MAX-MARGIN WORKS WHILE LARGE MARGIN FAILS: GENERALIZATION WITHOUT UNIFORM CONVERGENCE

Margalit Glasgow, Colin Wei, Mary Wootters & Tengyu Ma
Department of Computer Science
Stanford University
Stanford, CA 94305, USA
{mglasgow,colinwei,marykw,tengyuma}@stanford.edu

ABSTRACT

A major challenge in modern machine learning is theoretically understanding the generalization properties of overparameterized models. Many existing tools rely on *uniform convergence* (UC), a property that, when it holds, guarantees that the test loss will be close to the training loss, uniformly over a class of candidate models. Nagarajan & Kolter (2019b) show that in certain simple linear and neural-network settings, any uniform convergence bound will be vacuous, leaving open the question of how to prove generalization in settings where UC fails. Our main contribution is proving novel generalization bounds in two such settings, one linear, and one non-linear. We study the linear classification setting of Nagarajan & Kolter (2019b), and a quadratic ground truth function learned via a two-layer neural network in the non-linear regime. We prove a new type of margin bound showing that above a certain signal-to-noise threshold, any near-max-margin classifier will achieve almost no test loss in these two settings. Our results show that near-max-margin is important: while any model that achieves at least a $(1-\epsilon)$-fraction of the max-margin generalizes well, a classifier achieving half of the max-margin may fail terribly. Our analysis provides insight on why memorization can coexist with generalization: we show that in this challenging regime where generalization occurs but UC fails, near-max-margin classifiers contain both some generalizable components and some overfitting components that memorize the data. The presence of the overfitting components is enough to preclude UC, but the near-extremal margin guarantees that sufficient generalizable components are present.

1 INTRODUCTION

A central challenge of machine learning theory is understanding the generalization of overparameterized models. While in many real-world settings deep networks achieve low test loss, their high capacity makes theoretical analysis with classical tools difficult, or sometimes impossible (Zhang et al., 2017; Nagarajan & Kolter, 2019b). Most classical theoretical tools are based on *uniform convergence* (UC), a property that, when it holds, guarantees that the test loss will be close to the training loss, uniformly over a class of candidate models. Many generalization bounds for neural networks are built on this property, e.g. Neyshabur et al. (2015; 2017b; 2018); Harvey et al. (2017); Golowich et al. (2018).

The seminal work of Nagarajan & Kolter (2019b) gives theoretical and empirical evidence that UC cannot hold in natural overparameterized linear and neural network settings. The impossibility results of Nagarajan and Kolter are strong: they rule out even UC on the smallest reasonable algorithm-dependent family of models, that is, any possible models output by learning algorithm on typical datasets. In particular, they prove that in an overparameterized linear classification problem, models found by gradient descent will achieve small test loss, but any UC bound over these models will be vacuous. In a two-layer neural network setting, Nagarajan & Kolter (2019b) empirically demonstrate the same phenomenon for the $0/1$ loss.

Many margin-based generalization bounds do not technically fit into the category of UC bounds defined by Nagarajan and Kolter, but still may be intrinsically limited for similar reasons. Classical

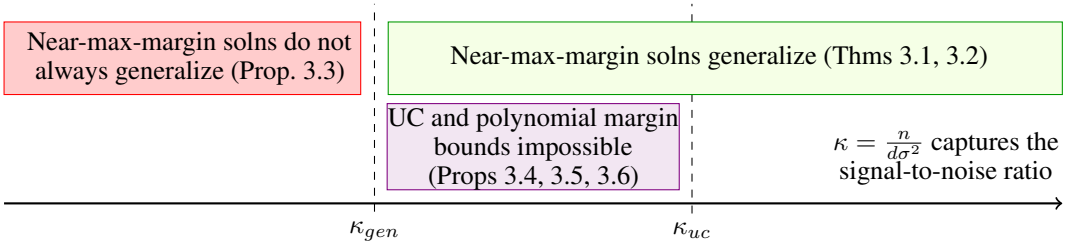

Figure 1: Thresholds for Uniform Convergence and Generalization.

margin-based generalization guarantees bounds (see eg. Shalev-Shwartz & Ben-David (2014); Kakade et al. (2009)) and related margin bounds for neural networks (Wei & Ma, 2019a; 2020; Bartlett et al., 2017; Golowich et al., 2018) scale inversely polynomially in the margin size, and are typically proved via uniform convergence on a surrogate loss (eg. the hinge loss or ramp loss) that upper bounds the $0/1$ misclassification loss. Nagarajan and Kolter's results show that any UC bound on the ramp loss is vacuous in an overparameterized linear setting, suggesting (though not proving) that classical margin bounds may not be useful. Muthukumar et al. (2021) shows empirically that such margin bounds are vacuous in a broader linear settings. In light of this, it is very important to develop theoretical tools to analyze generalization in settings where uniform convergence cannot yield meaningful bounds.

In this paper we establish novel margin-based generalization bounds in regimes where UC provably fails. These bounds guarantee generalization in the extremal case where the model has a near-maximal margin, and thus we call them *extremal margin bounds*.Indeed, near max-margin solutions are achievable by minimizing the logistic loss with weak ℓ_2-regularization (Wei et al., 2019), and minimizing the unregularized logistic loss with gradient descent converges to a stationary points of the max-margin objective (Lyu & Li, 2019; Lyu et al., 2021). In linear settings, SGD converges to the max-margin (Nacson et al., 2019).

Our results consider two settings, the linear setting of Nagarajan & Kolter (2019b), and a commonly studied quadratic problem learned on by a two-layer neural network (Wei et al., 2019; Frei et al., 2022b). In Theorems 3.1 and 3.2, we prove that above a certain signal-to-noise threshold κ_{gen}, near-max-margin solutions will generalize. Below this threshold, max-margin solutions may not generalize (Proposition 3.3). Below a second higher threshold, κ_{uc}, uniform convergence fails (Proposition 3.4). In Figure 1 we illustrate these three regions; the main significance of our results is in the challenging middle region between κ_{gen} and κ_{uc} where generalization occurs, but UC fails.

Additionally in this regime where UC fails, we show that classical margin bounds can only yield loose guarantees, even for the max-margin solution (Proposition 3.5 and 3.6). We prove this by showing the existence of models that achieve a large but non-near-max-margin (e.g., half the max-margin), but do not generalize at all. This phase transition between good-margin and near-max-margin cannot be captured by classical margin bounds where the generalization guarantee decays inversely polynomially in the margin. Our extremal margin bounds are fundamentally different from classical margin bounds and are not based on uniform convergence.

Prior works have also studied the challenging regime where uniform convergence does not work. In a linear regression setting, Zhou et al. (2020) and Koehler et al. (2021) show that the test loss can be uniformly bounded for all low-norm solutions that perfectly fit the data (this uses the data-dependent interpolation condition to improve upon UC bounds); nevertheless, Yang et al. (2021) shows that such bounds are still loose on the min-norm solution. Negrea et al. (2020) suggests an alternative framework based on uniform convergence over a less complex family of *surrogate* models; they use this technique to show generalization in a linear setting and in another high-dimensional problem amenable to analysis. To our knowledge, our results are the first instance of theoretically proving generalization in a neural network setting (that is not in the NTK regime) where UC provably fails.

We leverage near-max-margins in a unified way for both the linear and nonlinear settings, and we hope that this approach will be useful more broadly in overparameterized settings. In the challenging regime of generalization without UC, good learned models contain some generalizable signal components and some overfitting components that memorize the data. Our main technique is to show that any

near max-margin solution has to contain *both* signal components and overfitting components. The overfitting component causes UC to fail, but fortunately, has a reduced influence on a random test example, whereas the signal component has a similar influence on training and test examples.

1.1 Additional Related Work

A large body of work highlights challenges in using classical statistical theory to explain generalization in deep learning. Experimental results (Zhang et al., 2017; Neyshabur et al., 2017a) point out that despite being large in traditional capacity measures such as Rademacher complexities, deep networks still generalize well, and new explanations are needed to understand this behavior. Belkin et al. (2018) show that similar challenges hold in kernel methods. Beyond the work of Nagarajan & Kolter (2019b), Bartlett & Long (2021) prove that in a linear interpolation setting, model-dependent generalization bounds fail for the min-norm solution. Koren et al. (2022) show that SGD can exhibit a benign underfitting phenomenon where the test loss is small but empirical loss is large.

One related body of work has focused on characterizing "benign overfitting", where the model overfits to noise in labels of the training data but still attains good test performance. Our setting differs from benign overfitting because we study models that overfit prohibitively enough to preclude UC even with *clean* data. For models that overfit to noise, (i) it still may be possible to for algorithm-dependent notions of a UC bound to explain generalization on clean data, and (ii) if the overfitting is avoided with regularization, UC bounds may also be possible. Most of the results in this area concern linear models: Bartlett et al. (2020) analyze benign overfitting in regression problems by leveraging a closed form expression for the min-norm solution. Muthukumar et al. (2021); Shamir (2022); Cao et al. (2021); Wang & Thrampoulidis (2020); Chatterji & Long (2021) and Wang et al. (2021) study classification settings. The setting of Chatterji & Long (2021) is particularly similar to ours since it considers the max-margin solution under a Gaussians mixture. The works of Muthukumar et al. (2020) and Shamir (2022) reveal that is often possible to have benign overfitting in classification, whereas in regression for the same covariate distribution, the overfitting would imply poor generalization. Also closely related to our work on linear classification is the work of Montanari et al. (2019), which asymptotically characterizes the generalization of the max-margin solution as $n, d \to \infty$. Benign overfitting in neural networks has been shown in several simple settings. Frei et al. (2022a) analyzes two-layer neural networks trained by gradient descent on linearly-separable data. Cao et al. (2022) studies benign overfitting for a two-layer simplified convolutional network. Their techniques involve decomposing the output of the network into a sum of two terms, one involving the signal feature, and one involving the noise feature. Our techniques are very different because this decomposition is not possible for a fully connected 2-layer neural network.

More broadly, a variety of new generalization bounds have been derived in hopes of explaining generalization in deep learning. While none of these bounds have been explicitly proven to succeed in regimes where UC fails, they leverage additional properties of the training data or the optimization process and thus are not directly susceptible to the critiques of Nagarajan & Kolter (2019b). Among these are works that leverage properties such as Lipschitzness of the model on the training data (Arora et al., 2018; Nagarajan & Kolter, 2019a; Wei & Ma, 2019a;b), use algorithmic stability (Mou et al., 2018; Li et al., 2019a; Chatterjee & Zielinski, 2022), or information-theoretic perspectives (Negrea et al., 2019; Haghifam et al., 2021).

Finally, a body of work seeks to draw connections between optimization and generalization in deep learning by studying implicit regularization effects of the optimization algorithm (see e.g. (Gunasekar et al., 2017; Li et al., 2017; Gunasekar et al., 2018a;b; Woodworth et al., 2020; Damian et al., 2021; HaoChen et al., 2020; Li et al., 2019b; Wei et al., 2020) and related references). Most relevent in this literature is the aforementioned work connecting gradient descent and max-margin solutions.

