occur despite the long-standing intuition from statistical learning theory that overfitting to noise should result in poor out-of-sample prediction performance.

In this work, we provide several new results on benign overfitting in classification tasks, for both linear classifiers and two-layer leaky-ReLU neural networks. We consider gradient flow on the empirical risk with exponentially-tailed loss functions, such as the logistic loss. Under certain assumptions on the data distribution, we prove that gradient flow converges to solutions that exhibit benign overfitting: the predictors interpolate noisy training data and simultaneously generalize well to unseen test data. Our results extend existing work in two aspects: First, we prove benign overfitting in two-layer leaky ReLU networks, while existing results do not cover such models.[1]

* Equal contribution

1. Frei et al. (2022) showed benign overfitting in two-layer nets with *smooth* leaky ReLU activations, as we discuss later.

Second, we characterize benign overfitting in new distributional settings (i.e., assumptions on the data distributions).

The first distributional setting we consider is a noisy sub-Gaussian distribution $(x, y) \sim \mathsf{P}_{\mathsf{sg}}$ where labels are generated by a single component of x and are flipped to the opposite sign with probability η. We show that if the variance from this component is sufficiently large relative to the variance of the other components, and if the covariance matrix has a sufficiently high rank relative to the number of samples, then linear classifiers and leaky ReLU networks trained by gradient flow exhibit benign overfitting. In our second distributional setting, we consider a distribution $\mathsf{P}_{\mathsf{clust}}$ where inputs x are drawn uniformly from k nearly-orthogonal clusters, and labels are determined by the cluster and are flipped to the opposite sign with probability η. We show that under some assumptions on the scale and correlation of the clusters, gradient flow on linear classifiers and leaky ReLU networks produces classifiers which exhibit benign overfitting. This is a setting not covered by prior work on benign overfitting, and it essentially generalizes some previous results on benign overfitting in linear classification (Chatterji and Long, 2021; Wang and Thrampoulidis, 2021) and neural networks with *smooth* leaky activations (Frei et al., 2022).

Our proofs follow by analyzing the implicit bias of gradient flow. Lyu and Li (2020); Ji and Telgarsky (2020) showed that when training homogeneous neural networks with exponentially-tailed loss functions, gradient flow is biased towards solutions that maximize the margin in parameter space. Namely, if the empirical risk reaches a small enough value, then gradient flow converges in direction to a solution that satisfies the Karush–Kuhn–Tucker (KKT) conditions for the margin-maximization problem. We develop new proof techniques which show that in the aforementioned distributional settings, benign overfitting occurs for any solution that satisfies these KKT conditions. In a bit more detail, we show that every KKT point in our settings has a linear decision boundary, even in the case of leaky ReLU networks. This linear decision boundary can be expressed by a weighted sum of the training examples, where the weights of all examples are approximately balanced. Using this balancedness property, we are able to prove that benign overfitting occurs.

Related work

Benign overfitting. The benign overfitting phenomenon has recently attracted intense attention and was studied in various settings, such as linear regression (Hastie et al., 2020; Belkin et al., 2020; Bartlett et al., 2020; Muthukumar et al., 2020; Negrea et al., 2020; Chinot and Lerasle, 2020; Koehler et al., 2021; Wu and Xu, 2020; Tsigler and Bartlett, 2020; Zhou et al., 2022; Wang et al., 2022; Chatterji et al., 2021; Bartlett and Long, 2021; Shamir, 2022), kernel regression (Liang and Rakhlin, 2020; Mei and Montanari, 2019; Liang et al., 2020; Mallinar et al., 2022; Rakhlin and Zhai, 2019; Belkin et al., 2018), and classification (Chatterji and Long, 2021; Wang and Thrampoulidis, 2021; Cao et al., 2021; Muthukumar et al., 2021; Montanari et al., 2020; Shamir, 2022; Frei et al., 2022; Cao et al., 2022; McRae et al., 2022; Liang and Recht, 2021; Thrampoulidis et al., 2020; Wang et al., 2021; Donhauser et al., 2022). Below we discuss several works on benign overfitting in classification which are most relevant to our results.

In contrast to linear regression, in linear classification the solution to which gradient flow is known to converge, namely, the max-margin predictor, does not have a closed-form expression. Hence, analyzing benign overfitting in linear classification is more challenging. Chatterji and Long (2021); Wang and Thrampoulidis (2021) prove benign overfitting in linear classification for a high-dimensional sub-Gaussian mixture model. Our results imply as a special case benign overfitting in sub-Gaussian mixtures similar to their results. Cao et al. (2021) also study benign overfitting in

a sub-Gaussian mixture model, but they do not consider label flipping noise. Muthukumar et al. (2021) study the behavior of the overparameterized max-margin classifier in a discriminative classification model with label-flipping noise, by connecting the behavior of the max-margin classifier to the ordinary least squares solution. They show that under certain conditions, all training data points become support vectors of the maximum margin classifier (see also Hsu et al. (2021)). Montanari et al. (2020) studies a setting where the inputs are Gaussian, and the labels are generated according to a logistic link function. They derive an expression for the asymptotic prediction error of the max-margin linear classifier, assuming the ratio of the dimension and the sample size converges to some fixed positive limit. Shamir (2022) also studies linear classification and proves benign overfitting under a distributional setting which is different from the aforementioned works and from our setting.

Benign overfitting in nonlinear neural networks is even less well-understood. Frei et al. (2022) show benign overfitting in two-layer networks with *smooth* leaky ReLU activations for a high-dimensional sub-Gaussian mixture model; at the end of Section 5 we compare our results with theirs. Cao et al. (2022) study benign overfitting in training a two-layer *convolutional* neural network using the logistic loss, but they do not consider label-flipping noise as we do.

Implicit bias. The literature on implicit bias in neural networks has rapidly expanded in recent years (see Vardi (2022) for a survey). In what follows, we discuss results that apply either to linear classification using gradient flow, or to nonlinear two-layer networks trained with gradient flow in classification settings.

Soudry et al. (2018) showed that gradient descent on linearly-separable binary classification problems with exponentially-tailed losses (e.g., the exponential loss and the logistic loss), converges to the maximum ℓ_2-margin direction. This analysis was extended to other loss functions, tighter convergence rates, non-separable data, and variants of gradient-based optimization algorithms (Nacson et al., 2019a; Ji and Telgarsky, 2018; Ji et al., 2020; Gunasekar et al., 2018; Shamir, 2020; Ji and Telgarsky, 2021; Nacson et al., 2019b; Ji et al., 2021).

Lyu and Li (2020) and Ji and Telgarsky (2020) showed that homogeneous neural networks (and specifically two-layer leaky 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. We note that the aforementioned KKT point may not be a global optimum of the maximum-margin problem (Vardi et al., 2021; Lyu et al., 2021). Recently, Kunin et al. (2022) extended this result by showing bias towards margin maximization in a broader family of networks called *quasi-homogeneous*. Lyu et al. (2021); Sarussi et al. (2021); Frei et al. (2023) 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. Specifically, Frei et al. (2023) analyzed the implicit bias in leaky ReLU networks trained with nearly-orthogonal data, and our analysis of leaky ReLU networks builds on their result (see Section 3 for details). Moreover, implicit bias with nearly-orthogonal data was studied for ReLU networks in Vardi et al. (2022), where the authors prove bias towards networks that are not adversarially robust. Other works which consider the implicit bias of classification using gradient flow in nonlinear two-layer networks include Chizat and Bach (2020); Phuong and Lampert (2020); Safran et al. (2022); Timor et al. (2022).

2. Preliminaries

Notation. We use $\|x\|$ to denote the Euclidean norm of a vector x, while for matrices W we use $\|W\|_F$ to denote its Frobenius norm and $\|W\|_2$ its spectral norm. We use $\mathbb{1}(z)$ to denote the indicator function, so $\mathbb{1}(z) = 1$ if $z \geq 0$ and 0 otherwise. We use $\mathrm{sign}(z)$ as the function that is 1 when $z > 0$ and -1 otherwise. For integer $n \in \mathbb{N}$, we use $[n] = \{1, \ldots, n\}$. The Gaussian with mean a and variance σ^2 is denoted $\mathsf{N}(a, \sigma^2)$, while the multivariate Gaussian with mean μ and covariance matrix Σ is denoted $\mathsf{N}(\mu, \Sigma)$. We denote the minimum of two numbers a, b as $a \wedge b$, and the maximum $a \vee b$. For a vector $x \in \mathbb{R}^d$, we use $[x]_i \in \mathbb{R}$ to denote the i-th component of the vector, and $[x]_{i:j} \in \mathbb{R}^{j-i+1}$ as the vector with components $[x]_i, [x]_{i+1}, \ldots, [x]_j$. We use the standard big-Oh notation $O(\cdot), \Omega(\cdot)$ to hide universal constants, with $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ hiding logarithmic factors. We refer to quantities that are independent of the dimension d, number of samples n, the failure probability δ or number of neurons m in the network as constants.

The setting. We consider classification tasks where the training data $\{(x_i, y_i)\}_{i=1}^n$ are drawn i.i.d. from a distribution P over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$. We study two distinct models in this work. In the first, we consider maximum-margin *linear classifiers* $x \mapsto \mathrm{sign}(\langle w, x \rangle)$, which are solutions to the following constrained optimization problem:

$$\min_{w \in \mathbb{R}^d} \|w\|^2 \quad \text{such that for all } i \in [n], \quad y_i \langle w, x_i \rangle \geq 1. \tag{1}$$

By Soudry et al. (2018), gradient descent on exponentially-tailed losses such as the logistic loss has an implicit bias towards such solutions. We shall show that in a number of settings, any solution to Problem (1) will exhibit benign overfitting.

As our second model, we consider two-layer neural networks with leaky ReLU activations, where the first layer $W \in \mathbb{R}^{m \times d}$ is trained but the second layer weights $\{a_j\}_{j=1}^m$ fixed at random initialization:

$$f(x; W) := \textstyle\sum_{j=1}^m a_j \phi(\langle w_j, x \rangle), \quad \phi(q) = \max(\gamma q, q), \quad \gamma \in (0, 1). \tag{2}$$

For simplicity we assume m is an even number and that for half of the neurons, $a_j = 1/\sqrt{m}$, and the other half of the the neurons satisfy $a_j = -1/\sqrt{m}$. We consider a binary classification task with training data $S = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$. We define the *margin-maximization problem* for the neural network $f(x; W)$ over training data S as

$$\min_{W \in \mathbb{R}^{m \times d}} \|W\|_F^2 \quad \text{such that for all } i \in [n], \quad y_i f(x_i; W) \geq 1. \tag{3}$$

Recall the definition of the Karush–Kuhn–Tucker (KKT) conditions for non-smooth optimization problems (cf. Lyu and Li (2020); Dutta et al. (2013)). Let $h : \mathbb{R}^p \to \mathbb{R}$ be a locally Lipschitz function. The Clarke subdifferential (Clarke et al., 2008) at $\theta \in \mathbb{R}^p$ is the convex set

$$\partial^\circ h(\theta) := \mathrm{conv} \left\{ \lim_{s \to \infty} \nabla h(\theta_s) \;\middle|\; \lim_{s \to \infty} \theta_s = \theta, \; h \text{ is differentiable at } \theta_s \right\}.$$

If h is continuously differentiable at θ then $\partial^\circ h(\theta) = \{\nabla h(\theta)\}$. Given locally Lipschitz functions $h, g_1, \ldots, g_n : \mathbb{R}^p \to \mathbb{R}$, we say that $\theta \in \mathbb{R}^p$ is a *feasible point* of the problem

$$\min h(\theta) \quad \text{s.t.} \quad \text{for all } n \in [N], \; g_n(\theta) \leq 0,$$

if θ satisfies $g_n(\theta) \leq 0$ for all $n \in [N]$. We say that a feasible point θ is a *KKT point* if there exists $\lambda_1, \ldots, \lambda_N \geq 0$ such that

1. $0 \in \partial^\circ h(\theta) + \sum_{n\in[N]} \lambda_n \partial^\circ g_n(\theta)$;
2. For all $n \in [N]$ we have $\lambda_n g_n(\theta) = 0$.

We shall show that in a number of settings, any KKT point of Problem (3) will generalize well, even when a constant fraction of the training labels are uniformly random labels. Since any feasible point of Problem (3) interpolates the training data, this implies the network exhibits *benign overfitting*.

KKT points of Problem (3) appear naturally in the training of neural networks. For a loss function $\ell : \mathbb{R} \to [0, \infty)$ and for parameters W of the neural network $f(x; W)$, define the empirical risk under ℓ as

$$\widehat{L}(W) := \tfrac{1}{n} \textstyle\sum_{i=1}^n \ell\big(y_i f(x_i; W)\big).$$

Gradient flow for the objective function $\widehat{L}(W)$ is the trajectory $W(t)$ defined by an initial point $W(0)$, and is such that $W(t)$ satisfies the differential equation $\frac{\mathrm{d}}{\mathrm{d}t} W(t) \in -\partial^\circ \widehat{L}(W(t))$ with initial point $W(0)$. Since the network $f(x; \cdot)$ is 1-homogeneous, recent work by Lyu and Li (2020) and Ji and Telgarsky (2020) show that if ℓ is either the exponential loss $\ell(q) = \exp(-q)$ or logistic loss $\ell(q) = \log(1 + \exp(-q))$, then provided there exists a time t_0 for which $\widehat{L}(W(0)) < \log(2)/n$, gradient flow converges in direction to a KKT point of Problem (3), in the sense that for some KKT point W^* of Problem (3) it holds that $\frac{W(t)}{\|W(t)\|} \to \frac{W^*}{\|W^*\|}$. Thus, although there exist many neural networks which could classify the training data correctly, if gradient flow reaches a point with small enough loss then it will only produce networks which converge in direction to networks which satisfy the KKT conditions of Problem (3). Note that this need not imply that $W(t)$ converges in direction to a global optimum of Problem (3) (Vardi et al., 2021; Lyu et al., 2021). This is in contrast to the margin-maximization problem in linear classification given in Eq. (1), where the constraints and objective function are linear, and hence the KKT conditions are necessary and sufficient for global optimality.

3. Properties of KKT Points for Nearly Orthogonal Data

In this section we show that when the training data is nearly-orthogonal (in a sense to be formalized momentarily), then the decision boundaries of both (i) KKT points of the *linear* max-margin problem (1) and (ii) KKT points of the *nonlinear* leaky ReLU network (3) take the form of a weighted-average estimator $w = \sum_{i=1}^n s_i y_i x_i$ where $\{s_i\}_{i=1}^n$ are strictly positive and all of the same order, namely, w is a *nearly uniform average* of the training data. We will use this property in the next sections to show benign overfitting under certain distributional assumptions. We begin with our definitions of p-orthogonality and τ-uniform classifiers.

Definition 1 *Denote* $R_{min}^2 = \min_i \|x_i\|^2$, $R_{max}^2 = \max_i \|x_i\|^2$, *and* $R^2 = R_{max}^2/R_{min}^2$. *We call the training data* p-orthogonal *if* $R_{min}^2 \geq pR^2 n \max_{i\neq j} |\langle x_i, x_j\rangle|$.

Clearly, if the training data is exactly orthogonal then it is p-orthogonal for every $p > 0$. In contrast to exact orthogonality, p-orthogonality allows for the possibility that training data sampled i.i.d. from a broad class of distributions is p-orthogonal, as we shall see later.

Definition 2 *We say that* $w \in \mathbb{R}^d$ *is* τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{-1, 1\}$ *if* $w = \sum_{i=1}^n s_i y_i x_i$, *where the coefficients* $\{s_i\}_{i=1}^n$ *are strictly positive and* $\frac{\max_i s_i}{\min_i s_i} \leq \tau$.

Our first lemma shows that if the training data is p-orthogonal for large p and the norms of the training examples are all of the same order, then the linear max-margin classifier is given by a τ-uniform vector.

Proposition 3 *Suppose the training data are p-orthogonal for $p \geq 3$. Denote $R^2 = \max_{i,j} \|x_i\|^2/\|x_j\|^2$. Let $\widehat{w} = \operatorname{argmin}\{\|w\|^2 : y_i\langle w, x_i\rangle \geq 1 \,\forall i\}$ be the max-margin linear classifier. Then, $\widehat{w}$ is τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$ for $\tau = R^2\left(1 + \frac{2}{pR^2-2}\right)$.*

The proof for this proposition comes from an analysis of the KKT conditions for the max-margin problem and is provided in Appendix A. Observe that as $p \to \infty$ and as $R^2 \to 1$, we see that the linear max-margin becomes proportional to $\sum_{i=1}^n y_i x_i$, i.e. the classical sample average estimator.

Next, we show that when the training data are p-orthogonal for large enough p, then any KKT point of the leaky ReLU network margin maximization problem (3) has the same decision boundary as a τ-uniform linear classifier, despite the fact that two-layer leaky ReLU networks are in general nonlinear. The proof relies on a recent work by Frei et al. (2023), and is given in Appendix B.

Proposition 4 *Denote $R^2 = \max_{i,j} \|x_i\|^2/\|x_j\|^2$. Let f denote the leaky ReLU network (2) and let W denote a KKT point of Problem (3). Suppose the training data are p-orthogonal for $p \geq 3\gamma^{-3}$. Then, there exists $z \in \mathbb{R}^d$ such that for any $x \in \mathbb{R}^d$, $\operatorname{sign}\left(f(x; W)\right) = \operatorname{sign}\left(\langle z, x\rangle\right)$, and z is τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$ for $\tau = \frac{R^2}{\gamma^2}\left(1 + \frac{2}{\gamma pR^2-2}\right)$. Moreover, for any initialization $W(0)$, gradient flow on the logistic or exponential loss converges in direction to such a KKT point.*

Proposition 4 identifies an explicit formula for the limiting behavior of a neural network classifier trained by gradient flow in a *non-convex* setting. It is worth emphasizing that Proposition 4 does not make any assumptions on the width of the network or the initialization, and thus the characterization holds for neural networks in the feature-learning regime. Finally, note that as $p \to \infty$ and $R^2 \to 1$, KKT points of Problem (3) have the same decision boundary as a τ-uniform classifier for $\tau \to \gamma^{-2}$. In particular, if additionally the leaky parameter $\gamma \to 1$, the KKT points of leaky ReLU network margin-maximization problems become proportional as the sample average $\sum_{i=1}^n y_i x_i$, just as the linear max-margin predictor does.

Putting Proposition 3 and 4 together, we see that by understanding the behavior of τ-uniform classifiers $x \mapsto \operatorname{sign}\left(\langle \sum_{i=1}^n s_i y_i x_i, x\rangle\right)$, we can capture the behavior of both linear max-margin estimators as well as those of leaky ReLU networks trained by gradient flow with nearly-orthogonal data. In the following sections, we describe two distributional settings where we show that this estimator can exhibit *benign overfitting*: it achieves 0 training error on noisy datasets while simultaneously achieving test error near the noise rate.

4. Benign Overfitting for Sub-Gaussian Marginals

In this section we consider a distribution P_{sg} over (x, y) such that x has independent sub-Gaussian components, with a single high-variance component which determines the label y, while the remaining components of x have small variance. Let P_x be a distribution over $\mathbb{R}^d$. We assume the covariates $x \sim \mathsf{P}_x$ are mean-zero with covariance matrix $\Sigma = \mathbb{E}_{x\sim\mathsf{P}_x}[xx^\top]$ satisfying $\Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ where $\lambda_1 \geq \cdots \geq \lambda_d$. We assume that $z := \Sigma^{-1/2}x \sim \mathsf{P}_z$ where P_z is a sub-Gaussian random vector with independent components and sub-Gaussian norm at most σ_z (see Vershynin (2018) for more details on sub-Gaussian distributions). Given $x \sim \mathsf{P}_x$, labels are generated

as follows. For some label noise parameter $\eta \in (0, 1/2)$, we have $y = \mathrm{sign}([x]_1)$ with probability $1-\eta$ and $y = -\mathrm{sign}([x]_1)$ with probability η, where $[x]_1$ denotes the first component of x. Finally, we assume that for some absolute constant $\beta > 0$, we have $\mathbb{P}(|[z]_1| \leq t) \leq \beta t$ for all $t \geq 0$. In the remainder, we will assume that σ_z, η, and β are absolute constants, and our results will hold provided d and n are large enough relative to these and other universal constants.

The reader may be curious about the requirement that $\mathbb{P}_{z\sim \mathsf{P}_z}(|[z]_1| \leq t) \leq \beta t$. This is a technical assumption that ensures that the 'signal' in the model is large as it prevents the possibility that the mass of $[z]_1$ is highly concentrated near zero. Additionally, note that this assumption is satisfied if the distribution of either z or $[z]_1$ is (isotropic) log-concave by the anti-concentration property of isotropic log-concave distributions (Lovász and Vempala, 2007, Theorem 5.1 and Theorem 5.14).[2] This assumption also implies that $\mathbb{E}[|[z]_1|] \geq 1/(4\beta)$, since $\mathbb{P}(|z| \geq 1/(2\beta)) \geq 1/2$ by taking $t = 1/(2\beta)$. We can in principle accommodate more general conditions, such as $\mathbb{P}(|[z]_1| \leq t) \leq \beta t^p$ for some $p > 0$; this is a type of 'soft margin' condition which has been utilized in previous work on learning noisy halfspaces (Frei et al., 2021a,b).

We assume access to n i.i.d. training examples $\{(x_i, y_i)\} \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathsf{sg}}$. For a desired probability of failure $\delta \in (0, 1/2)$, we make the following assumptions on the problem parameters for a sufficiently large constant $C > 1$.

(SG1) The number of samples satisfies $n \geq C \log(6/\delta)$.

(SG2) The covariance matrix satisfies $\mathsf{StableRank}(\Sigma_{2:d}) \geq C \log(6n/\delta)$, where $\Sigma_{2:d}$ denotes the matrix $\mathrm{diag}(\lambda_2, \ldots, \lambda_d)$.

(SG3) The covariance matrix satisfies $\frac{\mathrm{tr}(\Sigma)}{\sqrt{\mathrm{tr}(\Sigma^2)}} \geq Cn \log(6n^2/\delta)$.