2 Preliminaries

Our work achieves results in two settings. The first is a linear setting previously studied by Nagarajan & Kolter (2019b) where both the ground truth and the trained model are linear. In the second nonlinear setting, studied before by Wei et al. (2019); Frei et al. (2022b), the ground truth is quadratic, and the trained model is a two-layer neural network. In both settings, the data is drawn from a product

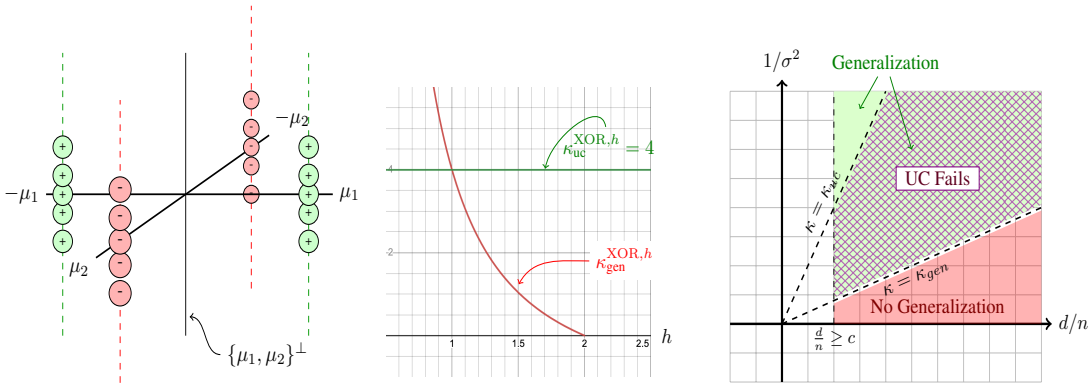

Figure 2: Left: Quadratic XOR Problem. Middle: $\kappa_{\text{gen}}^{\text{XOR},h}$ (red) and $\kappa_{\text{uc}}^{\text{XOR},h}$ (green) as a function of h. Right: Regions in which theorems hold. As shown in this figure, our results only hold when there is sufficient overparameterization, that is, $d \geq cn$ for a constant c.

distribution on features involved in the ground truth labeling function, and "junk" features orthogonal to the signal. We formalize the two settings below.

Linear setting

► **Data Distribution**. Fix some ground truth unit vector direction $\mu \in \mathbb{R}^d$. Let $x = z + \xi$, where $z \sim \text{Uniform}(\{\mu, -\mu\})$ and ξ is uniform on the sphere of radius $\sqrt{d-1}\sigma$ in $d-1$ dimensions, orthogonal to the direction μ. Let $y = \mu^T x$, such that $y = 1$ with probability $1/2$ and -1 with probability $1/2$. We denote this distribution of (x, y) on $\mathbb{R}^d \times \{-1, 1\}$ by $\mathcal{D}_{\mu,\sigma,d}$.

► **Model.** We learn a model $w \in \mathbb{R}^d$ that predicts $\hat{y} = \text{sign}(f_w(x))$ where $f_w(x) = w^T x$.

Setting for Two-Layer Neural Network Model with Quadratic "XOR" Ground Truth

► **Data Distribution.** Fix some orthogonal ground truth unit vector directions μ_1 and μ_2 in $\mathbb{R}^d$. Let $x = z + \xi$, where $z \sim \text{Uniform}(\{\mu_1, -\mu_1, \mu_2, -\mu_2\})$ and ξ is uniform on the sphere of radius $\sqrt{d-2}\sigma$ in $d-2$ dimensions, orthogonal to the directions μ_1 and μ_2. Let $y = (\mu_1^T x)^2 - (\mu_2^T x)^2$ for some orthogonal ground truth directions μ_1 and μ_2 (see Figure 2(left)). We denote this distribution of (x, y) on $\mathbb{R}^d \times \{-1, 1\}$ by $\mathcal{D}_{\mu_1,\mu_2,\sigma,d}$. We call this the XOR problem because $y = \text{XOR}\left((\mu_1 + \mu_2)^T x, (-\mu_1 + \mu_2)^T x\right)$. For instance, if $\mu_1 = e_1$ and $\mu_2 = e_2$, then $y = x_1^2 - x_2^2$. As can be seen in Figure 2(left), this distribution is not linearly separable, and so one must use nonlinear model to learn in this setting.

► **Model.** Fix $a \in \{-1, 1\}^m$ so that $\sum_i a_i = 0$. The model is a two-layer neural network with m hidden units and activation function ϕ, parameterized by $W \in \mathbb{R}^{m \times d}$. W (which will be learned) represents the weights of the first layer and a (which is fixed) is the second layer weights. The model predicts $f_W(x) = \sum_{i=1}^m a_i \phi(w_i^T x)$, where $w_i \in \mathbb{R}^d$ denotes the i'th column of W. We work with activations ϕ of the form $\phi(z) = \max(0, z)^h$ for $h \in [1, 2)$, and require that m is divisible by 4[1].

We define a *problem class* of distributions to be a set of data distributions. In this paper, we work with the linear problem class $\Omega_{\sigma,d}^{\text{linear}} := \{\mathcal{D}_{\mu,\sigma,d} : \mu \in \mathbb{R}^d, \|\mu\| = 1\}$, and the quadratic problem class $\Omega_{\sigma,d}^{\text{XOR}} := \{\mathcal{D}_{\mu_1,\mu_2,\sigma,d} : \mu_1 \perp \mu_2 \in \mathbb{R}^d, \|\mu_1\| = \|\mu_2\| = 1\}$. Here $\|\cdot\|$ denotes the ℓ_2 norm.

We will sometimes abuse notation and say that $x \sim \mathcal{D}$ instead of saying that $(x, y) \sim \mathcal{D}$.

Before proceeding, we make some comments on the parameter settings and compare to related work.

Large dimension assumption. In both the linear and non-linear settings, our focus is an overparameterized regime where the dimension d is at least a constant factor times larger than n, the number of training samples. Such an assumption is mild relative to the assumptions made in related work, which require $d = \omega(n)$ (see eg. (Cao et al., 2021; Wang & Thrampoulidis, 2020; Muthukumar et al., 2021; Shamir, 2022; Chatterji & Long, 2021) on linear models; for neural networks, the closest related works of Frei et al. (2022a) and Cao et al. (2022) assume that $d \geq n^2$ or stronger). When the dimen-

[1]The assumption that m is divisible by 4 is for convenience, and can be removed if m is large enough.

sion is sufficiently large (in particular, at least $\omega(n)$), with high probability, the max-margin solution coincides with the min-norm regression solution (see Hsu et al. (2021)), meaning the max-margin solution can be analyzed via a closed-form expression. Our work is fundamentally different from the work on linear classification which operates in the $d = \omega(n)$ regime, because in our setting when $d = \Theta(n)$, these two solutions do not coincide.

Assumption on Data Covariance. Many works on linear classification study more general data models which allow arbitrary decay of the eigenvalues of the covariance matrix (eg. Muthukumar et al. (2021); Wang & Thrampoulidis (2020); Cao et al. (2021)), or variance in the signal direction, that is, $x^T\mu \neq y$ (eg. Shamir (2022)). We work with a simpler distribution, which is still challenging, because it defies existing analyses built on UC or closed-form solutions.

2.1 Background and Definitions on Uniform Convergence

In this subsection, we provide some definitions from Nagarajan & Kolter (2019b) on algorithm-dependent UC bounds. We also provide some definitions and background on margin bounds.

For a loss function $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and a hypothesis h mapping from a domain $\mathcal{X}$ to $\mathbb{R}$, we define the test loss on a distribution $\mathcal{D}$ to be $\mathcal{L}_{\mathcal{D}}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(h(x), y)$. For a set of examples $S = \{(x_i, y_i)\}_{i\in[n]}$, we define $\mathcal{L}_S(h) := \mathbb{E}_{i\in[n]}\mathcal{L}(h(x_i), y_i)$ to be the empirical loss.

Unless otherwise specified, we will use $\mathcal{L}$ to denote the $0/1$ loss, which equals 1 if and only if the signs of the two labels disagree, that is, $\mathcal{L}(y, y') = \mathbf{1}(\text{sign}(y) \neq \text{sign}(y'))$.

Typically in machine learning one considers a global hypothesis class $\mathcal{G}$ that an algorithm may explore (e.g., the set of all two-layer neural networks). A uniform convergence bound, defined below, may hold over a smaller subset $\mathcal{H}$ of $\mathcal{G}$, eg. the subset of networks with bounded norm.

Definition 2.1 (Uniform Convergence Bound)**.** *A uniform convergence bound with parameter ϵ_{unif} for a distribution $\mathcal{D}$, a set of hypotheses $\mathcal{H}$, and loss $\mathcal{L}$ is a bound that guarantees that*

$$\Pr_{S\sim\mathcal{D}^n}[\sup_{h\in\mathcal{H}} |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| \geq \epsilon_{\text{unif}}] \leq \frac{1}{4}. \tag{2.1}$$

A uniform convergence bound can be customized to algorithms by choosing $\mathcal{H}$ to depend on the implicit bias of an algorithm. For instance, if an algorithm $\mathcal{A}$ favors low-norm solutions, one could choose $\mathcal{H}$ to be the set of all classifiers with bounded norm. Of course, if $\mathcal{H}$ is too small, it may not be useful for proving generalization, because $\mathcal{A}$ will never output a solution in $\mathcal{H}$. We formalize the notion of choosing a useful algorithm-dependent set $\mathcal{H}$ as follows.

Definition 2.2 (Useful Hypothesis Class)**.** *A hypothesis class $\mathcal{H}$ is* useful *with respect to an algorithm $\mathcal{A}$ and a distribution $\mathcal{D}$ if $\Pr_{S\sim\mathcal{D}^n}[\mathcal{A}(S) \in \mathcal{H}] \geq \frac{3}{4}$.*

Remark 2.3. *Our definition of a uniform convergence bound on a useful hypothesis class is essentially equivalent to the definition of algorithm-dependent uniform convergence bound in Nagarajan & Kolter (2019b). We introduce new terminology since we use it later in our results on margin bounds.*

More generally, we can have generalization bounds that do not yield the same generalization guarantee for all elements of $\mathcal{H}$. Instead, their guarantee scales with some property of the hypothesis h and the sample S. We call these *data-dependent* bounds. Such bounds are useful if the favorable property is satisfied with high probability by the algorithm of interest.

One specific type of data-dependent bound depends on the margin achieved by the classifier on the training sample. We recall the definition of a margin:

Definition 2.4 (Margin)**.** *The* margin $\gamma(h, S)$ *of a classifier h on a sample S equals* $\min_{(x,y)\in S} yh(x)$.

In certain parameterized hypothesis classes it is useful to define a normalized margin. If f_W is h-homogeneous, that is, $f_{cW}(x) = c^h f_W(x)$ for a positive scalar c, we define the *normalized margin*

$$\bar{\gamma}(f_W, S) := \frac{\gamma(f_W, S)}{\|W\|^h} = \gamma(f_{W/\|W\|}, S), \tag{2.2}$$

where we define the norm $\|W\|$ to equal $\sqrt{\mathbb{E}_{i\in[m]}[\|w_i\|^2]}$, where w_i is the i'th column of W.

We will use $\gamma^*(S)$ to denote the maximum normalized margin. When we are discussing the linear problem, we let $\gamma^*(S)$ be the max-margin over all vectors $w \in \mathbb{R}^d$ with norm 1, that is $\gamma^*(S) := \sup_{w:\|w\|_2 \leq 1} \gamma(S, f_w)$. In the XOR problem, we use $\gamma^*(S)$ to denote the max-margin over all weight matrices $W \in \mathbb{R}^{m \times d}$ with norm 1, that is $\gamma^*(S) := \sup_{W:\|W\| \leq 1} \gamma(S, f_W)$.

Most classical margin bounds prove that the generalization gap can be bounded by a term that scales inversely linearly or quadratically in the margin (Koltchinskii & Panchenko, 2002; Kakade et al., 2009). More generally, we will call margin bounds in which the generalization guarantee scales with $\left(\frac{1}{\gamma(S,f_W)}\right)^p$ for a constant p a *polynomial margin bound*. Such bounds usually rely on proving uniform convergence for a continuous loss that upper bounds the $0/1$ loss. As we will show in the next section, such bounds are also intrinsically limited in regimes where UC fails on the $0/1$ loss.

In contrast to this, in our work, we prove bounds for classifiers that achieve near-maximal margins.

Definition 2.5. *A classifier h is a* $(1-\epsilon)$-max-margin solution *for S if* $\gamma(h, S) \geq (1-\epsilon)\gamma^*(S)$.

We refer to a bound that holds for $(1-\epsilon)$-max-margin solutions as a *extremal margin bound.*

3 Main Results

In the following section, we state our main results for the linear and quadratic problems, and provide intuition for our findings. As illustrated in Figure 1, and in more detail in Figure 2(right), our results show different possibilities for a near max-margin solution depending on the size of $\kappa := \frac{n}{d\sigma^2}$, a signal-to-noise parameter, where σ, d are as in Section 2. When κ is smaller than some threshold κ_{gen} we are not guaranteed to have learning: even a near max-margin solution may not generalize. When κ exceeds κ_{gen} by an absolute constant and when $\sigma^2 \ll 1$, our results show that any near max-margin solution generalizes well. Finally, we show that if κ is smaller than a second threshold κ_{uc}, then uniform convergence approaches will fail to guarantee generalization. All of our results additionally include an overparameterization condition that $d \geq cn$ for a constant c, as is pictured in Fig 2(right).

The exact thresholds κ_{gen} and κ_{uc} depend on the problem class of interest, but in both the linear setting and the nonlinear setting we study, we show that $\kappa_{\text{uc}} > \kappa_{\text{gen}}$. Thus we observe a regime where uniform convergence fails, but generalization still occurs for near max-margin solutions.

For the linear problem, we define the universal constants

$$\kappa_{\text{gen}}^{\text{linear}} := 0 \text{ and } \kappa_{\text{uc}}^{\text{linear}} := 1. \tag{3.1}$$

For the XOR problem with activation relu^h, for $h \in [1, 2)$, we define the constants

$$\kappa_{\text{gen}}^{\text{XOR},h} := \text{ the solution to } \; 2^{\frac{1}{h}}\sqrt{\frac{2}{\kappa}} = \sqrt{\frac{\kappa}{4+\kappa}} + \sqrt{\frac{16}{\kappa\,(4+\kappa)}} \text{ and } \kappa_{\text{uc}}^{\text{XOR},h} := 4. \tag{3.2}$$

The constants are pictured in Figure 2(right) as a function of h. Observe that for $h \in (1, 2)$, we have $\kappa_{\text{gen}}^{\text{XOR},h} < \kappa_{\text{uc}}^{\text{XOR},h}$, and $\kappa_{\text{gen}}^{\text{XOR},h} > 0$. When $h = 1$ and the activation is relu, we have $\kappa_{\text{gen}}^{\text{XOR},h} = \kappa_{\text{uc}}^{\text{XOR},h}$, and thus we do not expect to have a regime where uniform convergence fails, but max-margin solutions generalize. We elaborate more intuitively on why $h > 1$ allows for generalization without UC in Section A.

Our first theorem states that when $\kappa > \kappa_{\text{gen}}$, any near-max-margin solution generalizes.

Theorem 3.1 (Extremal-Margin Generalization for Linear Problem)**.** *Let $\delta > 0$. There exist constants $\epsilon = \epsilon(\delta)$ and $c = c(\delta)$ such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega_{\sigma,d}^{\text{linear}}$ satisfying $\kappa_{gen}^{\text{linear}} + \delta \leq \kappa \leq \frac{1}{\delta}$, and $\frac{d}{n} \geq c$, then with probability $1 - 3e^{-n}$ over the randomness of a training set $S \sim \mathcal{D}^n$, for any $w \in \mathbb{R}^d$ that is a $(1-\epsilon)$-max-margin solution (as in Definition 2.5), we have $\mathcal{L}_{\mathcal{D}}(f_w) \leq e^{-\frac{n}{36d\sigma^4}} + e^{-n/8}$.*

Attentive readers may observe that since $\kappa_{\text{gen}}^{\text{linear}} = 0$, Theorem 3.1 can guarantee asymptotic generalization for some sequences of parameters $(n_i, d_i, \sigma_i)_{i \geq 1}$ even when $\kappa_i = \frac{n_i}{d_i \sigma_i^2} = o_{i\to\infty}(1)$, as long as σ_i^2 decays fast enough. In Theorem C.4 in the appendix, we state a more detailed version of this theorem which states the exact dependence of c and ϵ on δ, yielding precise results for $\kappa = o(1)$.

We prove a similar generalization result for XOR problem learned on two-layer neural networks.

Theorem 3.2 (Extremal-Margin Generalization for XOR on Neural Network)**.** *Let $h \in (1,2)$, and let $\delta > 0$. There exist constants $\epsilon = \epsilon(\delta)$ and $c = c(\delta)$ such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega_{\sigma,d}^{\mathrm{XOR}}$ satisfying $\kappa = \frac{n}{d\sigma^2} \geq \kappa_{gen}^{\mathrm{XOR},h} + \delta$ and $\frac{d}{n} \geq c$, then with probability $1 - 3e^{-n/c}$ over the training set $S \sim \mathcal{D}^n$, for any two-layer neural network with activation function relu^h and weight matrix W that is a $(1-\epsilon)$-max-margin solution (as in Definition 2.5), we have $\mathcal{L}_{\mathcal{D}}(f_W) \leq e^{-\frac{1}{c\sigma^2}}$.*

This theorem guarantees meaningful results whenever σ is small enough. To see this, note that the assumptions of the theorem require that $\frac{d}{n} \in \left[c, \frac{1}{\sigma^2(\kappa_{\mathrm{gen}}^{\mathrm{XOR},h}+\delta)}\right]$. If σ is small enough (in terms of δ), this interval is non-empty. Further, the generalization guarantee is good if σ is small enough (since $\exp(-1/(c\sigma^2))$ tends to 0 as σ approaches 0). For instance consider a setting where $d \gg n$, and $\sigma^2 = \frac{n}{d}$. Then our theorem guarantees that $\mathcal{L}_{\mathcal{D}}(f_W) \ll 1$.

Key intuitions for generalization theorems. We demonstrate the gist of the analysis for the linear problems with some simplifications. It turns out that two special solutions merit particular attention: (i) the good solution $w_{\mathrm{g}} = \mu$ that generalizes perfectly, and (ii) the bad overfitting solution $w_{\mathrm{b}} :\approx \frac{1}{\sqrt{nd}\sigma}\sum_j y_j \xi_j$ that memorizes the "junk" dimension of the data, and satisfies $\xi_i^T w_{\mathrm{b}} \approx \frac{1}{\sqrt{nd}\sigma} y_i |\xi_i|^2 = y_i\sqrt{\frac{d\sigma^2}{n}}$ for all i. [2] We examine the margin of the two solutions and have

$$\bar{\gamma}(w_{\mathrm{g}}, S) = 1 \text{ and } \bar{\gamma}(w_{\mathrm{b}}, S) \approx \sqrt{\frac{d\sigma^2}{n}}. \tag{3.3}$$

At first glance, one might conclude that when $\bar{\gamma}(w_{\mathrm{g}}, S) < \bar{\gamma}(w_{\mathrm{b}}, S)$, the max margin solution will be w_{b}, which does not generalize. However, our key observation is that any (near) max margin solution w always contains a mixture of both w_{g} and w_{b}. When the w_{g} component is small but non-trivial and the w_{b} component is large, the solution can simultaneously generalize but contain a large enough overfitting component to preclude UC.

More concretely, suppose we consider the margin of a linear mixture $w = \alpha w_{\mathrm{g}} + \beta w_{\mathrm{b}}$ satisfying $\alpha^2 + \beta^2 = 1$ so that $\|w\|_2 = 1$. It is easy to see that the margin on the training set is

$$\bar{\gamma}(w, S) = \alpha\bar{\gamma}(w_{\mathrm{g}}, S) + \beta\bar{\gamma}(w_{\mathrm{b}}, S) \tag{3.4}$$

Meanwhile, the margin on an test example x is only slightly affected by w_{b}:

$$\bar{\gamma}(w, x) \approx \alpha\bar{\gamma}(w_{\mathrm{g}}, S) \pm \beta w_{\mathrm{b}}^T x \approx \alpha\bar{\gamma}(w_{\mathrm{g}}, S) \pm \beta\bar{\gamma}(w_{\mathrm{b}}, S)\sqrt{\frac{n}{d}}. \tag{3.5}$$

The effect $w_{\mathrm{b}}^T x$ of the bad solution on the test sample is is smaller than $\bar{\gamma}(w, S)$ by a $\sqrt{\frac{n}{d}}$ factor because x is a high dimensional random vector, and thus mostly orthogonal to w_{b}. Therefore, even if the margin on the training set mostly stems from the bad overfitting solution, that is, $\alpha\bar{\gamma}(w_{\mathrm{g}}, S) < \beta\bar{\gamma}(w_{\mathrm{b}}, S)$, the model may still generalize as long as $\alpha\bar{\gamma}(w_{\mathrm{g}}, S) \geq \beta\bar{\gamma}(w_{\mathrm{b}}, S)\sqrt{\frac{n}{d}}$.