We remind the reader that the stable rank of a matrix $M \in \mathbb{R}^{m\times d}$ is $\mathsf{StableRank}(M) := \|M\|_F^2/\|M\|_2^2$. We note that the quantity $\mathrm{tr}(\Sigma)/\sqrt{\mathrm{tr}(\Sigma^2)}$ in (SG3) has appeared in previous work on benign overfitting: it is the square root of the "effective rank" $R_0(\Sigma)$ from Bartlett et al. (2020). Indeed, this quantity is large if $\sqrt{\mathsf{StableRank}(\Sigma^{1/2})}$ is large, since

$$\frac{\mathrm{tr}(\Sigma)}{\sqrt{\mathrm{tr}(\Sigma^2)}} \geq \frac{\mathrm{tr}(\Sigma)}{\sqrt{\|\Sigma\|_2 \mathrm{tr}(\Sigma)}} = \frac{\sqrt{\mathrm{tr}(\Sigma)}}{\|\Sigma^{1/2}\|_2} = \frac{\|\Sigma^{1/2}\|_F}{\|\Sigma^{1/2}\|_2} = \sqrt{\mathsf{StableRank}(\Sigma^{1/2})}.$$

Thus the Assumption (SG3) can be roughly understood as requiring that the matrix $\Sigma^{1/2}$ has sufficiently large rank. Additionally, we note that it is possible to have $\mathsf{StableRank}(\Sigma_{2:d}) = \Theta(1)$ while $\mathrm{tr}(\Sigma)/\sqrt{\mathrm{tr}(\Sigma^2)} = \Theta(\sqrt{d})$ (take $\Sigma = \mathrm{diag}(\sqrt{d}, \sqrt{d}, 1, \ldots, 1)$), and it is also possible for $\mathsf{StableRank}(\Sigma_{2:d}) = \Theta(\sqrt{d})$ while $\mathrm{tr}(\Sigma)/\sqrt{\mathrm{tr}(\Sigma^2)} = \Theta(1)$ (take $\Sigma = \mathrm{diag}(d, d^{1/4}, 1, \ldots, 1)$). Thus the Assumptions (SG2) and (SG3) are independent.

Our first lemma states that as the constant C in the preceding assumptions becomes larger, the training data becomes more orthogonal.

Lemma 5 *There exists an absolute constant $C_1 > 0$ (depending only on σ_z) such that for every large enough constant $C > 0$, for any $\delta \in (0, 1/2)$, under Assumptions (SG1) through (SG3) (defined*

2. For $z \sim \mathsf{P}_z$ where P_z is log-concave and isotropic, Lovász and Vempala (2007, Theorem 5.1) implies the one-dimensional marginal $[z]_1$ is isotropic and log-concave. Theorem 5.14 of the same reference shows that the density function of the (one-dimensional) $[z]_1$ is bounded from above by a constant, which implies $\mathbb{P}(|[z]_1| \leq t) \leq \beta t$ for an absolute constant $\beta > 0$.

for these C and δ), with probability at least $1-2\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$, the training data is ${}^{C}/_{C_1}$-orthogonal, and $\max_{i,j} {}^{\|x_i\|^2}/_{\|x_j\|^2} \leq (1+{}^{C_1}/_{\sqrt{C}})^4$.

The proof of Lemma 5, as well as all proofs for this section, appears in Appendix C. Recall from Propositions 3 and 4 that for p-orthogonal training data, as $R^2 = \max_{i,j} {}^{\|x_i\|^2}/_{\|x_j\|^2} \to 1$ and $p \to \infty$, solutions the linear max-margin problem (1) become τ-uniform for $\tau \to 1$. Similarly, KKT points of the leaky ReLU max margin problem behave like τ-uniform linear classifiers for $\tau \to \gamma^{-2}$ as $p \to \infty$ and $R \to 1$. In our main theorem for this section, we show that τ-uniform linear classifiers exhibit benign overfitting. We remind the reader that we refer to quantities that are independent of the dimension d, number of samples n, the failure probability δ or number of neurons m in the network as constants.

Theorem 6 *Let $\tau \geq 1$ be a constant, and suppose $\eta \leq \frac{1}{2\tau} - \Delta$ for some absolute constants $\eta, \Delta > 0$. There exist constants $C, C' > 0$ (depending only on $\eta, \sigma_z, \beta, \tau$, and Δ) such that for any $\delta \in (0, {}^1/_7)$, under Assumptions (SG1) through (SG3) (defined for these C and δ), with probability at least $1-7\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$, if $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$, then*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\left(\langle u, x_k\rangle\right), \qquad \text{while simultaneously,}$$
$$\eta \leq \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}\left(y \neq \operatorname{sign}(\langle u, x\rangle)\right) \leq \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

In particular, if $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$, then the linear classifier $x \mapsto \operatorname{sign}(\langle u, x\rangle)$ exhibits benign overfitting.

Theorem 6 shows that any τ-uniform estimator will exhibit benign overfitting, with the level of noise tolerated determined by the quantity τ. Moreover, by considering the 1-uniform estimator $\sum_{i=1}^n y_i x_i$, we see that there exists an estimator which can tolerate noise levels close to ${}^1/_2$.

Using Lemma 5 and Proposition 3, we can use Theorem 6 to characterize the linear max-margin predictor.

Corollary 7 *Suppose $0 < \eta \leq 0.49$. There exist constants $C, C' > 0$ such that for any $\delta \in (0, {}^1/_9)$, under Assumptions (SG1) through (SG3) (defined for these C and δ), with probability at least $1-9\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$, the max-margin linear classifier $w = \operatorname{argmin}\{\|w\|^2 : y_i\langle w, x_i\rangle \geq 1\ \forall i\}$ satisfies*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\left(\langle w, x_k\rangle\right), \qquad \text{while simultaneously,}$$
$$\eta \leq \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}\left(y \neq \operatorname{sign}(\langle w, x\rangle)\right) \leq \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

In particular, if $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$ then w exhibits benign overfitting.

The proof of Corollary 7 is the result of a simple calculation (for completeness it is provided in Appendix C): By Theorem 6, we can tolerate noise rates η close to $\frac{1}{2}$ if τ is close to one. By Lemma 5, as C gets larger the training data becomes more orthogonal and the ratio of the norms of the examples becomes closer to one. By Proposition 3 this implies $\tau \to 1$ as C increases.

We can similarly use Lemma 5 and Proposition 4 to show that KKT points of the max-margin problem for leaky ReLU networks from Problem (3) also exhibit benign overfitting.

Corollary 8 *Suppose that* $0 < \eta \le \frac{49\gamma^2}{100}$. *There exist constants* $C, C' > 0$ *such that for any* $\delta \in (0, 1/9)$, *under Assumptions (SG1) through (SG3) (defined for these* C *and* δ*), with probability at least* $1 - 9\delta$ *over* $\mathsf{P}_{\mathsf{sg}}^n$, *any KKT point* W *of Problem (3) satisfies*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\big(f(x_k; W)\big), \quad \text{while simultaneously,}$$

$$\eta \le \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}\Big(y \ne \operatorname{sign}\big(f(x;W)\big)\Big) \le \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

In particular, if $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$ *then the neural network* $f(x; W)$ *exhibits benign overfitting. Moreover, for any initialization* $W(0)$, *gradient flow converges in direction to a network satisfying the above.*

The proof of Corollary 8 similarly requires a small calculation which we provide in Appendix C. It is noteworthy that the only difference in the behavior of KKT points of the leaky ReLU max-margin problem (3) and the linear max-margin (1) is the level of noise that is tolerated: in the leaky ReLU case, smaller leaky parameters γ result in less noise tolerated, and as $\gamma \to 1$, we recover the behavior of the linear max-margin predictor from Corollary 7. Additionally, the generalization bound in Corollary 8 does not depend on the number m of neurons in the network.

From the above results, we see that in order for benign overfitting to occur in either the linear max-margin classifier or in two-layer leaky ReLU networks trained by gradient flow, the data needs to simultaneously satisfy two constraints: (1) the covariance matrix is sufficiently high rank in the sense of Assumptions (SG2) and (SG3), and (2) the variance in the first coordinate must be large relative to the variance of the last $d - 1$ coordinates. There is a tension here as can be seen by considering the covariance matrix $\Sigma = (\xi, 1, \ldots, 1)$ for $\xi \ge 1$: as $\xi \to \infty$, $\operatorname{tr}(\Sigma)/\sqrt{\operatorname{tr}(\Sigma^2)} \to 1$, and hence as the signal-to-noise ratio $\lambda_1^2/\operatorname{tr}(\Sigma_{2:d}^2)$ increases, it becomes more difficult to satisfy assumption (SG3). However, it is indeed possible to satisfy both (1) and (2). Consider the distribution $\mathsf{P}_{\mathsf{gaus}}$ over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ where $x \sim \mathsf{N}(0, \Sigma)$ with covariance matrix $\Sigma = \operatorname{diag}(d^\rho, 1, \ldots, 1)$ for some $\rho > 0$, and where $y = \operatorname{sign}([x]_1)$ with probability $1 - \eta$ and $y = -\operatorname{sign}([x]_1)$ with probability η for some constant $\eta > 0$. In the following corollary, we show that if $\rho \in (1/2, 1)$, then (1) and (2) are satisfied and so KKT points of the leaky ReLU max-margin problem (3) exhibit benign overfitting (an analogous result for the linear max-margin classifier holds as well).

Corollary 9 *Suppose* $0 < \eta \le \frac{49\gamma^2}{100}$. *Then for the distribution* $\mathsf{P}_{\mathsf{gaus}}$, *for any* $\delta \in (0, 1/9)$, *if* $\rho \in (1/2, 1)$, $d = \tilde{\Omega}(n^{1/(1-\rho)})$, *and* $n = \tilde{\Omega}(1)$, *then Assumptions (SG1) through (SG3) are satisfied. Moreover, with probability at least* $1 - 9\delta$ *over* $\mathsf{P}_{\mathsf{gaus}}^n$, *KKT points of Problem (3) exhibit benign overfitting:*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\big(f(x_k; W)\big),$$

$$\text{while simultaneously,} \quad \eta \le \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{gaus}}}\Big(y \ne \operatorname{sign}\big(f(x_k;W)\big)\Big) \le \eta + \tilde{O}\left(d^{\frac{1}{2}(1-2\rho)}\right) = \eta + o_d(1).$$

Furthermore, for any initialization $W(0)$, *gradient flow converges in direction to a network satisfying the above.*

We note that a similar result on benign overfitting for the linear max-margin classifier for data coming from $\mathsf{P}_{\mathsf{gaus}}$ has been shown by Muthukumar et al. (2021) with a rather different proof technique.

5. Benign Overfitting for Clustered Data

In this section we consider a distribution where data comes from multiple clusters and data from each cluster initially share the same label but then are flipped with some constant probability η. In particular, we consider a distribution $\mathsf{P}_{\mathsf{clust}}$ over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ defined as follows. Let $k \geq 2$ and $Q := [k]$. We are given cluster means $\mu^{(1)}, \ldots, \mu^{(k)}$ with cluster labels $\tilde{y}^{(1)}, \ldots, \tilde{y}^{(k)} \in \{\pm 1\}$. Cluster indices are sampled $q \sim \mathsf{Unif}(Q)$, after which $x|q \sim \mu^{(q)} + z$ where $z \sim \mathsf{P}'_z$ is such that: the components of z are mean-zero, independent, sub-Gaussian random variables with sub-Gaussian norm at most one; and $\mathbb{E}[\|z\|^2] = d$.[3] Finally, the (clean) label of x is $\tilde{y} = y^{(q)}$, and the observed label is $y = \tilde{y}$ with probability $1 - \eta$ and $y = -\tilde{y}$ with probability $1 - \eta$.

For a given $\delta \in (0, 1/2)$, we make the following assumptions on the parameters, for a sufficiently large constant $C > 1$:

(CL1) Number of samples $n \geq Ck^2 \log(k/\delta)$.

(CL2) Dimension $d \geq C \max\{n \max_q \|\mu^{(q)}\|^2, n^2 \log(n/\delta)\}$.

(CL3) The cluster means satisfy: $\min_q \|\mu^{(q)}\| \geq Ck\sqrt{\log(2nk/\delta)}$.

(CL4) The cluster means are nearly-orthogonal in the sense that: $\min_q \|\mu^{(q)}\|^2 \geq Ck \max_{q \neq r} |\langle \mu^{(q)}, \mu^{(r)} \rangle|$.

We shall show below that under these assumptions, the training data is linearly separable with high probability.

Our first lemma shows that under the preceding assumptions, the training data become more orthogonal and the ratio of the norms of the examples tends to one as C increases.

Lemma 10 *There exists an absolute constant $C_2 > 0$ such for every large enough constant $C > 0$, for any $\delta \in (0, 1/7)$, under Assumptions (CL1) through (CL4) (defined for these C and δ), with probability at least $1 - 7\delta$ over $\mathsf{P}_{\mathsf{clust}}$, the training data is C/C_2-orthogonal, and $\max_{i,j} \|x_i\|^2/\|x_j\|^2 \leq (1 + C_2/\sqrt{C})^2$.*

The proof of the above lemma, as well as all proofs for this section, appears in Appendix D. As before, Lemma 10 allows for us to utilize Propositions 3 and 4 to show that both KKT points of the linear max-margin problem (1) and of the leaky ReLU network max-margin problem (2) take the form $\sum_{i=1}^n s_i y_i x_i$. The following theorem characterizes the performance of this predictor.

Theorem 11 *Let $\tau \geq 1$ be a constant, and suppose $\eta \leq \frac{1}{1+\tau} - \Delta$ for some absolute constants $\eta, \Delta > 0$. There exist constants $C, C' > 0$ (depending only on η, τ, and Δ) such that for any $\delta \in (0, 1/14)$, under Assumptions (CL1) through (CL4) (defined for these C and δ), with probability at least $1 - 14\delta$ over $\mathsf{P}^n_{\mathsf{clust}}$, if $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$, then*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\left(\langle u, x_k \rangle\right),$$

$$\text{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{clust}}}\left(y \neq \operatorname{sign}(\langle u, x \rangle)\right) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

In particular, if $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$, then the linear classifier $x \mapsto \operatorname{sign}(\langle u, x \rangle)$ exhibits benign overfitting.

3. We can easily accommodate well-conditioned clusters, e.g. $\kappa d \leq \mathbb{E}\|z\|^2 \leq d$ for some absolute constant $\kappa > 0$, although the noise rate tolerated will then depend upon κ (smaller κ will require smaller η). We do not do so for simplicity of exposition.

In order for benign overfitting to occur, the above theorem requires that Assumptions (CL1) through (CL4) are satisfied while simultaneously $\min_q \|\mu^{(q)}\|^4 = \omega(k^2 d/n)$. This can be satisfied in a number of settings, such as:

(i) Orthogonal clusters with $\|\mu^{(q)}\| = \Theta(d^\beta)$ for each $q \in Q$, where $\beta \in (1/4, 1/2)$, $k = O(1)$, $n = \tilde{\Omega}(1)$ and $d = \tilde{\Omega}(n^{1/(1-2\beta)})$. In this setting the test error is at most $\eta + \exp\left(-\tilde{\Omega}(nd^{4\beta-1})\right)$.

(ii) Non-orthogonal clusters where $\max_{q \neq r} |\langle \mu^{(q)}, \mu^{(r)} \rangle| = O(d^{\frac{3}{5}})$ and $\|\mu^{(q)}\| = \Theta(d^{\frac{1}{3}})$ for each q, $n = \Theta(d^{\frac{1}{5}})$, and $k = \Theta(d^{0.05})$. In this setting the test error is at most $\eta + \exp\left(-\Omega(d^{0.43})\right)$.

Although neither of the above settings are explicitly covered by Theorem 6, the setting (i) is similar in flavor to that theorem, in that the labels are determined by a constant number of high-variance directions. By contrast, the setting (ii) is quite different, as it allows for (clean) labels to be determined by the output of a linear classifier over $k = \Theta(d^{0.05})$ components, namely $\operatorname{sign}\left(\langle \sum_{q=1}^k y^{(q)} \mu^{(q)}, x \rangle\right)$.

Just as in the case of Theorem 6, we have a number of corollaries of Theorem 11. The first is a consequence of Proposition 3.

Corollary 12 *Suppose $0 < \eta \leq 0.49$. There exist constants $C, C' > 0$ such that for any $\delta \in (0, 1/21)$, under Assumptions (CL1) through (CL4) (defined for these C and δ), with probability at least $1 - 21\delta$ over $\mathsf{P}_{\mathsf{clust}}^n$, the max-margin linear classifier $w = \operatorname{argmin}\{\|w\|^2 : y_i \langle w, x_i \rangle \geq 1 \, \forall i\}$ satisfies*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\left(\langle w, x_k \rangle\right),$$

$$\text{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y) \sim \mathsf{P}_{\mathsf{clust}}}\left(y \neq \operatorname{sign}(\langle w, x \rangle)\right) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

In particular, if $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$ then w exhibits benign overfitting.

Similarly, we can show that KKT points of the max-margin problem for leaky ReLU networks from Problem (3) also exhibit benign overfitting.

Corollary 13 *Suppose that $0 < \eta \leq \frac{49\gamma^2}{100}$. There exist constants $C, C' > 0$ such that for any $\delta \in (0, 1/21)$, under Assumptions (CL1) through (CL4) (defined for these C and δ), with probability at least $1 - 21\delta$ over $\mathsf{P}_{\mathsf{clust}}^n$, any KKT point W of Problem (3) satisfies*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\left(f(x_k; W)\right),$$

$$\text{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y) \sim \mathsf{P}_{\mathsf{clust}}}\left(y \neq \operatorname{sign}\left(f(x; W)\right)\right) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

In particular, if $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$ then the neural network $f(x; W)$ exhibits benign overfitting. Moreover, for any initialization $W(0)$, gradient flow converges in direction to a network which satisfies the above.

We would like to note that Theorem 11 (and the subsequent corollaries) does not explicitly cover the case that the data consists of two opposing clusters, i.e. when $\tilde{y} \sim \mathsf{Unif}(\{\pm 1\})$ and $x|\tilde{y} \sim \tilde{y}\mu + z$ for some random vector z and labels are flipped $\tilde{y} \mapsto -\tilde{y}$ with some probability $\eta < 1/2$. A number of recent works showed that the linear max-margin classifer (Chatterji and Long, 2021; Wang and Thrampoulidis, 2021) and two-layer neural networks with *smooth* leaky ReLU activations (Frei et al., 2022) exhibit benign overfitting for this distributional setting. However, our analysis can easily be extended to show benign overfitting of the linear max-margin and KKT points of the leaky ReLU network max-margin problem (3) for this distribution by using a small modification of the proof we use for Theorem 11. We also wish to emphasize that the results of Frei et al. (2022) are specific to networks with smooth leaky ReLU activations, which are not homogeneous and for which gradient flow does not have a known implicit bias towards satisfying the KKT conditions for margin-maximization. In particular, their analysis is based on tracking the generalization error of the neural network throughout the training trajectory, while ours relies upon the structure imposed by the KKT conditions for margin-maximization in homogeneous networks. Another difference between our work and theirs concerns the label noise model. We derive an explicit upper bound on the noise level tolerated, while their results hold for noise levels below an unspecified constant. Our analysis holds for labels flipped with constant probability, while theirs permits adversarial label noise.

6. Proof intuition

In this section we provide some intuition for how benign overfitting of the max-margin linear classifier is possible. We consider a distribution $\mathsf{P}_{\mathsf{opp}}$ defined by a mean vector $\mu \in \mathbb{R}^d$ and label noise parameter $\eta \in (0, 1/2)$, where examples $(x, y) \sim \mathsf{P}_{\mathsf{opp}}$ are sampled as follows:

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad z \sim \mathsf{N}(0, I_d), \quad x|\tilde{y} \sim \tilde{y}\mu + z, \quad \begin{cases} y = \tilde{y}, & \text{w.p. } 1-\eta, \\ y = -\tilde{y}, & \text{w.p. } \eta. \end{cases} \tag{4}$$

As we mentioned in the previous section, this distribution is not explicitly covered by Theorem 11 but the intuition and proof are essentially the same. Our starting point is Proposition 3, which shows the max-margin linear classifier is τ-uniform over the training data when the training data are nearly orthogonal. For simplicity let us consider the simplest estimator of this form, the 1-uniform vector $\widehat{\mu} = \sum_{i=1}^n y_i x_i$. Let us call the training examples (x_i, y_i) for which $y_i = \tilde{y}_i$ the *clean* examples, and denote the indices corresponding to such examples $\mathcal{C} \subset [n]$, with the examples with $y_i = -\tilde{y}_i$ the *noisy* examples, identified by $\mathcal{N} \subset [n]$ (so $\mathcal{C} \cup \mathcal{N} = [n]$). For the distribution (4) and training data $\{(x_i, y_i)\}_{i=1}^n$, the estimator $\widehat{\mu}$ thus takes the form

$$\widehat{\mu} = \sum_{i=1}^n y_i x_i = \sum_{i \in \mathcal{C}} (\mu + y_i z_i) + \sum_{i \in \mathcal{N}} (-\mu + y_i z_i) = (|\mathcal{C}| - |\mathcal{N}|)\,\mu + \sum_{i=1}^n y_i z_i \propto \mu + \frac{1}{|\mathcal{C}| - |\mathcal{N}|} \sum_{i=1}^n y_i z_i,$$

where $z_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$. Now, provided n is sufficiently large and if the noise rate is smaller than say $1/4$, then with high probability we have $9n/10 \geq |\mathcal{C}| - |\mathcal{N}| \geq n/10$. In particular, the estimator $\widehat{\mu}$ is proportional to a sum of two components: μ, which is the linear classifier which achieves optimal accuracy for the distribution, and $(|\mathcal{C}| - |\mathcal{N}|)^{-1} \sum_{i=1}^n y_i z_i$ which incorporates only the noise. This latter component is useless for prediction on fresh test examples, but is quite useful for achieving

small training error. Indeed, for training example $(x_k, y_k) = (\tilde{y}_k\mu + z_k, y_k)$,

$$\begin{aligned}\langle y_k x_k, \textstyle\sum_{i=1}^n y_i z_i\rangle &= \langle y_k\tilde{y}_k\mu + y_k z_k, y_k z_k + \textstyle\sum_{i\neq k} y_i z_i\rangle \\ &= \|z_k\|^2 + y_k\tilde{y}_k\langle\mu, \textstyle\sum_{i=1}^n y_i z_i\rangle + \langle y_k z_k, \textstyle\sum_{i\neq k} y_i z_i\rangle.\end{aligned}$$

Since $z_k \sim \mathsf{N}(0, I_d)$ and $\sum_{i=1}^n y_i z_i \sim \mathsf{N}(0, nI_d)$, standard concentration bounds show that $\|z_k\|^2 \gtrsim d$ while $|\langle\mu, \sum_{i=1}^n y_i z_i\rangle| \lesssim \sqrt{n}\|\mu\|$, and $|\langle y_k z_k, \sum_{i\neq k} y_i z_i\rangle| \lesssim \sqrt{nd}$ (ignoring log factors for simplicity). Thus, provided $d \gg \sqrt{nd}$ and $d \gg \sqrt{n}\|\mu\|$, the estimator $\xi := (|\mathcal{C}|-|\mathcal{N}|)^{-1}\sum_{i=1}^n y_i z_i$ satisfies

$$\langle \xi, y_k x_k\rangle \gtrsim d/n. \tag{5}$$

On the other hand, the effect of ξ on an independent test example (x, y) is,

$$|\langle yx, \xi\rangle| = \left|\tfrac{1}{|\mathcal{C}|-|\mathcal{N}|}\langle y\tilde{y}(z+\mu), \textstyle\sum_{i=1}^n y_i z_i\rangle\right| \lesssim \frac{1}{n}\left(\sqrt{n}\|\mu\| + \sqrt{dn}\right) = \frac{\|\mu\| + \sqrt{d}}{\sqrt{n}}. \tag{6}$$

Putting (5) and (6) together, we see that ξ has a significantly larger effect on the training data than on the test data performance as long as $d \gg \sqrt{n}\|\mu\| + \sqrt{dn}$. Assuming $\|\mu\| < \sqrt{d}$ this holds when $d \gg n^2$. In particular, the possibility of a component which enables benign overfitting becomes easier in high dimensions, at least when the signal $\|\mu\|$ is not too large.

The above sketch shows that the overfitting component ξ is useful for interpolating the (noisy) training data when $\|\mu\| < \sqrt{d}$ and $d \gg n^2$. However, the estimator $\widehat{\mu} \propto \mu + \xi$ also contains the component μ which is biased towards getting noisy training data *incorrect*. Thus, in order to show $\mu + \xi$ exhibits benign overfitting, we need to show (i) the signal strength from μ is not so strong as to prevent overfitting the noisy training data, but (ii) the signal strength from μ is large enough to enable good generalization from test data. For part (i), standard concentration bounds imply that

$$|\langle\mu, y_k x_k\rangle| = |y_k\tilde{y}_k\|\mu\|^2 + \langle y_k z_k, \mu\rangle| \lesssim \|\mu\|^2 + \|\mu\|. \tag{7}$$

In light of (5), the estimator $\mu+\xi$ will still interpolate the training data provided $d/n \gg \max(\|\mu\|, \|\mu\|^2)$. For part (ii), for a given *clean* test example $(x, \tilde{y})$,

$$\langle\tilde{y}x, \mu\rangle = \langle\mu + \tilde{y}z, \mu\rangle \gtrsim \|\mu\|^2 - C\|\mu\|. \tag{8}$$

Thus, provided $\|\mu\| \gg C$ and $\|\mu\|^2 \gg n^{-1/2}(\|\mu\| + \sqrt{d})$, we can be ensured that $\mu + \xi$ will also classify clean test examples correctly by putting together (8) and (6).

To summarize, we have identified settings under which we can guarantee that benign overfitting occurs for the estimator $\widehat{\mu} \propto \mu + \xi$ for the distribution $\mathsf{P}_{\mathsf{opp}}$. First, the training data must be sufficiently high-dimensional to ensure that the overfitting component ξ has a significant effect on the training data but little effect on future test data. Second, the underlying signal of $\widehat{\mu}$ (whose strength is measured by $\|\mu\|$) must not be so strong as to prevent overfitting to the noisy labels, yet must also be strong enough to ensure that future test data can be accurately predicted.

7. Discussion

We have characterized a number of new settings under which linear classifiers and two-layer neural networks can exhibit benign overfitting. We showed how the implicit bias of gradient flow imposes

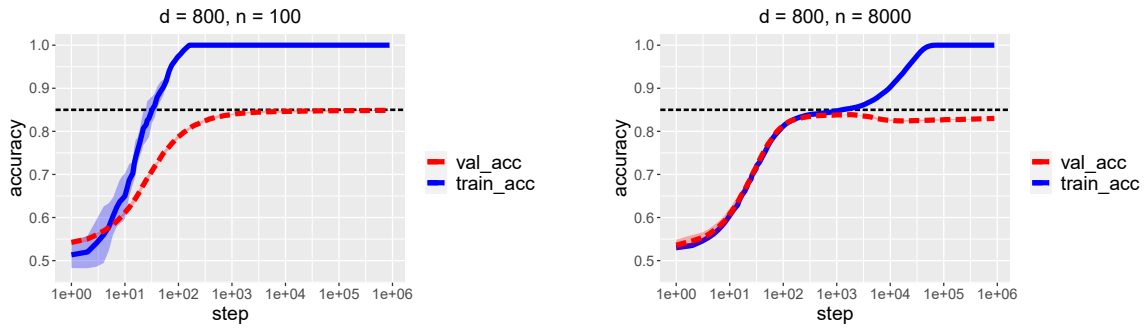

Figure 1: When training a two-layer leaky ReLU network with $m = 512$ neurons for data coming from the binary mixture distribution (4), there is qualitatively different generalization behavior in the $d \gg n$ vs. $n \gg d$ regimes. In this experiment, 15% of the labels are flipped to the opposing cluster's label, and $\|\mu\| = d^{0.26}$. In the high-dimensional regime (left), the network achieves 100% training accuracy and the optimal 85% validation accuracy, while in the low-dimensional regime (right) the validation accuracy is sub-optimal when it achieves 100% training accuracy.

significant structure on linear classifiers and neural networks trained by this method, and how this structure can be leveraged to understand the generalization of interpolating models in the presence of noisy labels.

Our analysis holds in the regime where the input dimension is significantly larger than the number of samples. This assumption is present in previous works on benign overfitting in neural networks (Frei et al., 2022; Cao et al., 2022; Xu and Gu, 2023; Kou et al., 2023; Kornowski et al., 2023). However, it is currently unknown whether benign overfitting is possible in neural network classification tasks when the dimension is fixed. In Figure 1, we examine the behavior of two-layer leaky ReLU networks trained by gradient descent for data coming from the binary mixture distribution (4) when 15% of the labels are flipped. (Further details on the experiment can be found in Appendix E.) We see that in this setting, the generalization of interpolating neural networks differs in the $d \gg n$ vs. the $n \gg d$ regime: in the high-dimensional setting, overfitting is benign as would be expected by Corollary 13, while in the low-dimensional setting, the validation accuracy is sub-optimal when the network interpolates the noisy training data. This suggests that new techniques would be needed to characterize generalization of interpolating networks in the low-dimensional setting.

There are a number of additional directions for future research. For instance, the larger class of homogeneous neural networks trained by the logistic loss also have an implicit bias towards satisfying the KKT conditions for margin-maximization. Can this implicit bias be leveraged to show benign overfitting in neural networks with ReLU activations of depth $L \geq 2$? Additionally, although two-layer leaky ReLU networks are in general nonlinear, our proof holds in settings where their decision boundaries are linear. It would be interesting to understand benign overfitting in neural networks when the learned decision boundary is nonlinear.

Acknowledgments

We gratefully 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.

References

Peter L Bartlett and Philip M Long. Failures of model-dependent generalization bounds for least-norm interpolation. *The Journal of Machine Learning Research*, 22(1):9297–9311, 2021.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022.

Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

Niladri S Chatterji, Philip M Long, and Peter L Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks. *arXiv preprint arXiv:2108.11489*, 2021.

Geoffrey Chinot and Matthieu Lerasle. On the robustness of the minimum ℓ_2 interpolator. *arXiv preprint arXiv:2003.05838*, 2020.

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.

Francis H Clarke, Yuri S Ledyaev, Ronald J Stern, and Peter R Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *International Conference on Machine Learning (ICML)*, 2022.

Joydeep Dutta, Kalyanmoy Deb, Rupesh Tulshyan, and Ramnik Arora. Approximate kkt points and a proximity measure for termination. *Journal of Global Optimization*, 56(4):1463–1499, 2013.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning (ICML)*, 2021a.

Spencer Frei, Yuan Cao, and Quanquan Gu. Provable generalization of SGD-trained neural networks of any width in the presence of adversarial label noise. In *International Conference on Machine Learning (ICML)*, 2021b.

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.

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*, 2023.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *Preprint, arXiv:1802.08246*, 2018.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Preprint, arXiv:1903.08560*, 2020.

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 91–99, 2021.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *Preprint, arXiv:1803.07300*, 2018.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

Ziwei Ji, Miroslav Dudik, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds, and benign overfitting. *arXiv preprint arXiv:2106.09276*, 2021.

Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From tempered to benign overfitting in relu neural networks. *Preprint, arXiv:2305.15141*, 2023.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting for two-layer relu convolutional networks. In *International Conference on Machine Learning (ICML)*, 2023.

Daniel Kunin, Atsushi Yamamura, Chao Ma, and Surya Ganguli. The asymmetric maximum margin bias of quasi-homogeneous neural networks. *arXiv preprint arXiv:2210.03820*, 2022.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory (COLT)*, 2020.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007. ISSN 1042-9832.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

Neil Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 2022.

Andrew D McRae, Santhosh Karnik, Mark Davenport, and Vidya K Muthukumar. Harmless interpolation in regression and classification with structured features. In *International Conference on Artificial Intelligence and Statistics*, pages 5853–5875. PMLR, 2022.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *Preprint, arXiv:1911.01544*, 2020.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019a.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019b.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272, 2020.

Mary Phuong and Christoph H Lampert. The inductive bias of relu networks on orthogonally separable data. In *International Conference on Learning Representations (ICLR)*, 2020.

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory (COLT)*, 2019.

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. *Preprint, arXiv:2205.09072*, 2022.

Roei Sarussi, Alon Brutzkus, and Amir Globerson. Towards understanding learning in neural networks with linear teachers. In *International Conference on Machine Learning (ICML)*, 2021.

Ohad Shamir. Gradient methods never overfit on separable data. *Preprint, arXiv:2007.00028*, 2020.

Ohad Shamir. The implicit bias of benign overfitting. In *Conference on Learning Theory*, pages 448–478. PMLR, 2022.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research (JMLR)*, 19(70):1–57, 2018.

Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. *Advances in Neural Information Processing Systems*, 33:8907–8920, 2020.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. *arXiv preprint arXiv:2201.12760*, 2022.

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *Preprint, arXiv:2009.14286*, 2020.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Preprint, arXiv:2208.12591*, 2022.

Gal Vardi, Ohad Shamir, and Nathan Srebro. On margin maximization in linear and relu networks. *arXiv preprint arXiv:2110.02732*, 2021.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Gradient methods provably converge to non-robust networks. *Preprint, arXiv:2202.04347*, 2022.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum l1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting and regularization. *Preprint, arXiv:2011.09148*, 2021.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Xingyu Xu and Yuantao Gu. Benign overfitting of non-smooth neural networks beyond lazy training. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Lijia Zhou, Frederic Koehler, Pragya Sur, Danica J Sutherland, and Nathan Srebro. A non-asymptotic moreau envelope theory for high-dimensional generalized linear models. *arXiv preprint arXiv:2210.12082*, 2022.

Appendix A. Proof of Proposition 3

Since $\widehat{w}$ satisfies the KKT conditions of the max-margin problem, we have $\widehat{w} = \sum_{i=1}^{n} \lambda_i y_i x_i$ where for all $i \in [n]$ we have $\lambda_i \geq 0$, and $\lambda_i = 0$ if $y_i \widehat{w}^\top x_i \neq 1$. We denote $R_{\min} = \min_i \|x_i\|$, $R_{\max} = \max_i \|x_i\|$, and $R = R_{\max}/R_{\min}$.

In the following lemma, we obtain an upper bound for the λ_i's:

Lemma 14 *Suppose the training data is p-orthogonal. Then for all $i \in [n]$ we have $\lambda_i \leq \frac{1}{R_{min}^2\left(1-\frac{1}{pR^2}\right)}$.*

Proof Let $j \in \operatorname{argmax}_{i\in[n]} \lambda_i$. We have

$$y_j \widehat{w}^\top x_j = \sum_{i=1}^{n} \lambda_i y_j y_i x_i^\top x_j = \lambda_j \|x_j\|^2 + \sum_{i\neq j} \lambda_i y_j y_i x_i^\top x_j \geq \lambda_j R_{\min}^2 - n \left(\max_{i\in[n]} \lambda_i\right) \left(\max_{i\neq j} |\langle x_i, x_j\rangle|\right) . \tag{9}$$

By the p-orthogonality assumption, we also have

$$n \max_{i\neq j} |\langle x_i, x_j\rangle| \leq \frac{R_{\min}^2}{pR^2} . \tag{10}$$

Suppose that

$$\lambda_j = \max_{i\in[n]} \lambda_i > \frac{1}{R_{\min}^2\left(1-\frac{1}{pR^2}\right)} . \tag{11}$$

Combining (9), (10) and (11), we get

$$y_j \widehat{w}^\top x_j \geq \lambda_j R_{\min}^2 - \lambda_j \cdot \frac{R_{\min}^2}{pR^2} = \lambda_j R_{\min}^2 \left(1 - \frac{1}{pR^2}\right) > 1 .$$

By the KKT conditions, if $y_j \widehat{w}^\top x_j > 1$ then we must have $\lambda_j = 0$, and thus we reach a contradiction. ■

Next, we obtain a lower bound on the λ_i's:

Lemma 15 *Suppose the training data is p-orthogonal. Then for all $i \in [n]$ we have $\lambda_i \geq \frac{1}{R_{max}^2}\left(1 - \frac{1}{pR^2-1}\right)$.*

Proof Let $j \in [n]$. By the definition of $\widehat{w}$ we have

$$\begin{aligned} 1 &\leq y_j \widehat{w}^\top x_j \\ &= \sum_{i=1}^{n} \lambda_i y_j y_i x_i^\top x_j \\ &= \lambda_j \|x_j\|^2 + \sum_{i\neq j} \lambda_i y_j y_i x_i^\top x_j \\ &\leq \lambda_j R_{\max}^2 + n \left(\max_{i\in[n]} \lambda_i\right) \left(\max_{i\neq j} |\langle x_i, x_j\rangle|\right) . \end{aligned} \tag{12}$$

By the p-orthogonality assumption, we have

$$n \max_{i \neq j} |\langle x_i, x_j \rangle| \leq \frac{R_{\min}^2}{pR^2} \,, \tag{13}$$

and by Lemma 14 we have

$$\max_{i \in [n]} \lambda_i \leq \frac{1}{R_{\min}^2 \left(1 - \frac{1}{pR^2}\right)} \,. \tag{14}$$

Combining (12), (13) and (14), we get

$$1 \leq \lambda_j R_{\max}^2 + \frac{1}{R_{\min}^2 \left(1 - \frac{1}{pR^2}\right)} \cdot \frac{R_{\min}^2}{pR^2} = \lambda_j R_{\max}^2 + \frac{1}{pR^2 - 1} \,.$$

Hence,

$$\lambda_j \geq \left(1 - \frac{1}{pR^2 - 1}\right) \frac{1}{R_{\max}^2} \,.$$

■

Combining Lemmas 14 and 15, we conclude that $\widehat{w} = \sum_{i=1}^n \lambda_i y_i x_i$ where

$$\begin{aligned}
\frac{\max_i \lambda_i}{\min_i \lambda_i} &\leq \frac{1}{R_{\min}^2 \left(1 - \frac{1}{pR^2}\right)} \cdot R_{\max}^2 \left(1 - \frac{1}{pR^2 - 1}\right)^{-1} \\
&= \frac{pR^2}{R_{\min}^2 (pR^2 - 1)} \cdot R_{\max}^2 \cdot \frac{pR^2 - 1}{pR^2 - 2} \\
&= \frac{R_{\max}^2}{R_{\min}^2} \cdot \frac{pR^2}{pR^2 - 2} \\
&= R^2 \left(1 + \frac{2}{pR^2 - 2}\right) \,.
\end{aligned}$$

Appendix B. Proof of Proposition 4

We start with some notations. For convenience, we will use different notations for positive neurons (i.e., where $a_j = 1/\sqrt{m}$) and negative neurons (i.e., where $a_j = -1/\sqrt{m}$). Namely,

$$f(x; W) = \sum_{j=1}^m a_j \phi(w_j^\top x) = \sum_{j=1}^{m/2} \frac{1}{\sqrt{m}} \phi(v_j^\top x) - \sum_{j=1}^{m/2} \frac{1}{\sqrt{m}} \phi(u_j^\top x) \,.$$

We denote $\zeta = \max_{i \neq j} |\langle x_i, x_j \rangle|$. Thus, our near-orthogonality assumption can be written as $n\zeta \leq \frac{R_{\min}^2}{pR^2}$. Since W satisfies the KKT conditions of Problem (3), then there are $\lambda_1, \ldots, \lambda_n$ (known as *KKT multipliers*) such that for every $j \in [m/2]$ we have

$$v_j = \sum_{i \in [n]} \lambda_i \nabla_{v_j} \left(y_i f(x_i; W)\right) = \frac{1}{\sqrt{m}} \sum_{i \in [n]} \lambda_i y_i \phi'_{i, v_j} x_i \,, \tag{15}$$

where ϕ'_{i,v_j} is a subgradient of ϕ at $v_j^\top x_i$, i.e., if $v_j^\top x_i > 0$ then $\phi'_{i,v_j} = 1$, if $v_j^\top x_i < 0$ then $\phi'_{i,v_j} = \gamma$ and otherwise ϕ'_{i,v_j} is some value in $[\gamma, 1]$. Also, we have $\lambda_i \geq 0$ for all i, and $\lambda_i = 0$ if $y_i f(x_i; W) \neq 1$. Likewise, for all $j \in [m/2]$ we have

$$u_j = \sum_{i\in[n]} \lambda_i \nabla_{u_j} \left(y_i f(x_i; W)\right) = \frac{1}{\sqrt{m}} \sum_{i\in[n]} \lambda_i (-y_i) \phi'_{i,u_j} x_i \,, \tag{16}$$

where ϕ'_{i,u_j} is defined similarly to ϕ'_{i,v_j}.

Our proof builds on the following lemma, which follows from Frei et al. (2023) (note that within their notation, in our setting we have $m_1 = m_2 = m/2$):

Lemma 16 (Frei et al. (2023), Theorem 3.2 & Corollary 3.5) *Denote $R^2_{min} = \min_i \|x_i\|^2$ and $R^2_{max} = \max_i \|x_i\|^2$. Let f denote the leaky ReLU network (2) and let W denote a KKT point of Problem (3). Let $\lambda_1, \ldots, \lambda_n \geq 0$ denote the corresponding KKT multipliers. Suppose the training data are p-orthogonal for $p \geq 3\gamma^{-3}$. Then, we have $\lambda_i \in \left(\frac{1}{2R^2_{max}}, \frac{3}{2\gamma^2 R^2_{min}}\right)$ for all $i \in [n]$, and for any $x \in \mathbb{R}^d$ we have $\operatorname{sign}\left(f(x; W)\right) = \operatorname{sign}\left(\langle z, x\rangle\right)$, where $z = \frac{\sqrt{m}}{2} v - \frac{\sqrt{m}}{2} u$ for*

$$v = \frac{1}{\sqrt{m}} \sum_{i:y_i=1} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i:y_i=-1} \lambda_i x_i \,,$$

and

$$u = \frac{1}{\sqrt{m}} \sum_{i:y_i=-1} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i:y_i=1} \lambda_i x_i \,.$$

Moreover, for any initialization $W(0)$, gradient flow on the logistic or exponential loss converges in direction to such a KKT point.

Note that the above lemma implies that $\operatorname{sign}\left(f(x; W)\right) = \operatorname{sign}\left(\langle z, x\rangle\right)$ for

$$\begin{aligned} z &= \frac{\sqrt{m}}{2}\left(\frac{1}{\sqrt{m}} \sum_{i:y_i=1} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i:y_i=-1} \lambda_i x_i\right) - \frac{\sqrt{m}}{2}\left(\frac{1}{\sqrt{m}} \sum_{i:y_i=-1} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i:y_i=1} \lambda_i x_i\right) \\ &= \frac{1+\gamma}{2} \sum_{i:y_i=1} \lambda_i x_i - \frac{1+\gamma}{2} \sum_{i:y_i=-1} \lambda_i x_i \\ &= \frac{1+\gamma}{2} \sum_{i=1}^{n} y_i \lambda_i x_i \,. \end{aligned} \tag{17}$$

The lemma also implies that $\lambda_i \in \left(\frac{1}{2R^2_{\max}}, \frac{3}{2\gamma^2 R^2_{\min}}\right)$ for all i. However, these bounds are not accurate enough for us. In the following lemmas we obtain bounds which give the explicit dependence on p for p-orthogonal data. The proofs of the lemmas follow similar arguments to the proof from Frei et al. (2023), with some required modifications.

Lemma 17 *Denote $R^2_{min} = \min_i \|x_i\|^2$, $R^2_{max} = \max_i \|x_i\|^2$, and $R^2 = R^2_{max}/R^2_{min}$. Let f denote the leaky ReLU network (2) and let W denote a KKT point of Problem (3). Suppose the training*

data are p-orthogonal for $p \geq 3\gamma^{-3}$. Using the notation from Eq. (15) and (16), for all $i \in [n]$ we have

$$\sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} \leq \frac{m}{R_{min}^2 \left(\gamma - \frac{1}{pR^2}\right)} ,$$

and

$$\lambda_i \leq \frac{1}{R_{min}^2 \gamma \left(\gamma - \frac{1}{pR^2}\right)} .$$

Proof Let $\xi = \max_{q\in[n]} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j}\right)$ and suppose that $\xi > \frac{m}{R_{\min}^2\left(\gamma - \frac{1}{pR^2}\right)}$. Let $r = \operatorname{argmax}_{q\in[n]} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j}\right)$. Since by our assumption $p \geq 3\gamma^{-3} \geq \frac{3}{\gamma}$ and $R \geq 1$, then $\xi > 0$ and therefore $\lambda_r > 0$. Hence, by the KKT conditions we must have $y_r f(x_r; W) = 1$.