The optimal α, β satisfying $\alpha^2 + \beta^2 = 1$ that maximize the margin turns out to be proportional to the original margin: $\frac{\alpha}{\beta} = \frac{\bar{\gamma}(w_{\mathrm{g}},S)}{\bar{\gamma}(w_{\mathrm{b}},S)}$, yielding a max-margin of $\sqrt{\bar{\gamma}(w_{\mathrm{g}}, S)^2 + \bar{\gamma}(w_{\mathrm{b}}, S)^2} \approx \sqrt{\frac{d\sigma^2+n}{n}}$.

Therefore, we have $\frac{\alpha\bar{\gamma}(w_{\mathrm{g}},S)}{\beta\bar{\gamma}(w_{\mathrm{b}},S)} = \frac{\bar{\gamma}(w_{\mathrm{g}},S)^2}{\bar{\gamma}(w_{\mathrm{b}},S)^2}$. In other words, we should expect reasonable generalization of near-max margin solutions as long as $\frac{\bar{\gamma}(w_{\mathrm{g}},S)}{\bar{\gamma}(w_{\mathrm{b}},S)} > (\frac{n}{d})^{1/4}$, which by eq. 3.3 occurs when $\frac{n}{d\sigma^4} \gg 1$.

In Appendix A, we describe the challenges that arise when adapting these intuitions to nonlinear setting, and our techniques for overcoming them.

Before proceeding to our lower bounds, observe that a typical margin bound for the linear setting would yield $|\mathcal{L}_{\mathcal{D}}(w) - \mathcal{L}_S(w)| \leq \frac{2|x|}{\sqrt{n}\bar{\gamma}(S,w)} \approx \frac{2\sqrt{d}\sigma\sqrt{1+1/\kappa}}{\sqrt{n}}$, which is at least 2 for $\kappa \leq \kappa_{\mathrm{uc}}^{\mathrm{linear}} = 1$.

[2] More precisely, we will choose w_{b} to be the rescaled min-norm vector satisfying $\xi_i^T w_{\mathrm{b}} = y_i$ for all i. This distinction is important in the case when d is only a constant factor larger than n, and the solution $\frac{1}{\sqrt{nd}\sigma}\sum_j y_j\xi_j$ does not necessarily correctly classify the training data.

We now proceed to present our lower bounds, which show when near max-margin solutions may not always generalize, and when UC bounds and polynomial margin bounds are impossible.

If $\kappa < \kappa_{\text{gen}}$, it is possible that a near-max margin solution does not generalize at all. Since $\kappa_{\text{gen}} = 0$ in the linear setting, we only state this result for the XOR problem.

Proposition 3.3 (Region where Max-Margin Generalization not Guaranteed)**.** *Let $h \in (1, 2)$, and let $\epsilon > 0$. There exists a constant $c = c(\epsilon)$ such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega^{\text{XOR}}_{\sigma,d}$ satisfying $\kappa \leq \kappa^{\text{XOR},h}_{gen} - \epsilon$ and $\frac{d}{n} \geq c$, with probability $1 - 3e^{-n/c}$ over $S \sim \mathcal{D}^n$, there exists some W with $\|W\| = 1$ and $\gamma(f_W, S) \geq (1-\epsilon)\gamma^*(S)$ such that $\mathcal{L}_{\mathcal{D}}(f_W) = \frac{1}{2}$.*

Theorems 3.2 and Prop. 3.3 demonstrate that in the XOR problem, there is a threshold in κ above which generalization occurs. If κ is above this threshold, we achieve generalization when $\sigma^2 \ll 1$.

The next proposition states that when $\kappa < \kappa_{\text{uc}}$, any algorithm-dependent uniform convergence bounds will be vacuous, that is, its generalization guarantee will be arbitrarily close to 1. We state our results for the linear and XOR neural network settings together; we state the more complicated XOR result in full and then mention how the linear result differs.

Proposition 3.4 (UC Bounds are Vacuous)**.** *Fix any $h \in (1, 2)$, and $\delta > 0$. For any n, d, σ and $\mathcal{D} \in \Omega^{\text{XOR}}_{\sigma,d}$, if $\kappa^{\text{XOR},h}_{gen} + \delta \leq \kappa \leq \kappa^{\text{XOR},h}_{uc} - \delta$, there exist strictly positive constants $\epsilon = \epsilon(\delta)$ and $c = c(\delta)$ such that the following holds. Let $\mathcal{A}$ be any algorithm that outputs a $(1-\epsilon)$-max-margin two-layer neural network f_W for any $S \in (\mathbb{R}^d \times \{1, -1\})^n$. Let $\mathcal{H}$ be any hypothesis class that is useful for $\mathcal{D}$ (as in Definition 2.2). Suppose that ϵ_{unif} is a uniform convergence bound for $\mathcal{D}$ and $\mathcal{H}$ that is, $\Pr_{S\sim\mathcal{D}^n}[\sup_{h\in\mathcal{H}} |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| \geq \epsilon_{\text{unif}}] \leq 1/4$. Then if $\frac{d}{n} \geq c$ and $n > c$, we must have $\epsilon_{\text{unif}} \geq 1 - \delta$.*

A similar result holds for the linear problem with $\kappa^{\text{linear}}_{gen} + \delta < \kappa < \kappa^{\text{linear}}_{uc} - \delta$ and any $\mathcal{D} \in \Omega^{\text{linear}}_{\sigma,d}$. In this case we achieve the guarantee that $\epsilon_{\text{unif}} \geq 1 - e^{-\frac{n}{36d\sigma^2}} - e^{-n/8}$.

Prop. 3.4 is proved using the same technique as in Nagarajan & Kolter (2019b): we show that with high probability over $S \sim \mathcal{D}^m$, the hypothesis $\mathcal{A}(S)$ has good generalization, but on an "oppositite" dataset $\psi(S)$ with the junk components reversed, the empirical error of $\mathcal{A}(S)$ is close to 1. This large gap between empirical error and generalization forces $\epsilon_{\text{unif-alg}}$ to be large.

Further extending this technique, we can also show the limitations of classical polynomial margin bounds which achieve an bound that scales inversely polynomially with $\gamma(h, S)$. We show that with high probability over $S \sim \mathcal{D}^m$, the hypothesis $\mathcal{A}(\psi(S))$ has a large margin on the set S (a constant fraction times the max-margin), but poor generalization on $\mathcal{D}$. Since any polynomial margin bound cannot predict much better generalization for the max-margin solution than for a solution with a constant-fraction of the max-margin, we conclude that any such margin bound is far from showing good generalization for the max-margin solution.

One subtlety to this approach is that here (unlike in the work of Nagarajan & Kolter (2019b)), the "opposite" data set $\psi(S)$ is defined to be the data set with the signal features reversed. Thus we can only show the limitations of polynomial margin bounds that are useful for *both* $\mathcal{D}$ and for its "oppostite" distribution $\psi(\mathcal{D})$, which has the opposite ground-truth vector(s), which is a slightly stronger assumption than in the work of Nagarajan & Kolter (2019b). [3] Formally, if $\mathcal{D} = \mathcal{D}^{\text{linear}}_{\mu,\sigma}$, then we define $\psi(\mathcal{D}) := \mathcal{D}^{\text{linear}}_{-\mu,\sigma}$. If $\mathcal{D} = \mathcal{D}^{\text{XOR}}_{\mu_1,\mu_2,\sigma}$, then $\psi(\mathcal{D}) := \mathcal{D}^{\text{XOR}}_{\mu_2,\mu_1,\sigma}$.

The following results state that if $\kappa < \kappa_{\text{uc}}$, then certain types of margin bounds cannot yield better than constant test loss on even the max-margin solution.

Proposition 3.5 (Polynomial Margin Bounds Fail for Linear Problem)**.** *Fix $\delta > 0$. For any n, d, σ and $\mathcal{D} \in \Omega^{\text{linear}}_{\sigma,d}$ such that $\kappa^{\text{linear}}_{gen} + \delta < \kappa < \kappa^{\text{linear}}_{uc} - \delta$ and $\frac{d}{n} \geq c$, the following holds. Let $\mathcal{A}$ be any algorithm so that $\mathcal{A}(S)$ outputs a $(1-\epsilon)$-max-margin solution f_w for any $S \in (\mathbb{R}^d \times \{1, -1\})^n$.*

[3] We believe considering such types of margin bounds is natural. Indeed, for most problems, the designer of the generalization bound would not know in advance the ground truth distribution, but might know that the data comes from some problem class, e.g., linearly separable distributions, or linearly separable distributions with a sparse ground truth vector. In such cases, they would likely have to design a generalization bound that holds for data coming from two distributions with opposite ground truths.

If we make this same stronger two-distribution assumption in Prop. 3.4, we can additionally rule out *one-sided* UC bounds, which only upper bound $\mathcal{L}_{\mathcal{D}} - \mathcal{L}_S$.

Let $\mathcal{H}$ be any hypothesis class that is useful for $\mathcal{A}$ (as in Definition 2.2) on both $\mathcal{D}^{\text{linear}}_{\mu,\sigma}$ and $\mathcal{D}^{\text{linear}}_{-\mu,\sigma}$. Suppose that there exists an polynomial margin bound of integer degree p: that is, there is some G that satisfies for $\tilde{D} \in \{\mathcal{D}, \psi(\mathcal{D})\}$,

$$\Pr_{S\sim\tilde{\mathcal{D}}^n}\left[\sup_{h\in\mathcal{H}} \mathcal{L}_{\tilde{\mathcal{D}}}(h) - \mathcal{L}_S(h) \geq \frac{G}{\gamma(h,S)^p}\right] \leq \frac{1}{4}.$$

Then with probability $\frac{1}{2} - 3e^{-n}$ over $S \sim \mathcal{D}^n$, the margin bound is weak even on the max-margin solution, that is, $\frac{G}{\gamma^*(S)^p} \geq \max\left(\frac{1}{c}, 1 - e^{-\frac{\kappa}{36\sigma^2}} - e^{-n/8} - \frac{3\kappa}{c}\right)^p$, which is more than an absolute constant.

This theorem says that no polynomial margin bound will be able to show that the test error of the max-margin solution is less than an absolute constant. We know however from Theorem 3.1 that in this same regime, the test error of the max-margin solution can be arbitrarily small for small enough σ. Thus no polynomial margin bound can predict this behaviour.

The attentive reader again may notice that if $\kappa \to 0$ as n and d grow, but generalization occurs, any such margin bound is vacuous, in that $\frac{G}{\gamma^*(S)^p} \to 1$. In Prop **??**, we prove a more precise version, yielding the exact dependence of c and ϵ on the gap between κ and the boundaries $\kappa^{\text{linear}}_{\text{uc}}$ and $\kappa^{\text{linear}}_{\text{gen}}$.

We achieve a similar result in the XOR setting.