We consider two cases:

Case 1: Assume that $y_r = -1$. Using (15) and (16), we have

$$\begin{aligned}
\sqrt{m} f(x_r; W) &= \sum_{j\in[m/2]} \phi(v_j^\top x_r) - \sum_{j\in[m/2]} \phi(u_j^\top x_r) \\
&= \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \sum_{q\in[n]} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right) - \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \sum_{q\in[n]} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right) \\
&= \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \lambda_r y_r \phi'_{r,v_j} x_r^\top x_r + \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right) \\
&\quad - \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \lambda_r (-y_r) \phi'_{r,u_j} x_r^\top x_r + \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right) \\
&\leq \sum_{j\in[m/2]} \phi\left(-\frac{1}{\sqrt{m}} \lambda_r \phi'_{r,v_j} R_{\min}^2 + \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right) \\
&\quad - \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \lambda_r \phi'_{r,u_j} R_{\min}^2 + \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right) .
\end{aligned}$$

Since the derivative of ϕ is lower bounded by γ, we know $\phi(z_1) - \phi(z_2) \geq \gamma(z_1 - z_2)$ for all $z_1, z_2 \in \mathbb{R}$. Using this and the definition of ξ, the above is at most

$$\begin{aligned}
&\sum_{j\in[m/2]} \left[\phi\left(\frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r \right) - \frac{1}{\sqrt{m}} \gamma \lambda_r \phi'_{r,v_j} R_{\min}^2 \right] \\
&\quad - \sum_{j\in[m/2]} \left[\phi\left(\frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r \right) + \frac{1}{\sqrt{m}} \gamma \lambda_r \phi'_{r,u_j} R_{\min}^2 \right] \\
&\leq -\frac{1}{\sqrt{m}} \gamma \xi R_{\min}^2 + \sum_{j\in[m/2]} \left| \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r \right| + \sum_{j\in[m/2]} \left| \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r \right| \\
&\leq -\frac{1}{\sqrt{m}} \gamma \xi R_{\min}^2 + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \sum_{q\in[n]\setminus\{r\}} \left| \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r \right| + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \sum_{q\in[n]\setminus\{r\}} \left| \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r \right| .
\end{aligned}$$

Using $|x_q^\top x_r| \leq \zeta$ for $q \neq r$, the above is at most

$$\begin{aligned}
&-\frac{1}{\sqrt{m}} \gamma \xi R_{\min}^2 + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \sum_{q\in[n]\setminus\{r\}} \lambda_q \phi'_{q,v_j} \zeta + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \sum_{q\in[n]\setminus\{r\}} \lambda_q \phi'_{q,u_j} \zeta \\
&= -\frac{1}{\sqrt{m}} \gamma \xi R_{\min}^2 + \frac{\zeta}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j} \right) \\
&\leq -\frac{1}{\sqrt{m}} \gamma \xi R_{\min}^2 + \frac{\zeta}{\sqrt{m}} \cdot n \cdot \max_{q\in[n]} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j} \right) \\
&= -\frac{1}{\sqrt{m}} \gamma \xi R_{\min}^2 + \frac{\zeta}{\sqrt{m}} n \xi \\
&= -\frac{\xi}{\sqrt{m}} (\gamma R_{\min}^2 - n\zeta) .
\end{aligned}$$

By our p-orthogonality assumption, the above expression is at most

$$-\frac{\xi}{\sqrt{m}} \left(\gamma R_{\min}^2 - \frac{R_{\min}^2}{pR^2} \right) = -\frac{\xi R_{\min}^2}{\sqrt{m}} \left(\gamma - \frac{1}{pR^2} \right) < -\frac{m}{R_{\min}^2 \left(\gamma - \frac{1}{pR^2} \right)} \cdot \frac{R_{\min}^2}{\sqrt{m}} \left(\gamma - \frac{1}{pR^2} \right) = -\sqrt{m} ,$$

where in the inequality we used the assumption on ξ, the assumption $p \geq 3\gamma^{-3} \geq \frac{3}{\gamma}$, and $R \geq 1$. Thus, we obtain $f(x_r; W) < -1$ in contradiction to $y_r f(x_r; W) = 1$.

Case 2: Assume that $y_r = 1$. A similar calculation to the one given in case 1 (which we do not repeat for conciseness) implies that $f(x_r; W) > 1$, in contradiction to $y_r f(x_r; W) = 1$. It concludes the proof of $\xi \leq \frac{m}{R_{\min}^2 \left(\gamma - \frac{1}{pR^2} \right)}$.

Finally, since $\xi \leq \frac{m}{R_{\min}^2 \left(\gamma - \frac{1}{pR^2} \right)}$ and the derivative of ϕ is lower bounded by γ, then for all $i \in [n]$ we have

$$\frac{m}{R_{\min}^2 \left(\gamma - \frac{1}{pR^2} \right)} \geq \sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} \geq m \lambda_i \gamma ,$$

and hence $\lambda_i \leq \frac{1}{R_{\min}^2 \gamma\left(\gamma - \frac{1}{pR^2}\right)}$. ■

Lemma 18 *Denote $R_{min}^2 = \min_i \|x_i\|^2$, $R_{max}^2 = \max_i \|x_i\|^2$, and $R^2 = R_{max}^2/R_{min}^2$. Let f denote the leaky ReLU network (2) and let W denote a KKT point of Problem (3). Suppose the training data are p-orthogonal for $p \geq 3\gamma^{-3}$. Using the notation from Eq. (15) and (16), for all $i \in [n]$ we have*

$$\sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} \geq \frac{m(\gamma p R^2 - 2)}{R_{max}^2(\gamma p R^2 - 1)} ,$$

and

$$\lambda_i \geq \frac{\gamma p R^2 - 2}{R_{max}^2(\gamma p R^2 - 1)} = \frac{1}{R_{max}^2}\left(1 - \frac{1}{\gamma p R^2 - 1}\right) .$$

Proof Suppose that there is $i \in [n]$ such that $\sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} < \frac{m(\gamma p R^2 - 2)}{R_{\max}^2(\gamma p R^2 - 1)}$. Using (15) and (16), we have

$$\begin{aligned}
\sqrt{m} \leq \left|\sqrt{m} f(x_i; W)\right| &= \left| \sum_{j\in[m/2]} \phi(v_j^\top x_i) - \sum_{j\in[m/2]} \phi(u_j^\top x_i) \right| \leq \sum_{j\in[m/2]} \left|v_j^\top x_i\right| + \sum_{j\in[m/2]} \left|u_j^\top x_i\right| \\
&= \sum_{j\in[m/2]} \left| \frac{1}{\sqrt{m}} \sum_{q\in[n]} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_i \right| + \sum_{j\in[m/2]} \left| \frac{1}{\sqrt{m}} \sum_{q\in[n]} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_i \right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \left(\left|\lambda_i y_i \phi'_{i,v_j} x_i^\top x_i\right| + \sum_{q\in[n]\setminus\{i\}} \left|\lambda_q y_q \phi'_{q,v_j} x_q^\top x_i\right| \right) \\
&\quad + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \left(\left|\lambda_i (-y_i) \phi'_{i,u_j} x_i^\top x_i\right| + \sum_{q\in[n]\setminus\{i\}} \left|\lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_i\right| \right) .
\end{aligned}$$

Using $|x_q^\top x_i| \leq \zeta$ for $q \neq i$ and $x_i^\top x_i \leq R_{\max}^2$, the above is at most

$$\begin{aligned}
&\frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \left(\lambda_i \phi'_{i,v_j} R_{\max}^2 + \sum_{q\in[n]\setminus\{i\}} \lambda_q \phi'_{q,v_j} \zeta \right) + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \left(\lambda_i \phi'_{i,u_j} R_{\max}^2 + \sum_{q\in[n]\setminus\{i\}} \lambda_q \phi'_{q,u_j} \zeta \right) \\
&= \frac{1}{\sqrt{m}} \left(\sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} R_{\max}^2 + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} R_{\max}^2 \right) + \\
&\quad \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{i\}} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} \zeta + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j} \zeta \right) \\
&= \frac{R_{\max}^2}{\sqrt{m}} \left(\sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} \right) + \frac{\zeta}{\sqrt{m}} \sum_{q\in[n]\setminus\{i\}} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j} \right) \\
&< \frac{R_{\max}^2}{\sqrt{m}} \cdot \frac{m(\gamma p R^2 - 2)}{R_{\max}^2(\gamma p R^2 - 1)} + \frac{\zeta}{\sqrt{m}} \cdot n \cdot \max_{q\in[n]} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j} \right) ,
\end{aligned}$$

where in the last inequality we used our assumption on i. Combining the above with our p-orthogonality assumption $n\zeta \leq \frac{R_{\min}^2}{pR^2}$, we get

$$\begin{aligned}\max_{q\in[n]}\left(\sum_{j\in[m/2]}\lambda_q\phi'_{q,v_j}+\sum_{j\in[m/2]}\lambda_q\phi'_{q,u_j}\right) &> m\left(1-\frac{\gamma pR^2-2}{\gamma pR^2-1}\right)\cdot\frac{1}{n\zeta}\\ &\geq m\left(1-\frac{\gamma pR^2-2}{\gamma pR^2-1}\right)\cdot\frac{pR^2}{R_{\min}^2}\\ &= m\left(\frac{1}{\gamma pR^2-1}\right)\cdot\frac{pR^2}{R_{\min}^2}\\ &= \frac{m}{R_{\min}^2}\left(\frac{1}{\gamma-\frac{1}{pR^2}}\right)\,,\end{aligned}$$

in contradiction to Lemma 17. It concludes the proof of $\sum_{j\in[m/2]}\lambda_i\phi'_{i,v_j}+\sum_{j\in[m/2]}\lambda_i\phi'_{i,u_j}\geq\frac{m(\gamma pR^2-2)}{R_{\max}^2(\gamma pR^2-1)}$.

Finally, since $\sum_{j\in[m/2]}\lambda_i\phi'_{i,v_j}+\sum_{j\in[m/2]}\lambda_i\phi'_{i,u_j}\geq\frac{m(\gamma pR^2-2)}{R_{\max}^2(\gamma pR^2-1)}$ and the derivative of ϕ is upper bounded by 1, then for all $i\in[n]$ we have

$$\frac{m(\gamma pR^2-2)}{R_{\max}^2(\gamma pR^2-1)}\leq\sum_{j\in[m/2]}\lambda_i\phi'_{i,v_j}+\sum_{j\in[m/2]}\lambda_i\phi'_{i,u_j}\leq m\lambda_i\,,$$

and hence

$$\lambda_i\geq\frac{\gamma pR^2-2}{R_{\max}^2(\gamma pR^2-1)}=\frac{1}{R_{\max}^2}\left(1-\frac{1}{\gamma pR^2-1}\right)\,.$$

■

Combining Eq. (17) with Lemmas 17 and 18, and letting $s_i=\frac{(1+\gamma)\lambda_i}{2}$ for all $i\in[n]$, we get that $\operatorname{sign}\left(f(x;W)\right)=\operatorname{sign}\left(\langle z,x\rangle\right)$ for $z=\sum_{i=1}^n s_iy_ix_i$, where for all $i\in[n]$ we have

$$s_i\in\left[\frac{(1+\gamma)}{2}\cdot\frac{1}{R_{\max}^2}\left(1-\frac{1}{\gamma pR^2-1}\right),\frac{(1+\gamma)}{2}\cdot\frac{1}{R_{\min}^2\gamma^2\left(1-\frac{1}{\gamma pR^2}\right)}\right]\,.$$

Therefore, z is τ-uniform with

$$\begin{aligned}\tau&=\frac{(1+\gamma)}{2}\cdot\frac{1}{R_{\min}^2\gamma^2\left(1-\frac{1}{\gamma pR^2}\right)}\cdot\frac{2}{(1+\gamma)}\cdot R_{\max}^2\left(1-\frac{1}{\gamma pR^2-1}\right)^{-1}\\ &=\frac{\gamma pR^2}{R_{\min}^2\gamma^2(\gamma pR^2-1)}\cdot R_{\max}^2\frac{\gamma pR^2-1}{\gamma pR^2-2}\\ &=\frac{R_{\max}^2}{R_{\min}^2\gamma^2}\cdot\frac{\gamma pR^2}{\gamma pR^2-2}\\ &=\frac{R^2}{\gamma^2}\left(1+\frac{2}{\gamma pR^2-2}\right)\,.\end{aligned}$$

Appendix C. Proofs for Sub-Gaussian Marginals

In this section we prove Lemma 5 and Theorem 6 as well as Corollary 7, Corollary 8, and Corollary 9. A rough outline of our proof strategy is as follows.

1. First, we show in Lemma 19 that in order for a linear classifier $x \mapsto \text{sign}(\langle w, x \rangle)$ to achieve a test error near the noise rate, it suffices for $\|[\Sigma^{1/2}w]_{2:d}\| / \sqrt{\lambda_1}[w]_1$ to be small.

2. Next, we show in Lemma 20 a number of properties of the training data $\{(x_i, y_i)\}_{i=1}^n$ that hold with high probability under Assumptions (SG1) through (SG3). Lemma 5 will hold as a deterministic consequence of this lemma, so that the training data are p-orthogonal for large p (recall Definition 1) and the norms of all of the examples are close to each other. This allows for us to apply Proposition 3 and Proposition 4, which show that the KKT points of both the linear max-margin problem (1) and the leaky ReLU max-margin (3) are τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$.

3. By the first step, to prove Theorem 6, it suffices to show that a τ-uniform $w \in \mathbb{R}^d$ is such that:

   (i) the norm $\|[\Sigma^{1/2}w]_{2:d}\|$ is small, and

   (ii) the first component $[w]_1$ is large and positive.

   Recall that a τ-uniform w takes the form $\sum_{i=1}^n s_i y_i x_i$ where $\max_{i,j} s_i/s_j \leq \tau$. For (i), Lemma 20 provides bounds on $\|[\Sigma^{1/2}x_i]_{2:d}\|$ for training examples x_i, which is the basic building block to this part. For (ii), note that for clean examples $i \in \mathcal{C} \subset [n]$ (where $y_i = \tilde{y}_i$), $[s_i y_i x_i]_1 = s_i |[x_i]_1|$, while for noisy examples $i \in \mathcal{N} \subset [n]$ (where $y_i = -\tilde{y}_i$), $[s_i y_i x_i]_1 = -s_i |[x_i]_1|$. Thus, it suffices to characterize the following,

$$\begin{aligned}\left[\sum_{i=1}^n s_i y_i x_i\right]_1 &= \sum_{i \in \mathcal{C}} s_i |[x_i]_1| - \sum_{i \in \mathcal{N}} s_i |[x_i]_1| \\ &= \sum_{i=1}^n s_i |[x_i]_1| - 2\sum_{i \in \mathcal{N}} s_i |[x_i]_1| \\ &\geq \min_i s_i \sum_{i=1}^n |[x_i]_1| - 2 \max_i s_i \sum_{i \in \mathcal{N}} |[x_i]_1|.\end{aligned}$$

   Lemma 23 directly bounds each of the terms above. The proof of Theorem 6 is then a direct calculation based on (i) and (ii) above.

4. Corollaries 7 and 8 follow by a direct calculation based on Lemma 5 and Theorem 6. Corollary 9 follows by a direct calculation that verifies $\mathsf{P}_{\mathsf{gaus}}$ satisfies the required properties.

C.1. Preliminary concentration inequalities

We first show that the test error of any linear classifier $w \in \mathbb{R}^d$ satisfying $[w]_1 > 0$ is close to the noise rate whenever $\|[\Sigma^{1/2}w]_{2:d}\| / \sqrt{\lambda_1}[w]_1$ is small.

Lemma 19 *There exists an absolute constant $c_1 \geq 2$ such that provided $w \in \mathbb{R}^d$ is such that $[w]_1 > 0$, then the following holds. If $[\Sigma^{1/2}w]_{2:d} = 0$ then $\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) \leq \eta$. Otherwise,*

$$\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) \leq \eta + \frac{c_1\|[\Sigma^{1/2}w]_{2:d}\|}{\sqrt{\lambda_1}[w]_1} \cdot \left(1 + \sqrt{0 \vee \log\left(\frac{\sqrt{\lambda_1}[w]_1}{\|[\Sigma^{1/2}w]_{2:d}\|}\right)}\right).$$

Proof By definition of P_{sg}, we have,

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) &= \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(y\langle w, x\rangle < 0) \\
&= \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(y\langle w, x\rangle < 0,\ y = -\operatorname{sign}([x]_1)) \\
&\quad + \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(y\langle w, x\rangle < 0,\ y = \operatorname{sign}([x]_1)) \\
&\leq \eta + \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(y\langle w, x\rangle < 0,\ y = \operatorname{sign}([x]_1)). \qquad (18)
\end{aligned}$$

Denoting $[u]_{2:d} \in \mathbb{R}^{d-1}$ as the last $d-1$ components of the vector $u \in \mathbb{R}^d$, we have,

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(y\langle w, x\rangle < 0,\ y = \operatorname{sign}([x]_1)) &= \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(|[x]_1|[w]_1 < -\operatorname{sign}([x]_1)\langle [w]_{2:d}, [x]_{2:d}\rangle) \\
&\overset{(i)}{=} \mathbb{P}(\sqrt{\lambda_1}[w]_1|[z]_1| < -\operatorname{sign}([z]_1)\langle [w]_{2:d}, [\Sigma^{1/2}z]_{2:d}\rangle) \\
&\overset{(ii)}{=} \mathbb{P}\left(|[z]_1| < -\frac{\operatorname{sign}([z]_1)\langle [\Sigma^{1/2}w]_{2:d}, [z]_{2:d}\rangle}{\sqrt{\lambda_1}[w]_1}\right). \qquad (19)
\end{aligned}$$

Equality (i) uses that $x = \Sigma^{1/2}z$. Equality (ii) uses the assumption that $[w]_1 > 0$. From here, we see that if $[\Sigma^{1/2}w]_{2:d} = 0$ then we have $\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(y\langle w, x\rangle < 0,\ y = \operatorname{sign}([x]_1)) = \mathbb{P}(|[z]_1| < 0) = 0$, which by (18) shows that $\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) \leq \eta$. Thus in the remainder of the proof we shall assume $[\Sigma^{1/2}w]_{2:d} \neq 0$.

Let us define the term

$$\rho := \frac{\operatorname{sign}([z]_1)\langle [\Sigma^{1/2}w]_{2:d}, [z]_{2:d}\rangle}{\sqrt{\lambda_1}[w]_1}.$$

This term is small in absolute value when $\sqrt{\lambda_1}[w]_1 \gg \|[\Sigma^{1/2}w]_{2:d}\|$. In particular, since z is a sub-Gaussian random vector with sub-Gaussian norm at most σ_z, by Hoeffding's inequality we have that for some $c > 0$ and any $t \geq 0$,

$$\mathbb{P}(|\rho| \geq t) = \mathbb{P}\left(\frac{|\langle [z]_{2:d}, [\Sigma^{1/2}w]_{2:d}\rangle|}{\sqrt{\lambda_1}[w]_1} \geq t\right) \leq 2\exp\left(-\frac{c\lambda_1[w]_1^2t^2}{\sigma_z^2\|[\Sigma^{1/2}w]_{2:d}\|^2}\right). \qquad (20)$$

Continuing from (19), we get for any $t \geq 0$,

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(y\langle w, x\rangle < 0,\ y = \operatorname{sign}([x]_1)) &= \mathbb{P}(|[z]_1| < -\rho) \\
&= \mathbb{P}(|[z]_1| < -\rho,\ |\rho| \geq t) + \mathbb{P}(|[z]_1| < -\rho, |\rho| < t) \\
&\leq \mathbb{P}(|\rho| \geq t) + \mathbb{P}(|[z]_1| < t) \\
&\overset{(i)}{\leq} 2\exp\left(-\frac{c\lambda_1[w]_1^2t^2}{\sigma_z^2\|[\Sigma^{1/2}w]_{2:d}\|^2}\right) + \beta t.
\end{aligned}$$

In inequality (i) we have used (20) as well as the assumption that $\mathbb{P}(|[z]_1| \le t) \le \beta t$ for any $t \ge 0$. In particular, if we let

$$\xi := \frac{\|[\Sigma^{1/2}w]_{2:d}\|}{\sqrt{\lambda_1}[w]_1}, \quad t := c^{-1/2}\sigma_z \xi \sqrt{0 \vee \log(1/\xi)},$$

then we have,

$$\begin{aligned}\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}(y\langle w, x\rangle < 0,\ y = \operatorname{sign}([x]_1)) &\le 2\exp\left(-\frac{ct^2}{\sigma_z^2\xi^2}\right) + \beta t \\ &= 2(1 \wedge \xi) + \xi\beta c^{-1/2}\sigma_z\sqrt{0 \vee \log(1/\xi)}) \\ &\le \xi \cdot \max(2, \beta c^{-1/2}\sigma_z)(1 + \sqrt{0 \vee \log(1/\xi)}).\end{aligned}$$

The proof follows by letting $c_1 = \max(2, \beta c^{-1/2}\sigma_z)$. ■

The following lemma characterizes a number of useful properties about the training data.

Lemma 20 *There exists an absolute constant $C_0 > 1$ such that for every large enough $C > 1$ (with C, C_0 depending only on σ_z) and for any $\delta \in (0, 1/2)$, under Assumptions (SG1) through (SG3) (defined for these C and δ), the following holds with probability at least $1 - 2\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$:*

1. *The norms of the samples satisfy,*

$$\left|\frac{\|x_i\|}{\sqrt{\operatorname{tr}(\Sigma)}} - 1\right| \le C_0\sqrt{\frac{\|\Sigma\|_2 \log(6n/\delta)}{\operatorname{tr}(\Sigma)}}, \quad \textit{for all } i \in [n],$$

   and

$$\left|\frac{\|[\Sigma^{1/2}x_i]_{2:d}\|}{\sqrt{\operatorname{tr}(\Sigma_{2:d}^2)}} - 1\right| \le C_0\sqrt{\frac{\|\Sigma_{2:d}^2\|_2 \log(6n/\delta)}{\operatorname{tr}(\Sigma_{2:d}^2)}}.$$

2. *The correlations of distinct samples satisfy,*

$$|\langle x_i, x_j\rangle| \le C_0\sqrt{\operatorname{tr}(\Sigma^2)}\log(6n^2/\delta), \quad \textit{for all } i \ne j.$$

3. *The samples satisfy,*

$$\min_i \|x_i\|^2 \ge \frac{\operatorname{tr}(\Sigma)}{C_0\sqrt{\operatorname{tr}(\Sigma^2)}\log(6n^2/\delta)} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot \max_{i\ne j} |\langle x_i, x_j\rangle|.$$

Proof

We prove the lemma in parts.