Proposition 3.6 (Polynomial Margin Bounds Fail for XOR on Neural Network)**.** *Fix an integer $p \geq 1$, and any $\epsilon > 0$. There exists $c = c(p, \epsilon)$ such that the following holds for any n, d, σ and $\mathcal{D} \in \Omega^{\text{XOR}}_{\sigma,d}$ with $\kappa^{\text{XOR},h}_{gen} + \epsilon < \kappa < \kappa^{\text{XOR},h}_{uc} - \epsilon$, $\frac{d}{n} \geq c$ and $n \geq c$. Let $\mathcal{H}$ be any hypothesis class such that for $\tilde{\mathcal{D}} \in \{\mathcal{D}, \psi(\mathcal{D})\}$,*

$$\Pr_{S\sim\tilde{\mathcal{D}}^n}[\textit{all } (1-\epsilon)\textit{-max-margin two-layer neural networks } f_W \textit{ for } S \textit{ lie in } \mathcal{H}] \geq 3/4.$$

Suppose that there exists an polynomial margin bound of degree p: that is, there is some G that satisfies for $\tilde{\mathcal{D}} \in \{\mathcal{D}, \psi(\mathcal{D})\}$,

$$\Pr_{S\sim\tilde{\mathcal{D}}^n}\left[\sup_{h\in\mathcal{H}} \mathcal{L}_{\tilde{\mathcal{D}}}(h) - \mathcal{L}_S(h) \geq \frac{G}{\gamma(h,S)^p}\right] \leq \frac{1}{4}.$$

Then with probability $\frac{1}{2} - 3e^{-n/c}$ over $S \sim \mathcal{D}^n$, on the max-margin solution, the generalization guarantee is no better than $\frac{1}{c}$, that is, $\frac{G}{\gamma^*(S)^p} \geq \frac{1}{c}$.

Remark 3.7. *The polynomial margin impossibility results is slightly weaker for the XOR problem. Namely, the hypothesis class $\mathcal{H}$ we consider is larger in the XOR problem: it must contain with probability $\frac{3}{4}$* any *near max-margin solution, instead of just the one output by $\mathcal{A}$.*

The combination of our generalization results and our margin possibility results suggest a phase transition in how the margin size affects generalization. If the margin is near-maximal, Theorems 3.1 and Prop. 3.2 show that we achieve generalization. Meanwhile, the proof of Props 3.5 and 3.6 suggest that solutions achieving a constant factor of the maximum margin may not generalize.

The proofs of all of our results concerning the linear problem are given in Section C. The proofs for the XOR problem are in Section D.

4 Conclusion

In the work, we give novel generalization bounds in settings where uniform convergence provably fails. We use a unified approach of leveraging the extremal margin in both a linear classification setting and a non-linear two-layer neural network setting. Our work provides insight on why memorization can coexist with generalization.

Going beyond our results, it is important to find broader tools for understanding the regime near the boundary of generalization and no generalization. We conclude with several concrete open directions in this vein. One question is how to prove generalization without UC when $d < n$, but the model itself (e.g. a neural network) is overparameterized, and thus can still overfit to the point of UC failing. A second direction asks if we can prove similar results in the non-linear network setting for the solution found by gradient descent, if this solution is not a near max-margin solution. Indeed, in a non-convex landscape, it not guaranteed that that gradient descent will find the max-margin solution.

ACKNOWLEDGMENTS

We thank Jason Lee for helpful discussions. MW acknowledges the support of NSF Grant CCF-1844628 and a Sloan Research Fellowship. TM is supported by NSF IIS 2045685.

REFERENCES

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.

Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

Peter L Bartlett and Philip M Long. Failures of model-dependent generalization bounds for least-norm interpolation. *Journal of Machine Learning Research*, 22(204):1–15, 2021.

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

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning (ICML)*, 2018.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

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

Satrajit Chatterjee and Piotr Zielinski. On the generalization mystery in deep learning. *arXiv preprint arXiv:2203.10036*, 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.

Alex Damian, Tengyu Ma, and Jason Lee. Label noise sgd provably prefers flat global minimizers, 2021.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *arXiv preprint arXiv:2202.05928*, 2022a.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *arXiv preprint arXiv:2202.07626*, 2022b.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

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

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9461–9471, 2018b.

Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Jeff Z HaoChen, Colin Wei, Jason D Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*, 2020.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension bounds for piecewise linear neural networks. In *Conference on Learning Theory*, pp. 1064–1068. PMLR, 2017.

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

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009.

Frederic Koehler, Lijia Zhou, Nathan Srebro, et al. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Tomer Koren, Roi Livni, Yishay Mansour, and Uri Sherman. Benign underfitting of stochastic gradient descent. *arXiv preprint arXiv:2202.13361*, 2022.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019a.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *arXiv preprint arXiv:1712.09203*, pp. 2–47, 2017.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pp. 11669–11680, 2019b.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

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, 2021.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 605–638. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/mou18a.html.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 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, 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*, pp. 3051–3059. PMLR, 2019.

Vaishnavh Nagarajan and J Zico Kolter. Deterministic pac-bayesian generalization bounds for deep networks via generalizing noise-resilience. *arXiv preprint arXiv:1905.13344*, 2019a.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *arXiv preprint arXiv:1902.04742*, 2019b.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pp. 11013–11023, 2019.

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*, pp. 7263–7272. PMLR, 2020.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401. PMLR, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017a.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017b.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Ohad Shamir. The implicit bias of benign overfitting. *arXiv preprint arXiv:2201.11489*, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: abundance of support vectors, benign overfitting and regularization. *arXiv preprint arXiv:2011.09148*, 2020.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34, 2021.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *Advances in Neural Information Processing Systems*, pp. 9722–9733, 2019a.

Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019b.

Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. In *International Conference on Learning Representations (ICLR)*, 2020.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pp. 9709–9721, 2019.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. *arXiv preprint arXiv:2002.12915*, 2020.

Blake Woodworth, Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Kernel and deep regimes in overparametrized models. *Conference on Learning Theory (COLT)*, 2020.

Zitong Yang, Yu Bai, and Song Mei. Exact gap between generalization error and uniform convergence in random feature models. In *International Conference on Machine Learning*, pp. 11704–11715. PMLR, 2021.

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, Danica J Sutherland, and Nati Srebro. On uniform convergence and low-norm interpolation learning. *Advances in Neural Information Processing Systems*, 33:6867–6877, 2020.

Appendices

LIST OF APPENDICES

A PROOF OVERVIEW

In our proof overview, we focus on the linear problem. While basic steps and intuitions remain the same for the more complicated neural network problem, we add explanation of where we need additional techniques or insights.

The starting observation is that any solution w can be decomposed into a *signal* component and a *overfitting* component. For the linear problem, lets call those components u and v respectively, where u is in the subspace containing μ, and v is orthogonal to μ, such that $w = u + v$. Conveniently in the linear problem, we have $f_w(x) = w^T x = f_u(x) + f_v(x)$. The proof of our main results can be divided into three main parts, which are sketched in the next three subsections.

A.1 THE OVERFITTING COMPONENT ONLY SLIGHTLY AFFECTS GENERALIZATION.

Since the "junk" features (orthognal to μ) are high dimensional and have smaller variance, on a random new sample drawn from the population distribution $\mathcal{D}$, we have $f_w(x) \approx f_u(x)$. We formalize the affect on generalization in the following lemma, which shows that non-trivial generalization can occur so long as $\|v\|_2 \lesssim \frac{1}{\sigma} u^T \mu$. Notice that if $\sigma^2 \ll 1$, this means the v component can be a great deal larger than the u component without affecting generalization.

Lemma A.1. *Fix a distribution $\mathcal{D}_{\mu,\sigma,d} \in \Omega^{\text{linear}}_{\sigma,d}$. For $w \in \mathbb{R}^d$, let $w = u + v$ as above. Let $q := \frac{u^T \mu}{\|v\|_2}$. Then for $x \sim \mathcal{D}_{\mu,\sigma,d}$,*

$$\Pr[|f_v(x)| \geq y f_u(x)] \leq 2e^{-\frac{q^2}{8\sigma^2}} + \exp(-8d).$$

We show a similar lemma for the two-layer neural network, proved in Section D.4.1. Recall that for a two-layer neural network with weight matrix $W \in \mathbb{R}^{m\times d}$, we define $f_W(x) = \mathbb{E}_{i\in[m]} a_i \phi(w_i^T x)$, where the weights a_i are fixed, and w_i is the ith row of W. Recall that $\phi(z) = \max(0,z)^h$ for $h \in (1,2)$.

Lemma A.2. *Fix a distribution $\mathcal{D}_{\mu_1,\mu_2,\sigma,d} \in \Omega^{h,\mathrm{XOR}}_{\sigma,d}$. For $W \in \mathbb{R}^{m\times d}$, let $W = U + V$ where V is orthogonal to the subspace containing μ_1 and μ_2. Then for some universal constant c, for any $t \geq 1$, with probability at least $1 - e^{-ct}$, on a random sample $x \sim \mathcal{D}_{\mu_1,\mu_2,\sigma,d}$,*

$$|f_W(x) - f_U(x)| \leq (8\|U\| + 3)\,(t+1)\sigma^2\|V\|^2 + 2\left((t+1)\sigma^2\|V\|^2\right)^{\frac{h}{2}}.$$

A.2 Any near-max-margin solution should leverage both signal and overfitting components.

A max-margin solution aims to maximize the minimum margin of any training example while holding the norm of the solution constant. To get a sense of why such a solution must leverage both signal and overfitting components, we consider first what would happen if we used a signal or overfitting component alone.

In the linear problem, setting v to be zero and only considering the signal direction, it is easy to check that the max-margin solution is achieved by setting $u = \mu$, leading to a margin γ_g of 1. We call this good solution w_{g}.

If we optimize in the direction orthogonal to μ alone and set u to zero, the max-margin solution can be shown to found by choosing v to be very near the vector $\frac{\sum_j y_j \xi_j}{\|\sum_j y_j \xi_j\|_2}$, which achieves a margin γ_b of roughly $\frac{\sigma}{\sqrt{n}}$. We call this bad overfitting solution w_{g}.

Depending on the choice of σ, the margin γ_b might be larger that γ_g. Fortunately, this does not preclude generalization. Indeed, we will show that combining these two solutions achieves an even larger margin! Consider constructing the solution $\hat{w} = \alpha w_{\mathrm{g}} + \beta w_{\mathrm{b}}$ and $\alpha^2 + \beta^2 = 1$. It is easy to check since $f_{\hat{w}}(x) = f_u(x) + f_v(x)$ that $\hat{w}$ achieves a margin of $\alpha + \beta$. The following simple optimization program characterizes the optimal trade-off between α and β:

$$\max \alpha\gamma_g + \beta\gamma_b \tag{A.1}$$

$$\alpha^2 + \beta^2 \leq 1. \tag{A.2}$$

Analyzing this program shows that optimal values are achieved by choosing $\frac{\alpha}{\beta} = \frac{\gamma_g}{\gamma_b}$, which suggests that the max-margin solution will include a significant component both of w_{g} and w_{b}.

While this alone is not enough to prove that *any* near-max-margin solution has a significant component of μ, we can extend this argument to show that any solution that achieves a margin larger than γ_b must include some component in the signal direction, and in particular, this component must be in the μ-direction. This is formalized in Lemma C.2

The linear problem had two nice properties which unfortunately we will not be able to leverage in the non-linear problem:

1. It is easy to understand the affect of linearly combining solutions from the signal space and the junk space. That is, for any $w = u + v$, we have $f_w(x) = f_u(x) + f_v(x)$.
2. Any component in the signal subspace which improves the margin is guaranteed to stand alone as a good enough solution. That is, for any v, if the margin of f_{u+v} is better than the margin of f_v, then it must be the case that f_u generalizes. This is because any component in the signal subspace which improves the margin must be in the direction of μ.

In the XOR problem, the failure of (1) to hold is challenging because it turns out that adding the max-margin solutions in the signal space (which we call the "good" solution W_g) to the max-margin solution in the orthogonal space (which we call the "bad" solution W_b) has the affect of partially *cancelling* each other's margins out. Ultimately, we will resolve this by showing that we can construct an alternate bad solution W_b' that does not cancel out the margin at all when combined with the good solution W_g. Even though alone W_b' has a slightly worse margin than W_b, we show that the benefit from combining W_g and W_b' overcomes this loss. This benefit scales with the convexity of the activation ϕ is the positive region. This is why we require that h, the power of the activation, is strictly greater than 1.

The failure of (2) to hold is challenging because there are many weight matrices U in the μ_1, μ_2 subspace that may improve the margin, but not stand alone well. For instance, U might just use the μ_1 direction and not the μ_2 direction, and thus still improve the margin, but not stand well alone. Fortunately, we can rule out this behavior by arguing, similarly to before, that any solution that does not use both signal directions equally cannot exceed a certain (non-maximal) margin. This argument takes the form of a series of lemmas presented in section D.1.

Even more challenging however is the potential to have a component W_1 that uses its component in the u_1-direction (or analogously in the u_2-direction) *only to improve the margin on points from one of the two positive clusters* (see Figure 2). We are able to rule out this behavior by reducing our understanding of max-margin solutions on the neural net to a simpler 3-variable optimization problem which concerns a single neuron $w \in \mathbb{R}^d$ and a pair of training examples x_j and $x_{j'}$ from the two positive clusters (see Figure 2(left)), that is, $x_j = \mu_1 + \xi_j$, and $x_{j'} = -\mu_1 + \xi_{j'}$.

Definition A.3 (Trivariate Subproblem)**.**

$$\max \phi(b+c) + \phi(-b+d) : \tag{A.3}$$

$$b^2 + \frac{\kappa}{4}(c^2+d^2) \leq 1 \tag{A.4}$$

In this problem, the variable b represents $\mu^T w$, the strength of the signal component. The variable c represents $w^T \xi_j$, and d represents $w^T \xi_{j'}$, the strengths of the overfitting components. This 3-variable optimization problem can be viewed as an analog of the optimization problem in eq. A.1.

The following lemma argues that if the neuron w is on average good for both x_j and $x_{j'}$, ie., the objective is large, then *for both* x_j and $x_{j'}$, a constant fraction of the activation must be explained by the component of w in the μ_1 direction.

Lemma A.4 (Simplification of Lemma D.15)**.** *Suppose $\phi(x) = \max(0,x)^h$ for $1 < h < 2$ and $\kappa > \kappa_{gen}^{XOR,h}$. Then for any a sufficiently small constant $\epsilon = \epsilon(\kappa)$, any $(1-\epsilon)$-optimal solution to the program in Definition A.3 satisfies*

1. *$\phi(b) \geq \Omega(1)\phi(b+c)$; and*
2. *$\phi(-b) \geq \Omega(1)\phi(-b+d)$.*

We note that this lemma is also the key step in arguing that signal solutions and in the overfitting solutions can be combined efficiently enough that using a non-zero component in the signal-subspace is effective.

In Section D, we flesh out this argument in detail, making rigorous the reduction from the full neural net and the full data-set to a the sub-problem in Definition D.6.

A.3 Any near-max-margin solution should have a large enough overfitting component to preclude UC.

In order to show the failure of standard uniform convergence bounds on the problems we consider, we must argue that the overfitting component of any near-max-margin solution roughly exceeds the size of the signal component, as outlined in the Intuition section. The argument described in the previous subsection guarantees this: for this linear problem it is enough to show that $\gamma_b \geq \gamma_g$, and for the non-linear problem, we will need to leverage the full version of Lemma D.15, which shows that c and d are large relative to b.

A larger overfitting component than signal component ensures the phenomenon similar to the one described in the work of Nagarajan and Kolter: if the signal component of the data were changed, but the junk feature stayed the same, the data would still be classified correctly by the classifier learned on the original data. Showing this phenomenon is the key step in proving Theorems 3.4. To prove the margin lower bounds, we also need to show that the this "opposite" dataset is still classified with a large margin by the original classifier.

B Auxiliary Lemmas

If M is a matrix, we use $\|M\|_2$ to denote the spectral norm of M.

We will use the following lemma throughout, on the concentration of random covariance matrices in both the linear and XOR problems.

Lemma B.1 (Concentration of Random Covariance Matrix)**.** *There exists a universal constant $C_{B.1}$ such that the following holds. Suppose $d > C_{B.1}^2 n$ and $\Xi \in \mathbb{R}^{d \times n}$, where each column of Ξ is a random vector distributed uniformly on the sphere of radius $\sqrt{d}\sigma$ in d dimensions. There exists a universal constant $C_{B.1}$ such that with probability at least $1 - 3e^{-n}$ the following two events hold:*

1. *$\|\frac{1}{\sigma^2 d}\Xi^T\Xi - I\|_2 \leq C_{B.1}\sqrt{\frac{n}{d}}$*
2. *For any $c \in \mathbb{R}^n$, if $v \in \mathbb{R}^d$ is the minimum-norm vector v satisfying $\Xi^T v = c$ exists and has $\|v\|_2^2 \in \frac{\|c\|_2^2}{\sigma^2 d}\left[\frac{1}{1+C_{B.1}\sqrt{\frac{n}{d}}}, \frac{1}{1-C_{B.1}\sqrt{\frac{n}{d}}}\right]$.*

The lemma still holds if the rows of Ξ are instead drawn i.i.d. from a sphere of dimension $d-a$ of radius $\sqrt{d-a}\sigma$ in any subspace of dimension $d-a$, for $a \in \{1,2\}$.

Proof. We prove this using a similar result for matrices with i.i.d. entries. Observe that $\Xi = ZD$, where $Z \in \mathbb{R}^{d\times n}$ is a random matrix with i.i.d. normal entries from $\mathcal{N}(0,\sigma^2)$, and $D \in \mathbb{R}^{n\times n}$ is the diagonal matrix with $D_{jj} = \frac{\sigma\sqrt{d}}{\|z_j\|_2}$, where z_j is the jth column of Z.

Thus

$$\left\|\frac{1}{\sigma^2 d}\Xi^T\Xi - I\right\|_2 = \left\|\frac{1}{\sigma^2 d}DZ^TZD - I\right\|_2 \tag{B.1}$$

$$\leq \|D\|_2^2 \left\|\frac{1}{\sigma^2 d}Z^TZ - D^{-2}\right\|_2 \tag{B.2}$$

$$\leq \|D\|_2^2 \left(\left\|\frac{1}{\sigma^2 d}Z^TZ - I\right\|_2 + \left\|D^{-2} - I\right\|_2\right) \tag{B.3}$$

Vershynin (2018) (see Theorem 3.1.1 and the discussion thereafter), states that for any j, for some universal constant C, with probability $1-e^{-2n}$, $|\frac{\|z_j\|^2}{\sigma^2 d} - 1| \leq C\sqrt{\frac{n}{d}}$. Vershynin (2018) (Ex. 4.7.3) also guarantees that with probability $1-2e^{-n}$, $\left\|\frac{1}{\sigma^2 d}Z^TZ - I\right\|_2 \leq C\sqrt{\frac{n}{d}}$. Unioning over the the first event occurring for all j, and the matrix concentration even happening, we have (for some new constant C), with probability

$$1 - 2e^{-n} - ne^{-2n} \geq 1 - 3e^{-n}, \tag{B.4}$$

$\left\|\frac{1}{\sigma^2 d}\Xi^T\Xi - I\right\|_2 \leq C\sqrt{\frac{n}{d}}$.

For the second conclusion, it well-known that the min-norm solution to overparameterized linear regression problem $X^Ta = b$ satisfies $\arg\min_{a:X^Ta=b}\|a\|_2 = X(X^TX)^\dagger b$, where $\dagger$ denotes the pseudo-inverse (see eg. Bartlett et al. (2020), page 5). Observe that $\Xi^T\Xi$ is invertible since $d > C_{B.1}^2 n$ and thus $C_{B.1}\sqrt{\frac{n}{d}} < 1$. It follows by that we can solve explicitly for v, yielding $v = \Xi(\Xi^T\Xi)^{-1}c$. Hence,

$$\|v\|_2^2 = c^T(\Xi^T\Xi)^{-1}\Xi^T\Xi(\Xi^T\Xi)^{-1}c = c^T(\Xi^T\Xi)^{-1}c \in \frac{\|c\|_2^2}{\sigma^2 d}\left[\frac{1}{1+C_{B.1}\sqrt{\frac{n}{d}}}, \frac{1}{1-C_{B.1}\sqrt{\frac{n}{d}}}\right]. \tag{B.5}$$

To see that the lemma applies if the rows of Ξ are orthogonal to some subspace, it suffices to assume the subspace is spanned by e_1 or e_1 and e_2. Thus one can view the matrix Ξ as being in $\mathbb{R}^{d-a\times n}$, and the conclusion follows by replacing d with $d-a$. The constant $C_{B.1}$ can be adjusted so that the conclusion written holds for $a \in \{1,2\}$. □

Lemma B.2. *Let $\phi(x) = \max(0,x)^h$ for $h \in (1,2)$. For any values s, t, we have $\phi(s+t) \leq (\phi(s)+\phi(t))2^{h-1}$*

Proof. By homogeneity of ϕ,

$$\frac{\phi(s)+\phi(t)}{\phi(s+t)} = \frac{\phi(\frac{s}{s+t})+\phi(\frac{t}{s+t})}{\phi(1)}, \tag{B.6}$$

so we need to show that $\phi(a)+\phi(b) \geq 2^{1-h}$ for any $a+b=1$. Since ϕ is convex, subject to this linear constraint, by the KKT condition, the minimum is attained when $\phi'(a) = \phi'(b)$ which occurs when $a = b = \frac{1}{2}$. (Note, we cannot have $\phi'(a) = \phi'(b) = 0$, since at least one of a and b must be positive. □

C Proofs for Linear Problem

Throughout this section, since we are only concerned with the linear problem, we will abbreviate $\Omega = \Omega_{\sigma,d}^{\text{linear}}$, $\kappa_{\text{uc}} = \kappa_{\text{uc}}^{\text{linear}}$, and $\kappa_{\text{gen}} = \kappa_{\text{gen}}^{\text{linear}}$.