Part 1: norms of samples. We first show concentration of the norms. Let $x \in \{x_1, \dots, x_n\}$. Recall that $x = \Sigma^{1/2} z$ where the components of z are independent, mean-zero, and z has sub-Gaussian norm at most σ_z, with $\mathbb{E}[zz^\top] = I_d$. We can thus apply Hanson-Wright inequality (Vershynin, 2018, Theorem 6.3.2), so that there is an absolute constant $c > 0$ such that we have for any $t \geq 0$,

$$\mathbb{P}\left(\left|\|x\| - \sqrt{\operatorname{tr}(\Sigma)}\right| > t\right) \leq 2\exp\left(-\frac{ct^2}{\sigma_z^4 \|\Sigma\|_2}\right),$$

where we have used that $\|\Sigma^{1/2}\|_F^2 = \sum_{i=1}^d \lambda_i = \operatorname{tr}(\Sigma)$. Choosing $t = c^{-1/2}\sigma_z^2\sqrt{\|\Sigma\|_2 \log(6n/\delta)}$ and using a union bound over $x \in \{x_1, \dots, x_n\}$, we see that,

$$\mathbb{P}\left(\exists i \in [n] : \left|\|x_i\| - \sqrt{\operatorname{tr}(\Sigma)}\right| > c^{-1/2}\sigma_z^2\sqrt{\|\Sigma\|_2 \log(6n/\delta)}\right) < \delta/3. \tag{21}$$

We now show a bound on $\|\Sigma^{1/2}x_i\|^2 = z_i^\top \Sigma^2 z_i$. Again fix $x \in \{x_1, \dots, x_n\}$. We can employ a nearly identical argument to above: since $\|\Sigma\|_F^2 = \operatorname{tr}(\Sigma^2)$, by Hanson-Wright inequality, for any $t \geq 0$,

$$\mathbb{P}\left(\left|\|\Sigma^{1/2}x\| - \sqrt{\operatorname{tr}(\Sigma^2)}\right| > t\right) \leq 2\exp\left(-\frac{ct^2}{\sigma_z^4 \|\Sigma^2\|_2}\right), \tag{22}$$

Choosing $t = c^{-1/2}\sigma_z^2\sqrt{\|\Sigma^2\|_2 \log(6n/\delta)}$ and noting that $\operatorname{tr}(\Sigma^2) \geq \|\Sigma^2\|_2$ implies

$$\sqrt{\operatorname{tr}(\Sigma^2)} + c^{-1/2}\sigma_z^2\sqrt{\|\Sigma^2\|_2 \log(6n/\delta)} \leq (1 + c^{-1/2}\sigma_z^2)\sqrt{\operatorname{tr}(\Sigma^2)\log(6n/\delta)},$$

by a union bound we get

$$\mathbb{P}\left(\exists i \in [n] : \|\Sigma^{1/2}x_i\| > (1 + c^{-1/2}\sigma_z^2)\sqrt{\operatorname{tr}(\Sigma^2)\log(6n/\delta)}\right) \leq \delta/3. \tag{23}$$

Using a completely identical argument used to derive (21), we also have

$$\mathbb{P}\left(\exists i \in [n] : \left|\|[\Sigma^{1/2}x_i]_{2:d}\| - \sqrt{\operatorname{tr}(\Sigma_{2:d}^2)}\right| > c^{-1/2}\sigma_z^2\sqrt{\|\Sigma_{2:d}^2\|\log(6n/\delta)}\right) \leq \delta/3. \tag{24}$$

Part 2: correlations of samples. We now bound the correlation between distinct samples. Let us fix $j \in [n]$ and consider $i \in [n] \setminus \{j\}$. Then there is a sub-Gaussian random vector z_i with sub-Gaussian norm at most σ_z such that $\langle x_i, x_j\rangle = z_i^\top \Sigma^{1/2} x_j$. In particular, $\langle x_i, x_j\rangle = z_i^\top \xi \cdot \|\Sigma^{1/2}x_j\|$ where ξ is a unit-norm vector. Since z_i is a sub-Gaussian random vector, this means that for some $c > 0$ and any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(|\langle x_i, x_j\rangle| > t \middle| \|\Sigma^{1/2}x_j\| \leq (1 + c^{-1/2}\sigma_z^2)\sqrt{\operatorname{tr}(\Sigma^2)\log(6n/\delta)}\right) \\
&\quad\leq 2\exp\left(-c \cdot \frac{t^2}{\sigma_z^2(1 + c^{-1/2}\sigma_z^2)^2 \operatorname{tr}(\Sigma^2)\log(6n/\delta)}\right).
\end{aligned} \tag{25}$$

Letting $c' = c/[\sigma_z^2(1 + c^{-1/2}\sigma_z^2)^2]$, we can thus bound,

$$\begin{aligned}
&\mathbb{P}\left(\exists i \neq j : |\langle x_i, x_j\rangle| > t\right) \\
&\quad\leq \mathbb{P}\left(\exists i \neq j : |\langle x_i, x_j\rangle| > t \middle| \|\Sigma^{1/2}x_j\| \leq (1 + c^{-1/2}\sigma_z^2)\sqrt{\operatorname{tr}(\Sigma^2)\log(6n/\delta)}\right) \\
&\qquad + \mathbb{P}\left(\exists j : \|\Sigma^{1/2}x_j\| > (1 + c^{-1/2}\sigma_z^2)\sqrt{\operatorname{tr}(\Sigma^2)\log(6n/\delta)}\right) \\
&\quad\overset{(i)}{\leq} 2n^2\exp\left(-\frac{c't^2}{\operatorname{tr}(\Sigma^2)\log(6n/\delta)}\right) + \frac{\delta}{3},
\end{aligned} \tag{26}$$

where (i) uses (23). Choosing $t = (c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta)$ and using (25), we get

$$\mathbb{P}\left(\exists i \neq j : \ |\langle x_i, x_j\rangle| > (c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta)\right) \leq \frac{2\delta}{3}. \tag{27}$$

Combining the above display with (21) and (24) and using a union bound, we get that with probability at least $1 - 2\delta$,

$$\begin{cases} \left|\frac{\|x_i\|}{\sqrt{\mathrm{tr}(\Sigma)}} - 1\right| \leq c^{-1}\sigma_z^2\sqrt{\frac{\|\Sigma\|_2\log(6n/\delta)}{\mathrm{tr}(\Sigma)}}, & \text{for all } i \in [n], \\ \left|\frac{\|[\Sigma^{1/2}x_i]_{2:d}\|}{\sqrt{\mathrm{tr}(\Sigma^2_{2:d})}} - 1\right| \leq c^{-1}\sigma_z^2\sqrt{\frac{\|\Sigma^2_{2:d}\|_2\log(6n/\delta)}{\mathrm{tr}(\Sigma^2_{2:d})}}, & \text{for all } i \in [n], \\ |\langle x_i, x_j\rangle| \leq (c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta), & \text{for all } i \neq j. \end{cases} \tag{28}$$

This completes the first two parts of the lemma.

Part 3: near-orthogonality of samples. We now show an upper bound on $R = \max_{i,j}\|x_i\|/\|x_j\|$. Note that with probability at least $1 - 2\delta$, (28) holds, and we shall show that this implies that for an absolute constant $C_0 > 0$ we have,

$$\frac{\min_i \|x_i\|^2}{R^2\max_{i\neq j}|\langle x_i, x_j\rangle|} \geq \frac{\mathrm{tr}(\Sigma)}{C_0\sigma_z\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta)}.$$

By Assumption (SG3) we have

$$\xi := \frac{\|\Sigma\|_2\log(6n/\delta)}{\mathrm{tr}(\Sigma)} \leq \frac{\|\Sigma\|_F\log(6n/\delta)}{\mathrm{tr}(\Sigma)} = \frac{\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n/\delta)}{\mathrm{tr}(\Sigma)} \leq \frac{1}{C}. \tag{29}$$

Thus by (28), we see that the quantity $R = \max_{i,j}\|x_i\|/\|x_j\|$ satisfies

$$\begin{aligned} R &= \max_{i,j}\frac{\|x_i\|}{\|x_j\|} \\ &\leq \frac{1 + c^{-1}\sigma_z^2\sqrt{\xi}}{1 - c^{-1}\sigma_z^2\sqrt{\xi}} \\ &\overset{(i)}{\leq} \frac{1 + c^{-1}\sigma_z^2/\sqrt{C}}{1 - c^{-1}\sigma_z^2/\sqrt{C}} \\ &\overset{(ii)}{\leq} \left(1 + \frac{2c^{-1}\sigma_z^2}{\sqrt{C}}\right)^2. \end{aligned} \tag{30}$$

Inequality (i) follows by (29), while (ii) uses the inequality $1/(1-x) \leq 1 + 2x$ on $[0, 1/2]$ and holds for $C > 1$ large enough. In particular, by taking C larger we can guarantee R is closer to one.

Next, we have by part 1 and part 2 of this lemma,

$$\frac{\min_i \|x_i\|^2}{\max_{i\neq j}|\langle x_i, x_j\rangle|} \geq \frac{\mathrm{tr}(\Sigma)\left(1 - c^{-1}\sigma_z^2\sqrt{\xi}\right)}{(c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta)} \geq \frac{\mathrm{tr}(\Sigma)\left(1 - c^{-1}\sigma_z^2/\sqrt{C}\right)}{(c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta)}.$$

For $C > 100c^{-2}\sigma_z^4$, by (30) we have $1 - c^{-1}\sigma_z^2/\sqrt{C} \geq 0.9$ and $R^{-2} \geq 0.68$. We therefore see that for C large enough,

$$\frac{\min_i \|x_i\|^2}{R^2 \max_{i\neq j} |\langle x_i, x_j\rangle|} \geq \frac{\mathrm{tr}(\Sigma)}{2(c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta)}.$$

■

C.2. Proof of Lemma 5

We next prove Lemma 5: as C grows in Assumptions (SG1) through (SG3), the training data become p-orthogonal for large p and $\max_{i,j} \|x_i\|^2/\|x_j\|^2 \to 1$.

Lemma 21 *There exists an absolute constant $C_1 > 0$ (depending only on σ_z) such that for every large enough constant $C > 0$, for any $\delta \in (0, 1/2)$, under Assumptions (SG1) through (SG3) (defined for these C and δ), with probability at least $1 - 2\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$, the training data is C/C_1-orthogonal, and $\max_{i,j} \|x_i\|^2/\|x_j\|^2 \leq (1 + C_1/\sqrt{C})^4$.*

Proof First, note that all of the results in Lemma 20 hold with probability at least $1 - 2\delta$. We shall show that the training data being C/C_1-orthogonal and that $\max_{i,j} \|x_i\|^2/\|x_j\|^2$ are a deterministic consequence of this high-probability event. By Lemma 20,

$$\min_i \|x_i\|^2 \geq \frac{\mathrm{tr}(\Sigma)}{C_0\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta)} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot \max_{i\neq j} |\langle x_i, x_j\rangle|.$$

By Assumption (SG3), this means

$$\min_i \|x_i\|^2 \geq \frac{C}{C_0} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot n \max_{i\neq j} |\langle x_i, x_j\rangle|.$$

In particular, the training data is C/C_0-orthogonal (see Definition 1).

For the ratio $R = \max_{i,j} \|x_i\|/\|x_j\|$, if we let $\xi := \frac{\|\Sigma\|_2 \log(6n/\delta)}{\mathrm{tr}(\Sigma)}$ then by part 1 of Lemma 20 we have

$$\sqrt{\mathrm{tr}(\Sigma)}(1 - C_0\sqrt{\xi}) \leq \|x_i\| \leq \sqrt{\mathrm{tr}(\Sigma)}(1 + C_0\sqrt{\xi}).$$

By Assumption (SG3) we know $\xi \leq 1/C$ (see (29)). Therefore for $C > 1$ large enough, using $1/(1-x) \leq 1 + 2x$ for $x \in [0, 1/2]$,

$$R = \max_{i,j} \frac{\|x_i\|}{\|x_j\|} \leq \frac{1 + C_0\sqrt{\xi}}{1 - C_0\sqrt{\xi}} \leq \frac{1 + C_0/\sqrt{C}}{1 - C_0/\sqrt{C}} \leq \left(1 + \frac{2C_0}{\sqrt{C}}\right)^2.$$

This completes the claimed upper bound for R. ■

C.3. Proof of Theorem 6

We now begin to prove that if w is τ-uniform, i.e. there are strictly positive s_i, $i = 1, \dots, n$ such that $w = \sum_{i=1}^n s_i y_i x_i$ and $\max_{i,j} {}^{s_i}/_{s_j} \leq \tau$, then the first component of w is large and positive. By Lemma 19, this is one step towards showing the test error of this linear predictor is close to the noise rate.

To begin, note that since $y_i = \operatorname{sign}([x_i]_1)$ for $i \in \mathcal{C}$ and $y_i = -\operatorname{sign}([x_i]_1)$ for $i \in \mathcal{N}$, we have,

$$\left[\sum_{i=1}^n s_i y_i x_i\right]_1 = \sum_{i\in\mathcal{C}} s_i |[x_i]_1| - \sum_{i\in\mathcal{N}} s_i |[x_i]_1| = \sum_{i=1}^n s_i |[x_i]_1| - 2\sum_{i\in\mathcal{N}} s_i |[x_i]_1|.$$

Thus, in order to show that this quantity is large and positive, we would like to show the first term is large and positive while the second term is not too negative. We do so in the following lemma.

Lemma 22 *There exists a universal constant $C_1' > 1$ (depending only on η and σ_z) such that for any $\delta \in (0, 1/3)$, if $n \geq C_1' \log(2/\delta)$ then with probability at least $1 - 3\delta$ over the training data $\{(x_i, y_i)\}_{i=1}^n \sim \mathsf{P}_{\mathsf{sg}}^n$, the following holds:*

$$\sum_{i=1}^n |[x_i]_1| \geq n\sqrt{\lambda_1}\,\mathbb{E}[|[z]_1|]\left(1 - C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right), \quad \textit{and}$$

$$\sum_{i\in\mathcal{N}} |[x_i]_1| \leq n\sqrt{\lambda_1}\,\mathbb{E}[|[z]_1|]\left(\eta + C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right).$$

Proof By definition, there are i.i.d. $z_i \sim \mathsf{P}_z$ such that $x_i = \Sigma^{1/2} z_i$. In particular, $[x_i]_1 = \sqrt{\lambda_1}[z_i]_1$, so thus it suffices to bound the sum $\sum_{i=1}^n |[z_i]_1| = \lambda_1^{-1/2} \sum_{i=1}^n |[x_i]_1|$ from below and the sum $\sum_{i\in\mathcal{N}} |[z_i]_1|$ from above.

Let us denote $\alpha := \mathbb{E}\,|[z_i]_1|$. Note that since $\mathbb{P}(|[z]_1| \leq t) \leq \beta t$, by taking $t = 1/(2\beta)$ we see that

$$\alpha = \mathbb{E}\,|[z_i]_1| \geq \frac{1}{4\beta}. \tag{31}$$

The quantity $|[z_i]_1| - \alpha$ with sub-Gaussian norm at most $c_1\sigma_z$ for some absolute constant $c_1 > 0$ (Vershynin, 2018, Lemma 2.6.8), and is i.i.d. over indices $i \in [n]$. Therefore, by Hoeffding's inequality, this means that for some absolute constant $c > 0$ and any $t \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (|[z_i]_1| - \alpha)\right| \geq t\right) \leq 2\exp\left(-\frac{cnt^2}{\sigma_z^2}\right).$$

Choosing $t = c^{-1/2}\sigma_z\sqrt{\log(2/\delta)/n}$ and using (31) we get that with probability at least $1 - \delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n (|[z_i]_1| - \alpha)\right| \leq c^{-1/2}\sigma_z\sqrt{\frac{\log(2/\delta)}{n}} \implies \sum_{i=1}^n |[z_i]_1| \geq n\alpha\left(1 - 4c^{-1/2}\sigma_z\beta\sqrt{\frac{\log(2/\delta)}{n}}\right). \tag{32}$$

Using the same argument (and assuming without loss of generality that $|\mathcal{N}| > 0$, since otherwise we can just ignore this term entirely), we get with probability at least $1 - \delta$,

$$\left|\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} (|[z_i]_1| - \alpha)\right| \leq c^{-1/2}\sigma_z\sqrt{\frac{\log(2/\delta)}{|\mathcal{N}|}}. \tag{33}$$

From here we see it is necessary to control the number of noisy points. The number of noisy points $|\mathcal{N}|$ is the sum of n independent, identically distributed random variables with mean η. Thus, by Hoeffding's inequality, for any $u \geq 0$,

$$\mathbb{P}\left(||\mathcal{N}| - n\eta| \geq u\right) \leq 2\exp\left(-\frac{2u^2}{n}\right).$$

In particular, selecting $u = \sqrt{n\log(2/\delta)/2}$, we see that with probability at least $1-\delta$,

$$\left|\frac{|\mathcal{N}|}{n} - \eta\right| \leq \sqrt{\frac{\log(2/\delta)}{n}}.$$

Rearranging we see that

$$\eta n - \sqrt{n\log(2/\delta)} \leq |\mathcal{N}| \leq \eta n + \sqrt{n\log(2/\delta)}.$$

Since η is an absolute constant, using the lemma's assumption that $n \geq C_1' \log(2/\delta)$ we get for C_1' large enough relative to η^{-2},

$$\eta n - \sqrt{n\log(2/\delta)} = \eta n\left(1 - \sqrt{\frac{\eta^{-2}\log(2/\delta)}{n}}\right) \geq \frac{1}{2}\eta n,$$

and therefore

$$\frac{1}{2}\eta n \leq |\mathcal{N}| \leq \eta n + \sqrt{n\log(2/\delta)} \tag{34}$$

Substituting the two previous displays into (33) we get

$$\begin{aligned}
\sum_{i\in\mathcal{N}} |[z_i]_1| &\leq |\mathcal{N}|\alpha\left(1 + c^{-1/2}\sigma_z\alpha^{-1}\sqrt{\frac{\log(2/\delta)}{|\mathcal{N}|}}\right) \\
&\leq n\alpha\left(\eta + \sqrt{\frac{\log(2/\delta)}{n}}\right)\cdot\left(1 + 4c^{-1/2}\sigma_z\beta\sqrt{\frac{2\eta^{-1}\log(2/\delta)}{n}}\right) \\
&\leq n\alpha\left(\eta + 12c^{-1/2}\sigma_z\beta\sqrt{\frac{2\eta^{-1}\log(2/\delta)}{n}}\right).
\end{aligned}$$

The second inequality uses (31). The last inequality uses the lemma's assumption that $n \geq C_1'\log(2/\delta)$ for a large enough C_1' and that η^{-1} is an absolute constant. Taking a union bound over the three events and taking C_1' large enough completes the proof since σ_z and η are absolute constants. ■

We now show that a τ-uniform classifier u has a large and positive first component while $\|[\Sigma^{1/2}u]_{2:d}\|$ is small with high probability. By Lemma 27, this suffices for showing generalization error near the noise rate.

Lemma 23 *Let $\tau \geq 1$ be a constant, and suppose $\eta \leq \frac{1}{2\tau} - \Delta$ for some absolute constants $\eta, \Delta > 0$. There exists an absolute constant $C > 1$ (depending only on $\eta, \sigma_z, \beta, \tau$, and Δ) such that for any*

$\delta \in (0, 1/5)$, under Assumptions (SG1) through (SG3) (defined for these C and δ), with probability at least $1 - 5\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$, if $u = \sum_{i=1}^n s_i y_i x_i \in \mathbb{R}^d$ is τ-uniform w.r.t $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathsf{sg}}$ then

$$[u]_1 \geq \frac{\tau \Delta n \alpha \sqrt{\lambda_1}}{8\beta} \left(\min_i s_i \right), \quad \textit{and} \quad \|[\Sigma^{1/2} u]_{2:d}\| \leq \frac{3}{2} n \left(\max_i s_i \right) \sqrt{\operatorname{tr}(\Sigma_{2:d}^2)}.$$

In particular, if $[\Sigma^{1/2} u]_{2:d} \neq 0$ then

$$\frac{[u]_1}{\|[\Sigma^{1/2} u]_{2:d}\|} \geq \frac{\Delta}{12\beta} \cdot \sqrt{\frac{\lambda_1}{\operatorname{tr}(\Sigma_{2:d}^2)}}.$$

Proof First, by a union bound, for C sufficiently large the results of both Lemma 22 and Lemma 20 hold with probability at least $1 - 5\delta$. In the remainder of the proof we will work on this high-probability event and we shall show the lemma holds as a deterministic consequence of this.

By definition, $y_i = \operatorname{sign}([x_i]_1)$ for $i \in \mathcal{C}$ and $y_i = -\operatorname{sign}([x_i]_1)$ for $i \in \mathcal{N}$. Since u is τ-uniform, there exist strictly positive numbers s_i such that $u = \sum_{i=1}^n s_i y_i x_i$ with $\max_{i,j} \frac{s_i}{s_j} \leq \tau$. Thus we can write,

$$\begin{aligned}
[u]_1 = \left[\sum_{i=1}^n s_i y_i x_i \right]_1 &= \sum_{i \in \mathcal{C}} s_i |[x_i]_1| - \sum_{i \in \mathcal{N}} s_i |[x_i]_1| \\
&= \sum_{i=1}^n s_i |[x_i]_1| - 2 \sum_{i \in \mathcal{N}} s_i |[x_i]_1| \\
&\geq \min_i s_i \sum_{i=1}^n |[x_i]_1| - 2 \max_i s_i \sum_{i \in \mathcal{N}} |[x_i]_1|. \qquad (35)
\end{aligned}$$

Let us denote $\alpha := \mathbb{E}\,|[z]_1|$. Recall by (31) that the assumption of anti-concentration on $[z]_1$ implies $\alpha \geq 1/(4\beta)$. Now using Lemma 22, we have,

$$\begin{aligned}
[u]_1 &\geq n\alpha\sqrt{\lambda_1} \left(\min_i s_i \right) \left(1 - C_1' \beta \sqrt{\frac{\log(2/\delta)}{n}} \right) \\
&\quad - n\alpha\sqrt{\lambda_1} \left(\max_i s_i \right) \left(2\eta + 2C_1' \beta \sqrt{\frac{\log(2/\delta)}{n}} \right) \\
&\geq n\alpha\sqrt{\lambda_1} \left(\min_i s_i \right) \left[1 - C_1' \beta \sqrt{\frac{\log(2/\delta)}{n}} - \tau \left(2\eta + 2C_1' \beta \sqrt{\frac{\log(2/\delta)}{n}} \right) \right] \\
&\overset{(i)}{\geq} n\alpha\sqrt{\lambda_1} \left(\min_i s_i \right) \left[1 - C_1' \beta \sqrt{\frac{\log(2/\delta)}{n}} - \tau \left(\frac{1}{\tau} - 2\Delta + 2C_1' \beta \sqrt{\frac{\log(2/\delta)}{n}} \right) \right].
\end{aligned}$$

The inequality (i) uses the lemma's assumption that $\eta \leq 1/(2\tau) - \Delta$. Rearranging the above, we see

$$\begin{aligned}
[u]_1 &\geq n\alpha\sqrt{\lambda_1} \left(\min_i s_i \right) \left[2\tau\Delta - C_1' \beta \sqrt{\frac{\log(2/\delta)}{n}} - 2\tau C_1' \beta \sqrt{\frac{\log(2/\delta)}{n}} \right] \\
&\geq \frac{1}{2} \tau \Delta n \alpha \sqrt{\lambda_1} \left(\min_i s_i \right) > 0. \qquad (36)
\end{aligned}$$

The final inequality uses that τ, Δ, β are absolute constants and by taking C large enough so that $n \geq C \log(2/\delta)$ implies the inequality.