C.1 Technical Lemmas

Throughout the following subsection, we assume $\mathcal{D}_{\mu,\sigma,d}$ is fixed. For a vector $w \in \mathbb{R}^d$, we let $u = \mu\mu^T w$ and $v = (I - \mu\mu^T)w$, such that $w = u + v$.

Our training data is given by the matrix (X, y), where $X \in \mathbb{R}^{d\times n}$ and $y \in \mathbb{R}^n$, where $x_j = Xe_j$, and (x_j, y_j) denotes the jth training sample. Recall that we have $x_j = z_j + \xi_j$, where $z_j = \mu y_j$, $y_j \sim \text{Uniform}(-1,1)$, and ξ_j is uniformly distributed on the sphere of radius $\sigma\sqrt{d-1}$ in the $d-1$ dimensions orthogonal to μ. We use $\Xi \in \mathbb{R}^{d\times n}$ to denote the matrix with columns ξ_j.

Lemma C.1. *On the event that the conclusion of Lemma B.1 holds for Ξ, for any v,*

$$\min_j y_j v^T \xi_j \leq \frac{\|v\|_2 \sqrt{\left(1 + C_{B.1}\sqrt{\frac{n}{d}}\right)}}{\sqrt{\kappa}} \tag{C.1}$$

Proof. Let $\gamma := \min_j y_j v^T \xi_j$ and suppose $\gamma > 0$; otherwise the lemma is immediate. Observe that $\|v\|_2$ must be larger than the norm of least norm vector (call it w) achieving $w^T \xi_j y_j \geq \gamma$ for all j, which implies that $|w^T \xi_j| \geq \gamma$ for all j. Let $c := w^T \Xi$, such that $\|c\|^2 \geq n\gamma^2$. By Lemma B.1, on this event of the lemma,

$$\|v\|^2 \geq \|w\|^2 \geq \frac{1}{1 + C_{B.1}\sqrt{\frac{n}{d}}} \frac{\|c\|^2}{d\sigma^2} \geq \frac{1}{1 + C_{B.1}\sqrt{\frac{n}{d}}} \frac{n\gamma^2}{d\sigma^2} \geq \frac{\kappa}{1 + C_{B.1}\sqrt{\frac{n}{d}}} \left(\min_j y_j v^T \xi_j\right)^2. \tag{C.2}$$

The conclusion follows. □

Let $\kappa_{\text{gen}} := 0$, and let $\kappa_{\text{uc}} := 1$.

The following is our main technical lemma. It shows that if the margin is near-optimal and $\kappa > \kappa_{\text{gen}}$, then w must have a large component in the μ direction. If additionally $\kappa < \kappa_{\text{uc}}$, then the spurious component must explain more than half of the margin on every data point, or even more if κ is very small.

Lemma C.2 (Main Technical Lemma)**.** *For any $\kappa > \kappa_{gen}$, there exist a universal constant c such that if $\epsilon \leq \frac{1}{c}\min\left(\kappa, \frac{1}{\kappa}\right)$ and $\sqrt{\frac{n}{d}} \leq \frac{1}{c}\min\left(\kappa, \frac{1}{\kappa}\right)$, with probability $1 - 3e^{-n}$ over (X, y), for any $(1-\epsilon)$-margin maximizing solution w with $\|w\| = 1$ we have*

$$\frac{w^T \mu}{\|v\|_2 \sqrt{\kappa}} \geq \frac{2\sqrt{2}}{3}. \tag{C.3}$$

If additionally $\kappa < \kappa_{uc}$ and $\epsilon \leq \frac{1}{c}\min\left(\kappa, \frac{1}{\kappa}, (\kappa_{uc} - \kappa)^2\right)$ and $\sqrt{\frac{n}{d}} \leq \frac{1}{c}\min\left(\kappa, \frac{1}{\kappa}, (\kappa_{uc} - \kappa)^2\right)$, then for every j,

$$y_j v^T \xi_j \geq \max\left(1 + \frac{1}{c}, \frac{1}{2\kappa}\right) w^T \mu = \max\left(1 + \frac{1}{c}, \frac{1}{2\kappa}\right) y_j w^T z_j. \tag{C.4}$$

Proof. We condition on the events that the outcome of Lemma B.1 (and hence Lemma C.1) hold of Ξ, which occurs with probability at least $1 - 3e^{-n}$. By Lemma C.1,

$$\gamma(w, S) = w^T \mu + \min_j y_j v^T \xi_j \tag{C.5}$$

$$\leq w^T \mu + \frac{\|v\|_2 \sqrt{\left(1 + C_{B.1}\sqrt{\frac{n}{d}}\right)}}{\sqrt{\kappa}}. \tag{C.6}$$

Further, we can lower bound the max-margin by constructing a solution $\hat{w}$ in the following way:

Let $\hat{w} = \frac{\hat{u}+\hat{v}}{\|\hat{u}+\hat{v}\|}$, where $\hat{u} = n\mu$, and $\hat{v}$ is the min-norm solution to $\hat{v}^T \Xi = (d\sigma^2)y$. Since the conclusion of Lemma B.1 holds, we have $\|\hat{v}\|^2 \leq \frac{n\sigma^2 d}{1 - C_{B.1}\sqrt{\frac{n}{d}}}$.

Thus we have

$$\gamma^*(S) \geq \left(1 - C_{B.1}\sqrt{\frac{n}{d}}\right) \frac{n + d\sigma^2}{\sqrt{n^2 + n\sigma^2 d}} \tag{C.7}$$

$$= \left(1 - C_{B.1}\sqrt{\frac{n}{d}}\right) \sqrt{1 + \frac{1}{\kappa}}. \tag{C.8}$$

We prove the first conclusion of the lemma first. Putting together Equations C.5 and C.7, we have

$$w^T \mu + \frac{\|v\|_2 \sqrt{\left(1 + C_{B.1}\sqrt{\frac{n}{d}}\right)}}{\sqrt{\kappa}} \geq (1-\epsilon)\gamma^*(S) \geq (1-\epsilon)\sqrt{1 + \frac{1}{\kappa}} \left(1 - C\sqrt{\frac{n}{d}}\right). \tag{C.9}$$

Thus for some (different) universal constant C, we have

$$w^T \mu + \frac{\|v\|_2}{\sqrt{\kappa}} \geq \left(1 - \epsilon - C\sqrt{\frac{n}{d}}\right) \sqrt{1 + \frac{1}{\kappa}} \tag{C.10}$$

$$= \left(1 - \epsilon - C\sqrt{\frac{n}{d}}\right) \sqrt{1 + \frac{1}{\kappa}} \left(\sqrt{(w^T \mu)^2 + \|v\|^2}\right) \tag{C.11}$$

For the remainder of the proof, let $\epsilon' := \epsilon + C\sqrt{\frac{n}{d}}$. Letting $q = \frac{w^T\mu}{\|v\|_2}$, we have

$$q + \frac{1}{\sqrt{\kappa}} \geq \left(1-\epsilon'\right)\sqrt{1+\frac{1}{\kappa}}\left(\sqrt{1+q^2}\right). \tag{C.12}$$

Squaring and rearranging terms, we have

$$q^2\left(1-\left(1-\epsilon'\right)^2\left(1+\frac{1}{\kappa}\right)\right) + q\left(\frac{2}{\sqrt{\kappa}}\right) + \left(\frac{1}{\kappa} - \left(1-\epsilon'\right)^2\left(1+\frac{1}{\kappa}\right)\right) \geq 0, \tag{C.13}$$

or equivalently,

$$a\left(\frac{q}{\sqrt{\kappa}}\right)^2 + 2\left(\frac{q}{\sqrt{\kappa}}\right) + c \geq 0, \tag{C.14}$$

where

$$a := \left(\kappa\left(1-\left(1-\epsilon'\right)^2\right) - \left(1-\epsilon'\right)^2\right) \tag{C.15}$$

$$c := \frac{1}{\kappa} - \left(1-\epsilon'\right)^2\left(1+\frac{1}{\kappa}\right), \tag{C.16}$$

Claim C.3. *For a small enough constant δ, for any $\kappa, n, d, \epsilon > 0$ such that $\sqrt{\frac{n}{d}} \leq \frac{\delta\min(\kappa,1/\kappa)}{2C}$ and $\epsilon \leq \frac{\delta\min(\kappa,1/\kappa)}{2}$, we have $a < 0$, and $c \leq -\frac{2\sqrt{2}}{3}$.*

Proof. Not that the conditions of the claim imply that $\epsilon' \leq \delta\min(\kappa, 1/\kappa)$. Thus for a small enough δ, we have

$$\left(1-\epsilon'\right)^2 \geq \left(1-\delta\min\left(\kappa,\frac{1}{\kappa}\right)\right)^2 \geq 1 - 3\delta\min\left(\kappa,\frac{1}{\kappa}\right). \tag{C.17}$$

Thus for a small enough δ,

$$c = \frac{1}{\kappa} - \left(1-\epsilon'\right)^2\left(1+\frac{1}{\kappa}\right) \tag{C.18}$$

$$\leq \frac{1}{\kappa} - \left(1-3\delta\min\left(\kappa,\frac{1}{\kappa}\right)\right)\left(1+\frac{1}{\kappa}\right) \tag{C.19}$$

$$\leq -1 + 3\delta\min\left(\kappa,\frac{1}{\kappa}\right)\left(1+\frac{1}{\kappa}\right) \tag{C.20}$$

$$\leq -1 + 6\delta < -\frac{2\sqrt{2}}{3}, \tag{C.21}$$

and similarly,

$$a = \kappa\left(1-\left(1-\epsilon'\right)^2\right) - \left(1-\epsilon'\right)^2 \tag{C.22}$$

$$\leq \kappa\left(3\delta\min(\kappa,1/\kappa)\right) - \left(1-3\delta\min(\kappa,1/\kappa)\right) \tag{C.23}$$

$$\leq -1 + 6\delta. \tag{C.24}$$

□

Viewing Equation C.14 as a quadratic function f of $\frac{q}{\sqrt{\kappa}}$, if the claim above holds, then

$$\frac{q}{\sqrt{\kappa}} \geq -\frac{f(0)}{f'(0)} = \frac{-c}{2} \geq \frac{\sqrt{2}}{3}. \tag{C.25}$$

Since q was defined to be $\frac{w^T\mu}{\|v\|}$, this proves the first part of the lemma for ϵ, $\sqrt{\frac{n}{d}} \leq \frac{\min(\kappa,\frac{1}{\kappa})}{c}$ for a constant c large enough.