Next, we want to bound $\|[\Sigma^{1/2}u]_{2:d}\|$. We will use the first part of Lemma 20 to do so. We have,

$$\begin{aligned}
\left\|\left[\Sigma^{1/2}u\right]_{2:d}\right\| &= \left\|\left[\textstyle\sum_{i=1}^n s_i y_i \Sigma^{1/2} x_i\right]_{2:d}\right\| \\
&\leq n\left(\max_i s_i\right) \max_i \left\|\left[\Sigma^{1/2}x_i\right]_{2:d}\right\| \\
&\leq n\left(\max_i s_i\right)\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}\left(1 + C_0\sqrt{\frac{\|\Sigma_{2:d}^2\|_2 \log(6n/\delta)}{\mathrm{tr}(\Sigma_{2:d}^2)}}\right) \\
&= n\left(\max_i s_i\right)\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}\left(1 + C_0\sqrt{\frac{\log(6n/\delta)}{\mathsf{StableRank}(\Sigma_{2:d})}}\right) \\
&\leq \frac{3}{2}n\left(\max_i s_i\right)\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}. \qquad (37)
\end{aligned}$$

The final inequality uses Assumption (SG2) so that $\mathsf{StableRank}(\Sigma_{2:d}) > C\log(6n/\delta)$ and follows by taking C large enough. Putting (36) and the above together, if $[\Sigma^{1/2}u]_{2:d} \neq 0$ we get

$$\frac{[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|} \geq \frac{1}{3}\tau\Delta\alpha \cdot \frac{\min_i s_i}{\max_i s_i} \cdot \sqrt{\frac{\lambda_1}{\mathrm{tr}(\Sigma_{2:d}^2)}} \geq \frac{\Delta\alpha}{3} \cdot \sqrt{\frac{\lambda_1}{\mathrm{tr}(\Sigma_{2:d}^2)}}.$$

Since by (31) we have $\alpha \geq 1/(4\beta)$, this completes the proof. ■

We are now in a position to prove Theorem 6. For the reader's convenience, we re-state it below.

Theorem 6 *Let $\tau \geq 1$ be a constant, and suppose $\eta \leq \frac{1}{2\tau} - \Delta$ for some absolute constants $\eta, \Delta > 0$. There exist constants $C, C' > 0$ (depending only on $\eta, \sigma_z, \beta, \tau$, and Δ) such that for any $\delta \in (0, 1/7)$, under Assumptions (SG1) through (SG3) (defined for these C and δ), with probability at least $1 - 7\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$, if $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$, then*

$$\text{for all } k \in [n], \quad y_k = \mathrm{sign}\left(\langle u, x_k\rangle\right), \qquad \text{while simultaneously,}$$

$$\eta \leq \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}\left(y \neq \mathrm{sign}(\langle u, x\rangle)\right) \leq \eta + C'\sqrt{\frac{\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\mathrm{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

In particular, if $\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$, then the linear classifier $x \mapsto \mathrm{sign}(\langle u, x\rangle)$ exhibits benign overfitting.

Proof By a union bound, with probability at least $1 - 7\delta$, the results of both Lemma 23 and Lemma 20 hold, and we showed previously that Lemma 5 is a deterministic consequence of Lemma 20 and the Assumptions (SG1) through (SG3). In the remainder of the proof we will work on this event and show that the theorem holds as a consequence of these lemmas and Assumptions (SG1) through (SG3).

Since u is τ-uniform, there exist strictly positive constants s_i such that $u = \sum_{i=1}^n s_i y_i x_i$. We first show that u interpolates the training data: for any $k \in [n]$ we have

$$\begin{aligned}
\langle u, y_k x_k \rangle &= s_k \|x_k\|^2 + \sum_{i \neq k} \langle s_i y_i x_i, y_k x_k \rangle \\
&\geq s_k \|x_k\|^2 - n \max_i s_i \cdot \max_{i \neq j} |\langle x_i, x_j \rangle| \\
&= s_k \|x_k\|^2 \left(1 - \frac{n \max_i s_i \cdot \max_{i \neq j} |\langle x_i, x_j \rangle|}{s_k \|x_k\|^2}\right) \\
&\geq s_k \|x_k\|^2 \left(1 - \frac{n\tau \max_{i \neq j} |\langle x_i, x_j \rangle|}{\|x_k\|^2}\right) \\
&\overset{(i)}{\geq} s_k \|x_k\|^2 \left(1 - \frac{C_1 \tau}{C}\right) \\
&\overset{(ii)}{\geq} \frac{1}{2} s_k \|x_k\|^2. \qquad (38)
\end{aligned}$$

The inequality (i) uses that the training data is C/C_1-orthogonal by Lemma 5, while (ii) follows by taking $C \geq 2C_1\tau$. This last quantity is strictly positive by Lemma 20. Thus, u interpolates the training data.

We now show the generalization error is close to the noise rate. By Lemma 19, if $[\Sigma^{1/2}u]_{2:d} = 0$ then since $[u]_1 > 0$ by Lemma 23, we have $\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle u, x \rangle)\big) \leq \eta$ and the proof is complete.

Thus consider the case that $[\Sigma^{1/2}u]_{2:d} \neq 0$. Let $c := \Delta/(12\beta)$, where $c < 1$ is an absolute constant (assuming w.l.o.g. $\beta \geq 1$) as Δ, β are absolute constants by assumption. Then by Lemma 23 we have,

$$\frac{[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|} \geq \frac{\Delta}{12\beta} \sqrt{\frac{\lambda_1}{\operatorname{tr}(\Sigma_{2:d}^2)}} = c \sqrt{\frac{\lambda_1}{\operatorname{tr}(\Sigma_{2:d}^2)}}. \qquad (39)$$

Applying Lemma 19 there exists $c_1 \geq 2$ such that

$$\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle u, x \rangle)\big) \leq \eta + \frac{c_1 \|[\Sigma^{1/2}u]_{2:d}\|}{\sqrt{\lambda_1}[u]_1} \left(1 + \sqrt{0 \vee \log\left(\frac{\sqrt{\lambda_1}[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|}\right)}\right). \qquad (40)$$

We now consider two cases.

Case 1: $c^{-1}\sqrt{\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2} \leq 1/2$. Since the function $\xi \mapsto \xi(1 + \sqrt{\log(1/\xi)})$ is monotone increasing on the interval $(0, 1/2]$, for any $\xi, \xi' \in [0, 1/2]$ satisfying $\xi \leq \xi'$ we have $\xi(1 + \sqrt{\log(1/\xi)}) \leq \xi'(1 + \sqrt{\log(1/\xi')})$. By (39) and the case assumption, we have

$$\frac{\|[\Sigma^{1/2}u]_{2:d}\|}{\sqrt{\lambda_1}[u]_1} \leq \sqrt{\frac{c^{-2}\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}} \leq \frac{1}{2}. \qquad (41)$$

Thus continuing from (40),

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle u, x\rangle)\big) &\leq \eta + \frac{c_1 \|[\Sigma^{1/2}u]_{2:d}\|}{\sqrt{\lambda_1}[u]_1}\left(1 + \sqrt{\log\left(0 \vee \frac{\sqrt{\lambda_1}[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|}\right)}\right) \\
&\overset{(i)}{=} \eta + \frac{c_1 \|[\Sigma^{1/2}u]_{2:d}\|}{\sqrt{\lambda_1}[u]_1}\left(1 + \sqrt{\log\left(\frac{\sqrt{\lambda_1}[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|}\right)}\right) \\
&\overset{(ii)}{\leq} \eta + c_1\sqrt{\frac{c^{-2}\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{\log\left(\sqrt{\frac{\lambda_1^2}{c^{-2}\operatorname{tr}(\Sigma_{2:d}^2)}}\right)}\right) \\
&\overset{(iii)}{\leq} \eta + c_1 c^{-1}\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).
\end{aligned}$$

Equality (i) uses that $\log(x) \geq 0$ for $x \geq 1$. Inequality (ii) uses (41). The final inequality (iii) uses that $c < 1$ and $a \leq a \vee b$ for any $a, b \in \mathbb{R}$.

Case 2: $c^{-1}\sqrt{\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2} > 1/2$**.** In this case it is trivially true that

$$\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle u, x\rangle)\big) \leq \eta + c_1 c^{-1}\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right),$$

since $c_1 \geq 2$ and $c^{-1}\sqrt{\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2} > 1/2$ the right-hand-side is at least 1. From this we see that the theorem follows by taking $C' = c_1 c^{-1}$. ■

C.4. Proof of Corollary 7, Corollary 8, and Corollary 9

This section contains proofs of Corollary 7, Corollary 8, and Corollary 9.

Corollary 24 *Suppose $0 < \eta \leq 0.49$. There exist constants $C, C' > 0$ such that for any $\delta \in (0, 1/9)$, under Assumptions (SG1) through (SG3) (defined for these C and δ), with probability at least $1 - 9\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$, the max-margin linear classifier $w = \operatorname{argmin}\{\|w\|^2 : y_i\langle w, x_i\rangle \geq 1\, \forall i\}$ satisfies*

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(\langle w, x_k\rangle\big), \qquad \textit{while simultaneously,}$$

$$\eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) \leq \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

In particular, if $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$ then w exhibits benign overfitting.

Proof By a union bound, both Theorem 6 and Lemma 5 hold with probability at least $1 - 9\delta$ and any τ-uniform linear classifier exhibits benign overfitting in the sense described in the theorem, with the noise tolerance determined by τ. Thus, we need only verify that working on this high-probability event and using the assumptions, the linear max-margin solution is τ-uniform and that τ is small.

By Lemma 5, the training data is C/C_1-orthogonal and we have the following upper bound for R^2,

$$R^2 = \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \le \left(1 + \frac{C_1}{\sqrt{C}}\right)^4 \le \frac{100}{99}. \tag{42}$$

The last inequality follows by taking C to be a large enough absolute constant. Therefore Proposition 3 ensures that the linear max-margin w is τ-uniform with $\tau = R^2\left(1 + \frac{2}{pR^2-2}\right)$. In particular,

$$\tau \le R^2\left(1 + \frac{2}{CR^2/C_1 - 2}\right) \le \frac{201}{198}.$$

The final inequality uses that $R^2 \le 100/99$ and by taking $C > 1$ large enough. Thus the max-margin linear classifier is τ-uniform with $\tau \le \frac{201}{198}$. Since $\frac{1}{2\tau} \ge \frac{198}{402} \ge 0.492$, if $\eta \le 0.49 = 0.492 - 0.002$ we can apply Theorem 6. ■

Corollary 25 *Suppose that $0 < \eta \le \frac{49\gamma^2}{100}$. There exist constants $C, C' > 0$ such that for any $\delta \in (0, 1/9)$, under Assumptions (SG1) through (SG3) (defined for these C and δ), with probability at least $1 - 9\delta$ over $\mathsf{P}_{\mathsf{sg}}^n$, any KKT point W of Problem (3) satisfies*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\big(f(x_k; W)\big), \quad \text{while simultaneously,}$$
$$\eta \le \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}\Big(y \ne \operatorname{sign}\big(f(x;W)\big)\Big) \le \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

In particular, if $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$ then the neural network $f(x;W)$ exhibits benign overfitting. Moreover, for any initialization $W(0)$, gradient flow converges in direction to a network satisfying the above.

Proof Just as in the proof of the preceding corollary, by a union bound, with probability at least $1 - 9\delta$ both Theorem 6 and Lemma 5 hold and any τ-uniform linear classifier exhibits benign overfitting with probability, with the noise tolerance determined by τ. By Lemma 5, the training data is C/C_1-orthogonal, and thus for $C > 3C_1\gamma^{-3}$, we may apply Proposition 4 so that $\operatorname{sign}(f(x;W)) = \operatorname{sign}(\langle z, x\rangle)$ where z is τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$ for $\tau = R^2\gamma^{-2}\left(1 + \frac{2}{\gamma CR^2/C_1 - 2}\right)$. Lemma 5 also implies that $R^2 \le (1 + C_1/\sqrt{C})^4 \le \frac{100}{99}$ for C large enough. Hence, for C large enough, $\tau \le \frac{201}{198}\gamma^{-2}$. Since $\frac{1}{2\tau} \ge \frac{198\gamma^2}{402} > 0.492\gamma^2$, if $\eta \le 0.49\gamma^2 = 0.492\gamma^2 - 0.002\gamma^2$ we may apply Theorem 6 since γ is an absolute constant. ■

Corollary 26 *Suppose $0 < \eta \le \frac{49\gamma^2}{100}$. Then for the distribution $\mathsf{P}_{\mathsf{gaus}}$, for any $\delta \in (0, 1/9)$, if $\rho \in (1/2, 1)$, $d = \tilde{\Omega}(n^{1/(1-\rho)})$, and $n = \tilde{\Omega}(1)$, then Assumptions (SG1) through (SG3) are satisfied. Moreover, with probability at least $1 - 9\delta$ over $\mathsf{P}_{\mathsf{gaus}}^n$, KKT points of Problem (3) exhibit benign overfitting:*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\big(f(x_k; W)\big),$$
$$\text{while simultaneously,} \quad \eta \le \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{gaus}}}\Big(y \ne \operatorname{sign}\big(f(x_k;W)\big)\Big) \le \eta + \tilde{O}\left(d^{\frac{1}{2}(1-2\rho)}\right) = \eta + o_d(1).$$

Furthermore, for any initialization $W(0)$, gradient flow converges in direction to a network satisfying the above.

Proof First, it is clear that $\mathsf{P}_{\mathsf{gaus}}$ is an instance of P_{sg}, since $\Sigma^{-1/2}x$ is an isotropic Gaussian which clearly satisfies the anti-concentration property $\mathbb{P}(|[z]_1| \leq t) \leq \beta t$ for $\beta = 1/\sqrt{2\pi}$. We thus need only verify that assumptions (SG1) through (SG3) are satisfied and that $\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2$ is small. Clearly, $\mathsf{StableRank}(\Sigma_{2:d}) = d - 1$ and

$$\frac{\mathrm{tr}(\Sigma)}{\sqrt{\mathrm{tr}(\Sigma^2)}} = \frac{d^\rho + d - 1}{\sqrt{d^{2\rho} + d - 1}}.$$

By assumption, $\rho \in (1/2, 1)$, so $d^\rho + d - 1 = \Theta(d)$, while $d^{2\rho} + d - 1 = \Theta(d^{2\rho})$. Therefore,

$$\frac{\mathrm{tr}(\Sigma)}{\sqrt{\mathrm{tr}(\Sigma^2)}} = \Theta(d^{1-\rho}).$$

Thus, we see that if $n = \tilde{\Omega}(1)$ and $d = \tilde{\Omega}(n^{1/(1-\rho)})$, then assumptions (SG1) through (SG3) are satisfied and hence Theorem 6 and Corollary 8 apply under the stated assumptions on the noise rate η. On the other hand,

$$\frac{\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2} = \frac{d-1}{d^{2\rho}} = \Theta(d^{1-2\rho}).$$

Since $\rho > 1/2$, we see that $\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o_d(1)$, and thus the test error of KKT points of Problem (3) are at most

$$\eta + C'\sqrt{\frac{\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\mathrm{tr}(\Sigma_{2:d}^2)}\right)}\right) = \eta + \tilde{O}(d^{\frac{1}{2}(1-2\rho)}) = \eta + o_d(1).$$

■

Appendix D. Proofs for clustered data

In this section we provide the proofs for Section 5. Our proof strategy mirrors that we used for the proof of Theorem 6 in Appendix C, and can be summarized as follows:

1. We first show that in order for a linear classifier $x \mapsto \mathrm{sign}(\langle w, x\rangle)$ to achieve small test error, it suffices to have $\langle w, y^{(q)}\mu^{(q)}\rangle$ be large and positive for each $q \in Q$.

2. Propositions 3 and 4 show that the max-margin solutions for linear classifiers and leaky ReLU networks correspond to τ-uniform classifiers when the training data is p-orthogonal. To use this result, we thus need to characterize the norms and pairwise correlations of the examples. Additionally, note that if $w \in \mathbb{R}^d$ is τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$, then $w = \sum_{i=1}^n s_i y_i x_i$ for some $s_i > 0$. Thus by the first step above, we see it will be helpful to characterize $\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle$ for samples $i \in [n]$ and clusters $q \in Q$. Lemma 28 provides some initial bounds that help us with these goals, and Lemma 29 collects all of the important properties of the training data that we will use. In particular, Lemma 10 will follow from Lemma 29, and the test error bound in Theorem 11 for τ-uniform classifiers will crucially rely on this lemma as well.

3. We then prove Theorem 11 by utilizing the above properties.
4. The proofs of Corollaries 12 and 13 then follow by a direct calculation.

D.1. Preliminary concentration inequalities

Our first lemma provides a generalization bound for any linear classifier over $\mathsf{P}_{\mathsf{clust}}$.

Lemma 27 *There exists an absolute constant $c > 0$ such that if $w \in \mathbb{R}^d$ is such that $\langle w, y^{(q)}\mu^{(q)}\rangle \geq 0$ for each $q \in [k]$, then*

$$\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) \leq \eta + \frac{1}{k}\sum_{q=1}^{k} \exp\left(-c\frac{\langle w, \mu^{(q)}\rangle^2}{\|w\|^2}\right).$$

Proof We use an identical proof to that of Lemma 19. By definition of $\mathsf{P}_{\mathsf{clust}}$, we have $y = \tilde{y}$ (the 'clean' label) with probability $1-\eta$ while $y = -\tilde{y}$ with probability η. Thus we can calculate,

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) &= \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0) \\
&= \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0,\ y = -\tilde{y}) \\
&\quad + \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0,\ y = \tilde{y}) \\
&\leq \eta + \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0,\ y = \tilde{y}). \qquad (43)
\end{aligned}$$

We can bound the second term above as follows,

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0,\ y = \tilde{y}) &= \frac{1}{k}\sum_{q=1}^{k} \mathbb{P}_{z\sim\mathsf{P}'_z}(\langle w, y^{(q)}\mu^{(q)} + y^{(q)}z\rangle < 0) \\
&= \frac{1}{k}\sum_{q=1}^{k} \mathbb{P}_{z\sim\mathsf{P}'_z}(\langle w, y^{(q)}z\rangle < -y^{(q)}\langle w, \mu^{(q)}\rangle) \\
&\leq \frac{1}{k}\sum_{q=1}^{k} \exp\left(-c\frac{\langle w, \mu^{(q)}\rangle^2}{\|w\|^2}\right).
\end{aligned}$$

In the last inequality we have used that $y^{(q)}\langle w, \mu^{(q)}\rangle \geq 0$, as well as the fact that $y^{(q)}z$ is sub-Gaussian (with sub-Gaussian norm at most the absolute constant σ_z) and Hoeffding's inequality. Substituting the above into (43) completes the proof. ∎

Due to Proposition 3 and 4, we are interested in the behavior of classifiers defined in terms of $w \in \mathbb{R}^d$ that are τ-uniform w.r.t. the training data. Such classifiers take the form $\sum_{i=1}^n s_i y_i x_i$, where $s_i > 0$. By Lemma 27, to show $x \mapsto \operatorname{sign}(\langle w, x\rangle)$ has small generalization error, it therefore helpful to characterize $\langle y_i x_i, \mu^{(q)}\rangle$ for different clusters q. We begin to do so with the following lemma.

Lemma 28 *Let P'_z be a distribution such that the components of $z \sim \mathsf{P}'_z$ are mean-zero, independent, sub-Gaussian random variables with sub-Gaussian norm at most one; and for some absolute constant $\kappa > 0$, $\kappa d \leq \mathbb{E}[\|z\|^2] \leq d$. Let $\delta \in (0,1)$. Suppose that $\{z_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{P}'_z$, and let $v_1, \ldots, v_k$ be any collection of vectors in $\mathbb{R}^d$. There are absolute constants $C, C_1 > 1$ such that provided $d \geq C\log(n/\delta)$, the following hold with probability at least $1 - 4\delta$.*

(i) For all i,

$$\kappa d\left(1 - C_1\sqrt{\frac{\kappa^{-2}\log(2n/\delta)}{d}}\right) \le \|z_i\|^2 \le d\left(1 + C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right).$$

(ii) For all $i \neq j$, $|\langle z_i, z_j\rangle| \le C_1\sqrt{d\log(2n/\delta)}$.

(iii) For all $i = 1,\dots,k$ and $j = 1,\dots,n$, $|\langle v_i, z_j\rangle| \le C_1\|v_i\|\sqrt{\log(2nk/\delta)}$.

Proof We prove the lemma in parts. We use an identical argument to Chatterji and Long (2021, Lemma 16).

For the first part, fix $i \in [n]$. The quantity $\|z_i\|^2$ is a sum of d independent random variables that are squares of sub-Gaussian random variables with norm at most one, and thus by Vershynin (2018, Lemma 2.7.6), this is the sum of d sub-exponential random variables with sub-exponential norm at most one. Thus by Bernstein's inequality (see (Vershynin, 2018, Theorem 2.8.1)), there is some absolute constant $c > 0$ such that for any $t \ge 0$,

$$\mathbb{P}(|\|z_i\|^2 - \mathbb{E}\,\|z_i\|^2| \ge t) \le 2\exp\left(-c\left(t \wedge \frac{t^2}{d}\right)\right).$$

Choosing $t = c^{-1}\sqrt{d\log(2n/\delta)}$, we see that

$$d \ge c^{-2}\log(2n/\delta) \implies t \wedge t^2/d = c^{-2}\log(2n/\delta).$$

Thus, we have

$$\mathbb{P}\left(\exists i : \left|\|z_i\|^2 - \mathbb{E}[\|z_i\|^2]\right| \ge c^{-1}\sqrt{d\log(2n/\delta)}\right) \le \delta.$$

By assumption, $\kappa d \le \mathbb{E}[\|z_i\|^2] \le d$. Using

$$\kappa d\left(1 - c^{-1}\sqrt{\frac{\kappa^{-2}\log(2n/\delta)}{d}}\right) = \kappa d - c^{-1}\sqrt{d\log(2n/\delta)},$$

$$d + c^{-1}\sqrt{d\log(2n/\delta)} = d\left(1 + c^{-1}\sqrt{\frac{\log(2n/\delta)}{d}}\right),$$

we thus have

$$\mathbb{P}\left(\exists i : \kappa d\left(1 - c^{-1}\sqrt{\frac{\kappa^{-2}\log(2n/\delta)}{d}}\right) > \|z_i\|^2 \text{ or } \|z_i\|^2 > d\left(1 + c^{-1}\sqrt{\frac{\log(2n/\delta)}{d}}\right)\right) \le \delta. \tag{44}$$

Next, note that for any $i, j \in [n]$, and any $t \ge 0$,

$$\mathbb{P}(|\langle z_i, z_j\rangle| \ge t) \le \mathbb{P}\left(|\langle z_i, z_j\rangle| \ge t \middle| \|z_j\| \le \sqrt{2d}\right) + \mathbb{P}(\|z_j\| > \sqrt{2d}).$$

For $i \neq j$, conditional on z_j, since z_i has independent sub-Gaussian components with sub-Gaussian norm at most one, the random variable $\langle z_i, z_j\rangle$ is mean-zero sub-Gaussian with sub-Gaussian norm

at most $c_1\|z_j\|$ for an absolute constant $c_1 > 0$ (Vershynin, 2018, Proposition 2.6.1). Thus by Hoeffding's inequality (Vershynin, 2018, Theorem 2.6.3) we have for some absolute constant $c_2 > 0$,

$$\mathbb{P}\left(|\langle z_i, z_j\rangle| \geq t \middle| \|z_j\| \leq \sqrt{2d}\right) \leq 2\exp\left(-c_2 \cdot \frac{t^2}{2d}\right).$$

Letting $t = c_2^{-1/2}\sqrt{2d\log(2n^2/\delta)}$ and we see that

$$\mathbb{P}\left(|\langle z_i, z_j\rangle| \geq c_2^{-1/2}\sqrt{2d\log(2n^2/\delta)} \middle| \|z_j\| \leq \sqrt{2d}\right) \leq \frac{\delta}{n^2}.$$

Using this and (44),

$$\begin{aligned}
&\mathbb{P}(\text{ for some } i \neq j, |\langle z_i, z_j\rangle| \geq c_2^{-1/2}\sqrt{2d\log(2n^2/\delta)}) \\
&\quad \leq n^2\mathbb{P}\left(|\langle z_i, z_j\rangle| \geq c_2^{-1/2}\sqrt{2d\log(2n^2/\delta)} \middle| \|z_j\| \leq \sqrt{2d}\right) + \mathbb{P}(\text{for some } j \in [n],\ \|z_j\| > \sqrt{2d}) \\
&\quad \leq 2\delta. \qquad (45)
\end{aligned}$$

In the last inequality we are using the lemma's assumption that $d \geq 4c^{-2}\log(2n/\delta)$ so that $\{\|z_i\|^2 > \sqrt{2}d\} \subset \{\|z_i\|^2 > d(1 + c^{-1}\sqrt{\log(2n/\delta)/d})\}$.

Finally, for $v \in \{v_1, \ldots, v_k\}$ and fixed j, since z_j has independent sub-Gaussian components we know $\langle z_j, v\rangle$ is a sub-Gaussian random variable with sub-Gaussian norm at most $c_1\|v\|$. Therefore, by Hoeffding's inequality we have for some constant $c_3 > 0$,

$$\mathbb{P}\left(|\langle z_j, v\rangle| \geq t\right) \leq 2\exp\left(-c_3 \cdot \frac{t^2}{\|v\|^2}\right).$$

Taking $t = c_3^{-1}\|v\|\sqrt{\log(2nk/\delta)}$ and a union bound over $j \in [n]$ and the k possible options for v, we see that

$$\mathbb{P}\left(\exists j \in [n], v \in \{v_1, \ldots, v_k\} \text{ s.t. } |\langle z_j, v\rangle| \geq c_3^{-1}\|v\|\sqrt{\log(2nk/\delta)}\right) \leq \delta.$$

Using a union bound with (44), (45) and the above yields a total failure probability of 4δ and completes the proof. ■

Next, we show how to use the above to say something about the training data. Recall that we observe samples $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathsf{clust}}$ which are noisy versions of $\{(x_i, \tilde{y}_i)\}_{i=1}^n$. We denote by $\mathcal{C} \subset [n]$ the clean samples and $\mathcal{N} \subset [n]$ the noisy examples, so that $\mathcal{C} \cup \mathcal{N} = [n] = I$. In particular, for $i \in \mathcal{N}$, $y_i = -\tilde{y}_i$, while for $i \in \mathcal{C}$, $y_i = \tilde{y}_i$. We further use the notation $\text{cluster}(i) = q_i$ and $I^{(q)} = \{i \in I : \text{cluster}(i) = q\}$ and

$$I_{\mathcal{C}}^{(q)} := \{i \in I \cap \mathcal{C} : \text{cluster}(i) = q\}, \quad I_{\mathcal{N}}^{(q)} := \{i \in I \cap \mathcal{N} : \text{cluster}(i) = q\},$$

so that $I^{(q)} = I_{\mathcal{C}}^{(q)} \cup I_{\mathcal{N}}^{(q)}$.

Lemma 29 *There is an absolute constant $C_1' > 1$ such that the following holds. For $C > 1$ sufficiently large under Assumptions (CL1) through (CL4), with probability at least $1 - 7\delta$, items (i) through (iii) of Lemma 28 hold (with $v_i = \mu^{(i)}$ for $i = 1, \ldots, k$), and we have the following.*

(i) For all i,

$$d\left(1 - C_1'\sqrt{\frac{\log(2n/\delta)}{d}}\right) \le \|x_i\|^2 \le d\left(1 + C_1'\sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right).$$

(ii) For each $q \in Q$ and $i \in I^{(q)}$,

$$\left|\langle \mu^{(q)}, x_i\rangle - \|\mu^{(q)}\|^2\right| \le C_1'\|\mu^{(q)}\|\sqrt{\log(2nk/\delta)}.$$

(iii) For each $q \in Q$, if $i, j \in I^{(q)}$ and $i \neq j$, then

$$\left|\langle x_i, x_j\rangle - \|\mu^{(q)}\|^2\right| \le C_1'\sqrt{d\log(2n/\delta)}.$$

(iv) For each $q, r \in Q$ with $q \neq r$, if $i \in I^{(q)}$ and $j \in I^{(r)}$, then

$$|\langle x_i, x_j\rangle| \le \max_{q\neq r}|\langle \mu^{(q)}, \mu^{(r)}\rangle| + C_1'\sqrt{d\log(2n/\delta)}.$$

(v) For all $q \in Q$,

$$\left|\frac{|I^{(q)}|}{n} - \frac{1}{k}\right| \le \sqrt{\frac{\log(2k/\delta)}{n}},$$

and

$$\left|\frac{|I^{(q)}_{\mathcal{N}}|}{|I^{(q)}|} - \eta\right| \le \sqrt{\frac{\log(2k/\delta)}{n}}, \quad \left|\frac{|I^{(q)}_{\mathcal{C}}|}{|I^{(q)}|} - (1-\eta)\right| \le \sqrt{\frac{\log(2k/\delta)}{n}}.$$

Proof By definition,

$$\|x_i\|^2 = \|z_i\|^2 + \|\mu^{(q_i)}\|^2 + 2\langle z_i, \mu^{(q_i)}\rangle.$$

We first note that since $d \ge Cn^2\log(n/\delta)$ by Assumption (CL2), with probability at least $1 - 4\delta$, all of the results of Lemma 28 hold, where $v_1, \ldots, v_k$ are taken to be the cluster means, $v_i = \mu^{(i)}$. We work on this high-probability event in the remainder of the proof.

By definition, since we have assumed $\mathbb{E}[\|z\|^2] = d$,

$$\begin{aligned}
\|x_i\|^2 &= \|z_i\|^2 + \|\mu^{(q_i)}\|^2 + 2\langle z_i, \mu^{(q_i)}\rangle \\
&\ge d\left(1 - C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right) + \|\mu^{(q_i)}\|^2 - 2C_1\|\mu^{(q_i)}\|\sqrt{\log(2nk/\delta)} \\
&\ge d\left(1 - C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right).
\end{aligned}$$

where we have used Assumption (CL3) (for $C > 1$ large enough) in the last inequality. On the other hand, by Lemma 28 we also have,

$$\begin{aligned}\|x_i\|^2 &= \|z_i\|^2 + \|\mu^{(q_i)}\|^2 + 2\langle z_i, \mu^{(q_i)}\rangle \\ &\le d\left(1 + C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right) + \|\mu^{(q_i)}\|^2 + 2C_1\|\mu^{(q_i)}\|\sqrt{\log(2nk/\delta)} \\ &\overset{(i)}{\le} d\left(1 + C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right) + 2\|\mu^{(q_i)}\|^2 \\ &\overset{(ii)}{\le} d\left(1 + C_1\sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right).\end{aligned}$$

The inequality (i) uses Assumption (CL3) and inequality (ii) uses Assumption (CL2).

For the second part of the lemma, note that for $i \in I^{(q)}$, $\langle \mu^{(q)}, x_i\rangle - \|\mu^{(q)}\|^2 = \langle z_i, \mu^{(q)}\rangle$. Lemma 28 thus bounds the absolute value of this quantity.

For the third part of the lemma, consider those $i \neq j$ that belong to the same cluster. For these, we have $\mu^{(q_i)} = \mu^{(q_j)}$ so that

$$\begin{aligned}\langle x_i, x_j\rangle &= \langle \mu^{(q_i)} + z_i, \mu^{(q_j)} + z_j\rangle \\ &= \|\mu^{(q_i)}\|^2 + \langle z_i, \mu^{(q_j)}\rangle + \langle \mu^{(q_j)}, z_i\rangle + \langle z_i, z_j\rangle.\end{aligned}$$

By Lemma 28, we thus have

$$\begin{aligned}|\langle x_i, x_j\rangle - \|\mu^{(q_i)}\|^2| &\le |\langle z_i, \mu^{(q_j)}\rangle| + |\langle \mu^{(q_i)}, z_j\rangle| + |\langle z_i, z_j\rangle| \\ &\le 2C_1 \max_q \|\mu^{(q)}\|\sqrt{\log(2nk/\delta)} + C_1\sqrt{d\log(2n/\delta)} \\ &\overset{(i)}{\le} 2C_1\sqrt{\frac{d\log(2nk/\delta)}{Cn}} + C_1\sqrt{d\log(2n/\delta)} \\ &\overset{(ii)}{\le} 2C_1\sqrt{d\log(2n/\delta)}. \qquad (46)\end{aligned}$$

The inequality (i) uses Assumption (CL2). Inequality (ii) follows since Assumption (CL1) implies that for $C > 1$ large enough, we have $n > 10C_1^2 k$ so that

$$\log(2nk/\delta) < \log(2n^2/\delta) < 2\log(2n/\delta).$$

For the fourth part of the lemma, if $i \in I^{(q)}$ and $j \in I^{(r)}$ for $q \neq r$,

$$\begin{aligned}|\langle x_i, x_j\rangle| &= |\langle \mu^{(q_i)} + z_i, \mu^{(q_j)} + z_j\rangle| \\ &\le |\langle \mu^{(q_i)}, \mu^{(q_j)}\rangle| + |\langle z_i, \mu^{(q_j)}\rangle| + |\langle \mu^{(q_i)}, z_j\rangle| + |\langle z_i, z_j\rangle| \\ &\le \max_{q\neq r} |\langle \mu^{(q)}, \mu^{(r)}\rangle| + 2C_1 \max_q \|\mu^{(q)}\|\sqrt{\log(2nk/\delta)} + C_1\sqrt{d\log(2n/\delta)} \\ &\le \max_{q\neq r} |\langle \mu^{(q)}, \mu^{(r)}\rangle| + 2C_1\sqrt{d\log(2n/\delta)}.\end{aligned}$$

where the second-to-last inequality uses Lemma 28, and the last inequality uses an identical argument to (46).

For the last part of the lemma, if $q \in Q$ then the quantity

$$|I^{(q)}| = \sum_{i=1}^{n} \mathbb{1}(\text{cluster}(i) = q)$$

is a sum of n i.i.d. random variables with mean $1/k$. By Hoeffding's inequality, for any $u \geq 0$,

$$\mathbb{P}\left(\left||I^{(q)}| - \frac{n}{k}\right| \geq u\right) \leq 2\exp\left(-\frac{2u^2}{n}\right).$$

In particular, selecting $u = \sqrt{n\log(2k/\delta)}$ and taking a union bound over the k clusters, we see that with probability at least $1-\delta$, for all $q \in Q$,

$$\left|\frac{|I^{(q)}|}{n} - \frac{1}{k}\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}.$$

Finally, let us denote by N_q the number of noisy examples within cluster q,

$$|I^{(q)}_{\mathcal{N}}| = N_q = \sum_{i \in I^{(q)}} \mathbb{1}(i \in \mathcal{N}).$$

Condintioned on the value of $|I^{(q)}|$, since we are considering random classification noise, N_q is the sum of $|I^{(q)}|$ independent, identically distributed random variables with mean

$$m_q := \mathbb{P}(i \in \mathcal{N}) = \eta.$$

By Hoeffding's inequality, for any $u \geq 0$,

$$\mathbb{P}\left(\left|N_q - |I^{(q)}|m_q\right| \geq u\right) \leq 2\exp\left(-\frac{2u^2}{|I^{(q)}|}\right).$$

In particular, selecting $u = \sqrt{|I^{(q)}|\log(2k/\delta)}$ and taking a union bound over the k clusters, we see that with probability at least $1-\delta$, for all $q \in Q$,

$$\left|\frac{|I^{(q)}_{\mathcal{N}}|}{|I^{(q)}|} - \eta\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}.$$

Since samples are 'clean' and in cluster q with probability $1-\eta$, a completely identical argument yields the bound for $|I^{(q)}_{\mathcal{C}}|$. Taking a union bound over the event in Lemma 28 and the three events above leads to a total failure probability of 7δ. ■

D.2. Proof of Lemma 10

We now show that under our assumptions on the problem parameters, the training data are p-orthogonal for large p and the norms of each example are quite close to each other.

Lemma 30 *There exists an absolute constant $C_2 > 0$ such for every large enough constant $C > 0$, for any $\delta \in (0, 1/7)$, under Assumptions (CL1) through (CL4) (defined for these C and δ), with probability at least $1-7\delta$ over $\mathsf{P}_{\mathsf{clust}}$, the training data is C/C_2-orthogonal, and $\max_{i,j} \|x_i\|^2/\|x_j\|^2 \leq (1 + C_2/\sqrt{C})^2$.*

Proof All of the results of Lemma 29 hold with probability at least $1 - 7\delta$. We shall show that the lemma is a deterministic consequence of this high-probability event.

First, if $i, j \in I^{(q)}$ and $i \neq j$, then by Lemma 29,

$$|\langle x_i, x_j \rangle| \leq \max_q \|\mu^{(q)}\|^2 + C_1' \sqrt{d \log(2n/\delta)} \leq 2C_1' \max\left(\max_q \|\mu^{(q)}\|^2, \sqrt{d \log(2n/\delta)} \right).$$

On the other hand, if $i \in I^{(q)}$ and $j \in I^{(r)}$ with $q \neq r$, then

$$\begin{aligned} |\langle x_i, x_j \rangle| &\leq \max_{q \neq r} |\langle \mu^{(q)}, \mu^{(r)} \rangle| + C_1' \sqrt{d \log(2n/\delta)} \\ &\overset{(i)}{\leq} \min_q \|\mu^{(q)}\|^2 + C_1' \sqrt{d \log(2n/\delta)} \\ &\leq 2C_1' \max\left(\max_q \|\mu^{(q)}\|^2, \sqrt{d \log(2n/\delta)} \right), \end{aligned}$$

where in (i) we use Assumption (CL4). Thus for any $i \neq j$ we have

$$|\langle x_i, x_j \rangle| \leq 2C_1' \max\left(\max_q \|\mu^{(q)}\|^2, \sqrt{d \log(2n/\delta)} \right). \tag{47}$$

On the other hand, by Lemma 29 we also have

$$d\left(1 - C_1' \sqrt{\frac{\log(2n/\delta)}{d}}\right) \leq \min_i \|x_i\|^2 \leq \max_i \|x_i\|^2 \leq d\left(1 + C_1' \sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right). \tag{48}$$

We can thus bound

$$\begin{aligned} \frac{\min_i \|x_i\|^4}{\max_i \|x_i\|^2} &\geq d \cdot \frac{\left(1 - C_1' \sqrt{\frac{\log(2n/\delta)}{d}}\right)^2}{\left(1 + C_1' \sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right)} \\ &\overset{(i)}{\geq} d \cdot \left(1 - C_1' \sqrt{\frac{\log(2n/\delta)}{d}}\right)^2 \cdot \left(1 - C_1' \sqrt{\frac{\log(2n/\delta)}{d}} - \frac{2}{Cn}\right) \\ &\overset{(ii)}{\geq} \frac{1}{2} d. \end{aligned} \tag{49}$$

In inequality (i) we have used that $1/(1+x) \geq 1 - x$ for $x > 0$, and in inequality (ii) we have taken $C > 1$ large enough in Assumption (CL2). Thus, we have

$$\frac{\min_i \|x_i\|^4}{\max_i \|x_i\|^2 \max_{i \neq j} |\langle x_i, x_j \rangle|} \geq \frac{d}{2 \max_{i \neq j} |\langle x_i, x_j \rangle|} \geq \frac{d}{2C_1' \max\left(\max_q \|\mu^{(q)}\|^2, \sqrt{d \log(2n/\delta)} \right)}.$$

Rearranging and using Assumption (CL2), this implies

$$\begin{aligned}\min_i \|x_i\|^2 &\geq \frac{1}{2C_1'} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot \frac{d}{\max(\max_q \|\mu^{(q)}\|^2, \sqrt{d\log(2n/\delta)})} \cdot \max_{i\neq j} |\langle x_i, x_j\rangle| \\ &\geq \frac{C}{2C_1'} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot n \max_{i\neq j} |\langle x_i, x_j\rangle|.\end{aligned}$$

In particular, the training data is C/C_2-orthogonal for $C_2 := 2C_1'$ (see Definition 1). Moreover, by (48) we have

$$\begin{aligned}R^2 &\leq \left(1 + C_1'\sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right) \cdot \left(1 - C_1'\sqrt{\frac{\log(2n/\delta)}{d}}\right)^{-1} \\ &\overset{(i)}{\leq} \left(1 + C_1'/\sqrt{C} + \frac{2}{Cn}\right) \cdot \left(1 - C_1'/\sqrt{C}\right)^{-1} \\ &\leq \left(1 + 2C_1'/\sqrt{C}\right)^2.\end{aligned}$$

The inequality (i) uses Assumption (CL2). The final inequality uses that $\frac{C_1'}{\sqrt{C}} + \frac{2}{Cn} \leq \frac{2C_1'}{\sqrt{C}}$ for C large enough and that $1/(1-x) \leq 1 + 2x$ for $x \in (0, 1/2)$. ■

D.3. Proof of Theorem 11

We now show that any τ-uniform linear classifier projected onto any direction of the form $y^{(q)}\mu^{(q)}$ is large and positive. By Lemma 27, this will be a key ingredient for a test error bound.

Lemma 31 *Let $u \in \mathbb{R}^d$ be τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$ for some absolute constant $\tau \geq 1$. Let $\Delta > 0$ be an absolute constant and assume $\eta \leq \frac{1}{1+\tau} - \Delta$. Then under Assumptions (CL1) through (CL4), provided $C > 1$ is a large enough absolute constant (depending only on η, τ, and Δ), then with probability at least $1 - 7\delta$ over $\mathsf{P}_{\mathsf{clust}}^n$, for each $q \in Q$,*

$$\frac{\langle u, y^{(q)}\mu^{(q)}\rangle}{\|u\|} \geq \frac{\sqrt{3}(1+\tau)\Delta}{4\sqrt{10}\tau} \cdot \frac{\sqrt{n}\|\mu^{(q)}\|^2}{k\sqrt{d}}.$$

Proof First note that with probability at least $1 - 7\delta$, the items in both Lemma 29 and Lemma 28 (with the vectors $v_i = \mu^{(i)}$) hold. We also showed that Lemma 10 holds as a deterministic consequence of these lemmas. In the remainder of the proof, we will work on this high-probability event and show that Lemma 31 follows as a deterministic consequence of Lemmas 28, 29, and 10 under Assumptions (CL1) through (CL4).

Since u is τ-uniform, there are strictly positive numbers s_i such that $u = \sum_{i=1}^n s_i y_i x_i$ and $\max_{i,j} {}^{s_i}/_{s_j} = \tau$. Our proof consists in two parts: first, we want to show that for each q, the quantity

$$\begin{aligned}\langle u, y^{(q)}\mu^{(q)}\rangle &= \left\langle \sum_{i=1}^n s_i y_i x_i, y^{(q)}\mu^{(q)} \right\rangle \\ &= \sum_{r=1}^k \sum_{i\in I^{(r)}} s_i \langle y_i x_i, y^{(q)}\mu^{(q)}\rangle \\ &= \sum_{i\in I^{(q)}} s_i \langle y_i x_i, y^{(q)}\mu^{(q)}\rangle + \sum_{r\neq q} \sum_{i\in I^{(r)}} s_i \langle y_i x_i, y^{(q)}\mu^{(q)}\rangle\end{aligned}$$

is large. We will do so by considering the two terms above. Intuitively, when $i \in I^{(q)}$ then the summands in the first term $\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle$ will be large and positive for clean points $i \in I_{\mathcal{C}}^{(q)}$ and negative for noisy points $i \in I_{\mathcal{N}}^{(q)}$, and so as long as there are more clean points than noisy ones, the first term will be large and positive. For the second term above, this term will not be too large in absolute value since the clusters are nearly-orthogonal. After we show that the above holds, we then want to provide an upper bound on $\|u\|^2$.

We will first show that the quantity $\langle u, y^{(q)}\mu^{(q)}\rangle$ is large and positive by considering the two terms in the above decomposition separately.

First term: $i \in I^{(q)}$. In this case, we have $\mu^{(q_i)} = \mu^{(q)}$. If $i \in I_{\mathcal{C}}^{(q)}$, then $y_i = y^{(q)}$, while if $i \in I_{\mathcal{N}}^{(q)}$, then $y_i = -y^{(q)}$. We will thus show a positive lower bound for clean points and an upper bound on the absolute value of noisy points.

We first provide a lower bound for clean samples $i \in \mathcal{C}$. For such samples, $x_i = \mu^{(q_i)} + z_i$ and $y_i = \tilde{y}^{(q_i)} = \tilde{y}$ and so

$$\begin{aligned}\langle s_i y_i x_i, y^{(q)}\mu^{(q)}\rangle &= s_i \langle \mu^{(q)} + z_i, \mu^{(q)}\rangle \\ &\geq s_i \left[\|\mu^{(q)}\|^2 - |\langle z_i, \mu^{(q)}\rangle|\right] \\ &= s_i \|\mu^{(q)}\|^2 \left(1 - \frac{|\langle z_i, \mu^{(q)}\rangle|}{\|\mu^{(q)}\|^2}\right) \\ &\overset{(i)}{\geq} s_i \|\mu^{(q)}\|^2 \left(1 - \frac{C_1\sqrt{\log(2nk/\delta)}}{\|\mu^{(q)}\|}\right).\end{aligned}$$

Inequality (i) uses Lemma 28. Using an identical sequence of calculations, we can derive a similar upper bound for $|\langle s_i y_i x_i, \tilde{y}(\mu^{(q)} + z)\rangle|$ for noisy examples: we have for $i \in I_{\mathcal{N}}^{(q)}$,

$$\begin{aligned}|\langle s_i y_i x_i, \tilde{y}\mu^{(q)}\rangle| &= s_i |\langle \mu^{(q)} + z_i, \mu^{(q)}\rangle| \\ &\leq s_i \left[\|\mu^{(q)}\|^2 + |\langle z_i, \mu^{(q)}\rangle|\right] \\ &\leq s_i \|\mu^{(q)}\|^2 \left(1 + \frac{C_1\sqrt{\log(2nk/\delta)}}{\|\mu^{(q)}\|}\right).\end{aligned}$$

Putting the two preceding displays together, we get,

$$\begin{cases} s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle \geq s_i\|\mu^{(q)}\|^2 \cdot \left(1 - C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right), & i \in I^{(q)}_{\mathcal{C}}, \\ |s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle| \leq s_i\|\mu^{(q)}\|^2 \cdot \left(1 + C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right), & i \in I^{(q)}_{\mathcal{N}}. \end{cases} \tag{50}$$

Second term: $i \in I^{(r)}$**,** $r \neq q$**.** Since $\mu^{(q_i)} \neq \mu^{(q)}$, we have for both noisy and clean examples,

$$\begin{aligned} |s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle| &= s_i|\langle \mu^{(r)} + z_i, \mu^{(q)}\rangle| \\ &\leq s_i\left(|\langle z_i, \mu^{(q)}\rangle| + \max_{q\neq r}|\langle \mu^{(q)}, \mu^{(r)}\rangle|\right) \\ &\overset{(i)}{\leq} s_i\left(C_1\|\mu^{(q)}\|\sqrt{\log(2nk/\delta)} + \max_{q\neq r}|\langle \mu^{(q)}, \mu^{(r)}\rangle|\right) \end{aligned} \tag{51}$$

Inequality (i) uses Lemma 28. Putting the above together, we get,

$$\begin{aligned} &\langle u, y^{(q)}\mu^{(q)}\rangle \\ &= \sum_{i\in I^{(q)}} s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle + \sum_{r\neq q}\sum_{i\in I^{(r)}} s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle \\ &\geq \sum_{i\in I^{(q)}_{\mathcal{C}}} s_i\|\mu^{(q)}\|^2\left(1 - C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right) - \sum_{i\in I^{(q)}_{\mathcal{N}}} s_i\|\mu^{(q)}\|^2\left(1 + C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right) \\ &\quad - \sum_{r\neq q}\sum_{i\in I^{(r)}} s_i\left(C_1\|\mu^{(q)}\|\sqrt{\log(2nk/\delta)} + \max_{q\neq r}|\langle \mu^{(q)}, \mu^{(r)}\rangle|\right) \\ &\geq \left(\min_i s_i\right)|I^{(q)}_{\mathcal{C}}|\|\mu^{(q)}\|^2\left(1 - C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right) - \left(\max_i s_i\right)|I^{(q)}_{\mathcal{N}}|\|\mu^{(q)}\|^2\left(1 + \sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right) \\ &\quad - \left(\max_i s_i\right)\cdot\left(n - |I^{(q)}|\right)\cdot\left(C_1\|\mu^{(q)}\|\sqrt{\log(2nk/\delta)} + \max_{q\neq r}|\langle \mu^{(q)}, \mu^{(r)}\rangle|\right). \end{aligned} \tag{52}$$

For notational simplicity let us define

$$\nu := C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}} \ll 1, \tag{53}$$

where ν small follows by Assumption (CL3). Since $\tau := \max_i s_i / \min_i s_i$ and $|I^{(q)}| = |I_{\mathcal{C}}^{(q)}| + |I_{\mathcal{N}}^{(q)}|$, we can then write the above inequality as

$$\begin{aligned}
&\langle u, y^{(q)}\mu^{(q)}\rangle \\
&\geq \left(\min_i s_i\right) \cdot \left(|I^{(q)}| - |I_{\mathcal{N}}^{(q)}|\right) \cdot \|\mu^{(q)}\|^2(1-\nu) - \left(\max_i s_i\right) \cdot |I_{\mathcal{N}}^{(q)}| \cdot \|\mu^{(q)}\|^2(1+\nu) \\
&\quad - \left(\max_i s_i\right) \cdot (n - |I^{(q)}|) \cdot \left(\|\mu^{(q)}\|^2\nu + \max_{q\neq r} |\langle \mu^{(q)}, \mu^{(r)}\rangle|\right) \\
&= \left(\min_i s_i\right) |I^{(q)}| \|\mu^{(q)}\|^2 \Bigg[1 - \nu - (1-\nu)\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} - (1+\nu)\tau \cdot \frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} \\
&\quad - \tau\left(\frac{n}{|I^{(q)}|} - 1\right) \cdot \left(\nu + \frac{\max_{q\neq r} |\langle \mu^{(q)}, \mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\Bigg] \\
&= \left(\min_i s_i\right) |I^{(q)}| \|\mu^{(q)}\|^2 \Bigg[1 - (1+\tau) \cdot \frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} - \left(1 - (1-\tau) \cdot \frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|}\right)\nu \\
&\quad - \tau\left(\frac{n}{|I^{(q)}|} - 1\right) \cdot \left(\nu + \frac{\max_{q\neq r} |\langle \mu^{(q)}, \mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\Bigg].
\end{aligned} \tag{54}$$

From here we see we need to control $|I^{(q)}|$ and $|I_{\mathcal{N}}^{(q)}|$. Using Lemma 29, we have

$$\left|\frac{|I^{(q)}|}{n} - \frac{1}{k}\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}, \qquad \left|\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} - \eta\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}.$$

In particular, we have

$$|I^{(q)}| \geq \frac{n}{k} - \sqrt{n\log(2k/\delta)} = \frac{n}{k}\left(1 - \sqrt{\frac{k^2\log(2k/\delta)}{n}}\right) \overset{(i)}{\geq} \frac{n}{2k}, \tag{55}$$

where inequality (i) uses Assumption (CL1) so that $n \geq 4k^2\log(2k/\delta)$. We therefore have

$$\frac{n}{|I^{(q)}|} - 1 \leq 2k.$$

Substituting these inequalities into (54) and using that $(1-(1-\tau)|I_{\mathcal{N}}^{(q)}|/|I^{(q)}|) \leq \tau$, we get,

$$\begin{aligned}
\langle u, y^{(q)}\mu^{(q)}\rangle \geq |I^{(q)}|\|\mu^{(q)}\|^2 \left(\min_i s_i\right) \cdot \Bigg[1 - (1+\tau) \cdot \left(\eta + \sqrt{\frac{\log(2k/\delta)}{n}}\right) - \tau\nu \\
- 2k\tau \cdot \left(\nu + \frac{\max_{q\neq r} |\langle \mu^{(q)}, \mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\Bigg].
\end{aligned} \tag{56}$$

Algebraic calculations to finish the bound on $\langle u, y^{(q)}\mu^{(q)}\rangle$. We now want to show that the quantity appearing in the brackets in (56) is positive. Since by assumption $\eta \leq \frac{1}{1+\tau} - \Delta$ for some

absolute constant $\Delta > 0$,

$$\begin{aligned}
&1 - (1+\tau)\cdot\left(\eta + \sqrt{\frac{\log(2k/\delta)}{n}}\right) - \tau\nu - 2k\tau\cdot\left(\nu + \frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right) \\
&\quad\geq 1 - (1+\tau)\cdot\left(\frac{1}{1+\tau} - \Delta + \sqrt{\frac{\log(2k/\delta)}{n}}\right) - \tau\nu - 2k\tau\cdot\left(\nu + \frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right) \\
&\quad= (1+\tau)\cdot\left[\Delta - \sqrt{\frac{\log(2k/\delta)}{n}} - \frac{\tau\nu}{1+\tau} - \frac{2k\tau}{1+\tau}\cdot\left(\nu + \frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\right] \\
&\quad= (1+\tau)\Delta\cdot\left[1 - \sqrt{\frac{\Delta^{-2}\log(2k/\delta)}{n}} - \frac{\tau\nu\Delta^{-1}}{1+\tau} - \frac{2k\tau\Delta^{-1}}{1+\tau}\cdot\left(\nu + \frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\right].
\end{aligned} \tag{57}$$

For the second term in the brackets, Assumption (CL1) implies

$$\sqrt{\frac{\Delta^{-2}\log(2k/\delta)}{n}} \leq \frac{1}{8}.$$

For the next two terms, note that $\frac{\tau}{1+\tau} \leq 1$ since $\tau \geq 1$. Since $\nu = C_1\sqrt{\log(2nk/\delta)/\|\mu^{(q)}\|^2}$, the second term can be driven to zero by taking $C > 1$ sufficiently large by Assumption (CL3) (namely, $\min_q \|\mu^{(q)}\| \geq Ck\sqrt{\log(2nk/\delta)}$):

$$\frac{\tau\nu\Delta^{-1}}{1+\tau} \leq \Delta^{-1}\cdot C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}} \leq \frac{1}{8}.$$

Again using Assumption (CL3), for $C > 1$ large enough we have,

$$\frac{2k\tau\Delta^{-1}}{1+\tau}\cdot\nu \leq 2k\Delta^{-1}\cdot C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}} \leq \frac{1}{32}.$$

Finally, Assumption (CL4) implies that for $C > 1$ large enough,

$$\frac{2k\tau\Delta^{-1}}{1+\tau}\cdot\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\left\|\mu^{(q)}\right\|^2} \leq \frac{1}{32}.$$

Putting the above into (57), we get

$$\begin{aligned}
&1 - (1+\tau)\cdot\left(\eta + \sqrt{\frac{\log(2k/\delta)}{n}}\right) - \tau\nu - 2k\tau\cdot\left(\nu + \frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right) \\
&\quad\geq (1+\tau)\Delta\left(1 - \frac{1}{8} - \frac{1}{8} - \frac{1}{32} - \frac{1}{32}\right) > \frac{(1+\tau)\Delta}{2}.
\end{aligned}$$

Substituting this into (56), we get

$$\begin{aligned}\langle u, y^{(q)}\mu^{(q)}\rangle &\geq |I^{(q)}|\|\mu^{(q)}\|^2\left(\min_i s_i\right)\cdot\left[1-(1+\tau)\cdot\left(\eta+\sqrt{\frac{\log(2k/\delta)}{n}}\right)-\tau\nu\right.\\ &\qquad\left.-2k\tau\cdot\left(\nu+\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\right]\\ &\geq\frac{1}{2}(1+\tau)\Delta|I^{(q)}|\|\mu^{(q)}\|^2\left(\min_i s_i\right)\\ &\overset{(i)}{\geq}\frac{(1+\tau)n\|\mu^{(q)}\|^2\Delta(\min_i s_i)}{4k}.\end{aligned}\tag{58}$$

The inequality (i) uses (55). This provides the requisite lower bound for $\langle u, y^{(q)}\mu^{(q)}\rangle$.

Upper bound on $\|u\|$. Here we use the fact that the samples are nearly-orthogonal: we have,

$$\begin{aligned}\left\|\sum_{i=1}^n s_i y_i x_i\right\|^2 &\leq \sum_{i=1}^n s_i^2\|x_i\|^2+\sum_{i\neq j}s_i s_j|\langle x_i,x_j\rangle|\\ &\leq n\left(\max_i s_i^2\right)\left(\max_i\|x_i\|^2\right)+n^2\left(\max_i s_i^2\right)\max_{i\neq j}|\langle x_i,x_j\rangle|\\ &= n\left(\max_i s_i^2\right)\left(\max_i\|x_i\|^2+n\max_{i\neq j}|\langle x_i,x_j\rangle|\right)\\ &\overset{(i)}{\leq}\frac{5}{4}n\left(\max_i s_i^2\right)\left(\max_i\|x_i\|^2\right)\\ &\overset{(ii)}{\leq}\frac{5}{4}n\left(\max_i s_i^2\right)\cdot d\left(1+C_1\sqrt{\frac{\log(2n/\delta)}{d}}+\frac{2}{Cn}\right)\\ &\overset{(iii)}{\leq}\frac{10}{3}nd\max_i s_i^2.\end{aligned}\tag{59}$$

Inequality (i) above uses Lemma 10. Inequality (ii) uses Lemma 29, and inequality (iii) follows by taking $C > 1$ large enough by Assumptions (CL1) and (CL2). Putting (59) and (58) together, we get,

$$\frac{\left\langle u, y^{(q)}\mu^{(q)}\right\rangle^2}{\|u\|^2}\geq\frac{(1+\tau)^2n^2\|\mu^{(q)}\|^4\Delta^2\min_i s_i^2}{16k^2\cdot\frac{10}{3}nd\max_i s_i^2}=\frac{3(1+\tau)^2\Delta^2}{160\tau^2}\cdot\frac{n\|\mu^{(q)}\|^4}{k^2d}.$$

Taking square roots of the above completes the proof. ■

Putting together Lemma 31 and Lemma 27, we can derive a generalization bound for the linear classifier $\sum_{i=1}^n s_i y_i x_i$.

Theorem 11 *Let $\tau \geq 1$ be a constant, and suppose $\eta \leq \frac{1}{1+\tau} - \Delta$ for some absolute constants $\eta, \Delta > 0$. There exist constants $C, C' > 0$ (depending only on η, τ, and Δ) such that for any*

$\delta \in (0, 1/14)$, under Assumptions (CL1) through (CL4) (defined for these C and δ), with probability at least $1 - 14\delta$ over $\mathsf{P}_{\mathsf{clust}}^n$, if $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $\{(x_i, y_i)\}_{i=1}^n$, then

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\big(\langle u, x_k\rangle\big),$$

$$\text{while simultaneously,} \quad \eta \le \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle u, x\rangle)\big) \le \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

In particular, if $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$, then the linear classifier $x \mapsto \operatorname{sign}(\langle u, x\rangle)$ exhibits benign overfitting.

Proof By a union bound, with probability at least $1 - 14\delta$, the results of Lemmas 31 and Lemma 10 hold. In the remainder of the proof we will work on this high-probability event and show that the theorem is a deterministic consequence of it and the Assumptions (CL1) through (CL4).

Since u is τ-uniform, there are strictly positive numbers s_i such that $u = \sum_{i=1}^n s_i y_i x_i$. We shall first show this estimator interpolates the training data. An identical calculation used as in (38) shows that

$$\begin{aligned}
\langle u, y_k x_k\rangle &= s_k \|x_k\|^2 + \sum_{i\neq k} \langle s_i y_i x_i, y_k x_k\rangle \\
&\ge s_k \|x_k\|^2 \left(1 - \frac{n\tau \max_{i\neq j} |\langle x_i, x_j\rangle|}{\|x_k\|^2}\right) \\
&\overset{(i)}{\ge} s_k \|x_k\|^2 \left(1 - \frac{C_2 \tau}{C}\right) \\
&\ge \frac{1}{2} s_k \|x_k\|^2 > 0.
\end{aligned}$$

The inequality (i) uses that the training data is C/C_2-orthogonal by Lemma 10, and we took C large relative to the absolute constants C_2, τ.

We now show the generalization error is close to the noise rate. Since $\eta \le \frac{1}{1+\tau} - \Delta$, by Lemma 31, we know that for each q we have,

$$\frac{\big\langle \sum_{i=1}^n s_i y_i x_i, y^{(q)} \mu^{(q)}\big\rangle}{\|\sum_{i=1}^n s_i y_i x_i\|} \ge \frac{\sqrt{3}(1+\tau)\Delta}{4\sqrt{10}\tau} \cdot \frac{\sqrt{n}\|\mu^{(q)}\|^2}{k\sqrt{d}}.$$

Now using Lemma 27, this implies that

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle \widehat{\mu}, x\rangle)\big) &\le \eta + \frac{1}{k}\sum_{q=1}^k \exp\left(-\frac{3c(1+\tau)^2 n \Delta^2 \|\mu^{(q)}\|^4}{160\tau^2 k^2 d}\right) \\
&\le \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right),
\end{aligned}$$

where C' is an absolute constant independent of d and n. ∎

D.4. Proof of Corollary 12 and Corollary 13

In this section we show how to use Theorem 11 and Lemma 10 to prove Corollary 12 and Corollary 13.

Corollary 32 *Suppose* $0 < \eta \leq 0.49$. *There exist constants* $C, C' > 0$ *such that for any* $\delta \in (0, 1/21)$, *under Assumptions (CL1) through (CL4) (defined for these* C *and* δ*), with probability at least* $1 - 21\delta$ *over* $\mathsf{P}_{\mathsf{clust}}^n$, *the max-margin linear classifier* $w = \operatorname{argmin}\{\|w\|^2 : y_i \langle w, x_i \rangle \geq 1 \, \forall i\}$ *satisfies*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\left(\langle w, x_k \rangle\right),$$

$$\text{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{clust}}}\left(y \neq \operatorname{sign}(\langle w, x \rangle)\right) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

In particular, if $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$ *then* w *exhibits benign overfitting.*

Proof The calculation is essentially identical to that used for the proof of Corollary 7. By a union bound, the results of Theorem 11 and Lemma 10 hold with probability at least $1 - 21\delta$, and any τ-uniform linear classifier exhibits benign overfitting with noise tolerance determined by τ. We therefore verify that the linear max-margin classifier is τ-uniform with small τ.

By Lemma 10, the training data is C/C_2-orthogonal and $R^2 = \max_{i,j} \|x_i\|^2/\|x_j\|^2 \leq (1 + C_2/\sqrt{C})^2$. Since for C large enough we have $C/C_2 \geq 3$, by Proposition 3 this means the linear max-margin w is τ-uniform with $\tau \leq R^2\left(1 + \frac{2}{CC_2^{-1}R^2 - 2}\right)$. In particular, we have

$$\tau \leq \left(1 + \frac{C_2}{\sqrt{C}}\right)^2 \cdot \left(1 + \frac{2}{CC_2^{-1}R^2 - 2}\right)^2 \leq \frac{100}{99} \cdot \frac{201}{200} = \frac{201}{198}.$$

The final inequality follows by taking $C > 1$ a large enough absolute constant. Thus the max-margin linear classifier is τ-uniform where $\tau \leq \frac{201}{198}$. Since $\frac{1}{1+\tau} \geq \frac{198}{399} \geq 0.496$, by taking $\eta \leq 0.49 = 0.496 - 0.006$ we may apply Theorem 11. ∎

Finally, we prove Corollary 13, again re-stated for convenience.

Corollary 33 *Suppose that* $0 < \eta \leq \frac{49\gamma^2}{100}$. *There exist constants* $C, C' > 0$ *such that for any* $\delta \in (0, 1/21)$, *under Assumptions (CL1) through (CL4) (defined for these* C *and* δ*), with probability at least* $1 - 21\delta$ *over* $\mathsf{P}_{\mathsf{clust}}^n$, *any KKT point* W *of Problem (3) satisfies*

$$\text{for all } k \in [n], \quad y_k = \operatorname{sign}\left(f(x_k; W)\right),$$

$$\text{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{clust}}}\left(y \neq \operatorname{sign}\left(f(x; W)\right)\right) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

In particular, if $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$ *then the neural network* $f(x; W)$ *exhibits benign overfitting. Moreover, for any initialization* $W(0)$, *gradient flow converges in direction to a network which satisfies the above.*

Proof As in the preceding corollary, with probability at least $1 - 21\delta$ the results of Theorem 11 and Lemma 10 hold and any τ-uniform linear classifier exhibits benign overfitting with noise tolerance determined by τ. By Lemma 10, the training data is C/C_2-orthogonal for $C > 3C_2\gamma^{-3}$ we may apply Proposition 4 so that $\operatorname{sign}(f(x;W)) = \operatorname{sign}(\langle z, x\rangle)$ where z is τ-uniform w.r.t. the training data for $\tau = R^2\gamma^{-2}\left(1 + \frac{2}{\gamma C R^2/C_2 - 2}\right)$. Lemma 10 shows that $R^2 \leq \frac{100}{99}$ for C large enough, and hence $\tau \leq \frac{201}{198}\gamma^{-2}$ for large C. Note that $\frac{1}{1+201\gamma^{-2}/198} \geq 0.496\gamma^2$. Hence, we may apply Theorem 6 with $\eta \leq 0.49\gamma^2 = 0.496\gamma^2 - 0.006\gamma^2$ since γ is an absolute constant. ■

Appendix E. Experiment details

In this section we provide details for the experiment in Figure 1. We consider two-layer leaky ReLU networks of the form (2) where $\gamma = 0.1$ with $m = 512$ neurons, so that half of the second-layer weights are fixed at $+1/\sqrt{m}$ and the other half at $-1/\sqrt{m}$. We initialize the first-layer weights using the TensorFlow default of Glorot uniform initialization. We assume data of the form (4), where $d = 800$, $\|\mu\| = d^{0.26}$ and $\eta = 0.15$. We consider two settings: either $n = 8000$ or $n = 100$. We train for 10^6 steps with learning rate $\alpha = 0.01$ in each setting. The figure plots the average over 3 random seeds over the random initialization and the sampling of the data, with the shaded area corresponding to the range of the minimum to the maximum accuracies across the three seeds.