We perform a similar argument for the second conclusion. Observe that by plugging in the contents of Equation C.5 into Equation C.7, we have for some constant C (whose value changes throughout this equation,

but does not depend on κ),

$$w^T\mu + \min_j y_j v^T \xi_j \geq \left(1 - \epsilon - C\sqrt{\frac{n}{d}}\right)\sqrt{1+\frac{1}{\kappa}}\sqrt{(w^T\mu)^2 + \|v\|^2} \tag{C.26}$$

$$\geq \left(1 - \epsilon - C\sqrt{\frac{n}{d}}\right)\sqrt{1+\frac{1}{\kappa}}\sqrt{(w^T\mu)^2 + \frac{\kappa\left(\min_j y_j v^T\xi_j\right)^2}{1 + C_{B.1}\sqrt{\frac{n}{d}}}} \tag{C.27}$$

$$\geq \left(1 - \epsilon - C\sqrt{\frac{n}{d}}\right)\sqrt{1+\frac{1}{\kappa}}\sqrt{(w^T\mu)^2 + \kappa\left(\min_j y_j v^T\xi_j\right)^2} \tag{C.28}$$

Let $r := \frac{\min_j y_j v^T \xi_j}{w^T\mu}$. Dividing through by $w^T\mu$ and squaring, we obtain:

$$(r+1)^2 \geq \left(1 - \epsilon - C\sqrt{\frac{n}{d}}\right)^2\left(1+\frac{1}{\kappa}\right)(1+\kappa r^2), \tag{C.29}$$

Rearranging, and multiplying by κ, we have

$$(r\kappa)^2\left(\frac{1}{\kappa} - \left(1+\frac{1}{\kappa}\right)\left(1 - \epsilon - C\sqrt{\frac{n}{d}}\right)^2\right) + 2(r\kappa) + \left(\kappa - (\kappa+1)\left(1 - \epsilon - C\sqrt{\frac{n}{d}}\right)^2\right) \geq 0 \tag{C.30}$$

Now, for $\kappa < \kappa_{\text{uc}} = 1$, for a small enough constant δ (independent of κ), if $\sqrt{\frac{n}{d}} \leq \frac{\delta\kappa(\kappa_{\text{uc}}-\kappa)^2}{2C}$ and $\epsilon \leq \frac{\delta\kappa(\kappa_{\text{uc}}-\kappa)^2}{2}$, we $q = r\kappa$, we have

$$q^2\left(\frac{1}{\kappa} - \left(1+\frac{1}{\kappa}\right)\left(1 - 3\delta(\kappa_{\text{uc}}-\kappa)^2\kappa\right)\right) + 2q + \left(\kappa - (\kappa+1)\left(1 - 3\delta(\kappa_{\text{uc}}-\kappa)^2\kappa\right)\right) \geq 0, \tag{C.31}$$

so

$$-q^2\left(1 - 6\delta(\kappa_{\text{uc}}-\kappa)^2\right) + 2q - \left(1 - 6\delta(\kappa_{\text{uc}}-\kappa)^2\right) \geq 0, \tag{C.32}$$

or

$$-q^2 + \frac{2q}{(1 - 6\delta(\kappa_{\text{uc}}-\kappa)^2)} - 1 \geq 0. \tag{C.33}$$

Let $x = 6\delta(\kappa_{\text{uc}} - \kappa)^2$. The smallest root of this equation is given by

$$\frac{-(2+x) + \sqrt{(2+x)^2 - 4}}{-2} = \frac{2 + x - \sqrt{4x - x^2}}{2} \geq 1 - \sqrt{x}, \tag{C.34}$$

so $q \geq 1 - \sqrt{6\delta}(\kappa_{\text{uc}} - \kappa)$.

Thus for a constant δ small enough, we have

$$\frac{\min_j y_j v^T \xi_j}{w^T\mu} = \frac{q}{\kappa} > \frac{1 + \sqrt{6\delta}(\kappa - 1)}{\kappa} > \max\left(1, \frac{1}{2\kappa}\right). \tag{C.35}$$

□

The following lemma shows that the influence of the v on the label of a test example is small.

Lemma A.1. *Fix a distribution $\mathcal{D}_{\mu,\sigma,d} \in \Omega^{\text{linear}}_{\sigma,d}$. For $w \in \mathbb{R}^d$, let $w = u + v$ as above. Let $q := \frac{u^T\mu}{\|v\|_2}$. Then for $x \sim \mathcal{D}_{\mu,\sigma,d}$,*

$$\Pr[|f_v(x)| \geq y f_u(x)] \leq 2e^{-\frac{q^2}{8\sigma^2}} + \exp(-8d).$$

Proof. For any x, we have $f_u(x) = u^T\mu$. Now $f_v(x)$ is distributed like $\|v\|\sigma\sqrt{d-1}\frac{X}{\sqrt{X^2+Y}}$, where $X \sim \mathcal{N}(0,1)$, and Y is a Chi-square random variable with $d-2$ degrees of freedom. Thus we can bound the probability that $|f_v(x)| \geq y f_u(x)$ by the probability that $\|v\|\sigma X \geq \frac{1}{2} y f_u(x)$ plus the probability that $X^2 + Y$ is smaller than $\frac{d-1}{4}$.

Thus we have

$$\Pr[|f_v(x)| \geq y f_u(x)] \leq \Pr[|\mathcal{N}(0, \sigma^2\|v\|_2^2)| \geq \frac{1}{2} y u^T\mu] + \Pr\left[X^2 + Y \leq \frac{d}{4}\right] \tag{C.36}$$

$$= \Pr[|\mathcal{N}(0,\sigma^2)| \geq |q|/2] + \exp(-d/8) \tag{C.37}$$

$$\leq 2e^{-\frac{q^2}{8\sigma^2}} + \exp(-d/8). \tag{C.38}$$

□

C.2 PROOF OF MAIN RESULTS

We use the results of the previous subsection to prove our main results in the linear setting.

First we prove our linear generalization result, Theorem 3.1, which we restate for the reader's convenience.

Theorem 3.1 (Extremal-Margin Generalization for Linear Problem)**.** *Let $\delta > 0$. There exist constants $\epsilon = \epsilon(\delta)$ and $c = c(\delta)$ such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega^{\text{linear}}_{\sigma,d}$ satisfying $\kappa^{\text{linear}}_{gen} + \delta \leq \kappa \leq \frac{1}{\delta}$, and $\frac{d}{n} \geq c$, then with probability $1 - 3e^{-n}$ over the randomness of a training set $S \sim \mathcal{D}^n$, for any $w \in \mathbb{R}^d$ that is a $(1-\epsilon)$-max-margin solution (as in Definition 2.5), we have $\mathcal{L}_{\mathcal{D}}(f_w) \leq e^{-\frac{n}{36d\sigma^4}} + e^{-n/8}$.*

We prove the following slightly stronger result, which implies Theorem 3.1, and gives the exact dependence of c on δ.

Theorem C.4. *There exists a universal constant c such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega^{\text{linear}}_{\sigma,d}$ such that $\kappa = \frac{n}{d\sigma^2} > \kappa^{\text{linear}}_{gen}$ and $\frac{d}{n} \geq c \max\left(\frac{1}{\kappa^2}, \kappa^2\right)$, with probability $1 - 3e^{-n}$ over the randomness of a training set $S \sim \mathcal{D}^n$, for any $w \in \mathbb{R}^d$ that is a $(1-\epsilon)$-max-margin solution (as in Definition 2.5), we have $\mathcal{L}_{\mathcal{D}}(f_w) \leq e^{-\frac{n}{36d\sigma^4}} + e^{-n/8}$, where $\epsilon = \frac{1}{c}\min\left(\frac{1}{\kappa}, \kappa\right)$.*

Proof. The proof follows from combining Lemmas C.2 and A.1. Let c be the universal constant from Lemmas C.2. For any $\kappa > \kappa_{\text{gen}}$, by Lemmas C.2, for constants $\epsilon = \frac{1}{c}\min\left(\kappa, \frac{1}{\kappa}\right)$, if $\frac{d}{n} > c^2 \max\left(\kappa^2, \frac{1}{\kappa^2}\right)$, with probability $1 - 3e^{-n}$, for any w which is a $(1-\epsilon)$ max-margin solution, we have

$$\frac{w^T \mu}{\|v\|_2 \kappa} \geq \frac{\sqrt{2}}{3}. \tag{C.39}$$

Now appealing to Lemma A.1, this means that $\mathcal{L}_{\mathcal{D}}(f_w) \leq 2e^{-\frac{\kappa}{36\sigma^2}} + \exp(-d/8) \leq 2e^{-\frac{\kappa}{36\sigma^2}} + \exp(-n/8)$. □

To prove our impossibility results, for any $\mathcal{D} = \mathcal{D}_{\mu,d,\sigma} \in \Omega$, we define the following mappings ψ and $\bar{\psi}$. For $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, where $x = \mu y + \xi$, define $\psi((x, y)) = (-\mu y + \xi, y)$. This mapping swaps the signal direction of the example, but maintains the label and junk component, and we will use it for our margin lower bound. Define $\bar{\psi}((x, y)) = (\mu y - \xi, y)$, which swaps the signal direction of the example, but maintains the label and junk component. We will use this for our UC lower bound.

For a set of training examples S, let $\psi(S)$ (resp. $\bar{\psi}(S)$) be the set where each example is mapped via ψ (resp. $\bar{\psi}$). Finally define $\psi(\mathcal{D}) := \mathcal{D}_{-\mu,d,\sigma}$. Thus $\psi(\mathcal{D})$ is the distribution with the opposite signal direction, and $\bar{\psi}(\mathcal{D}) = \mathcal{D}$. It is immediate to check that for any classifier $w \in \mathbb{R}^d$, $\mathcal{L}_{\mathcal{D}}(f_w) = 1 - \mathcal{L}_{\psi(\mathcal{D})}(f_w)$. Note that ψ and $\bar{\psi}$ implicitly depend on $\mathcal{D}$ through the parameter μ. If it is not clear from context that we are speaking about a specific $\mathcal{D}$, we will use $\psi_{\mathcal{D}}$ or $\bar{\psi}_{\mathcal{D}}$ to denote the mapping associated with $\mathcal{D}$.

We now prove the the linear part of Proposition 3.4. We restate a version which just includes the linear part, and gives more precise dependence of c and ϵ on the distance between κ and the boundaries κ_{uc} and κ_{gen}.

Proposition C.5 (UC Bounds are Vacuous for Linear Problem (From Proposition 3.4))**.** *There exists a universal constant c for which the following holds. For any n, d, σ and $\mathcal{D} \in \Omega^{\text{linear}}_{\sigma,d}$ such that $\kappa^{\text{linear}}_{gen} \leq \kappa \leq \kappa^{\text{linear}}_{uc}$, $\frac{d}{n} \geq \frac{c}{\kappa^2(\kappa_{uc}-\kappa)^4}$, the following holds for any $\epsilon \leq \frac{\kappa(\kappa_{uc}-\kappa)^2}{c}$. Let $\mathcal{A}$ be any algorithm that outputs $w \in \mathbb{R}^d$ which is a $(1-\epsilon)$-max-margin solution for any $S \in (\mathbb{R}^d \times \{1, -1\})^n$. Let $\mathcal{H}$ be any hypothesis class that is useful for $\mathcal{A}$ on $\mathcal{D}$ (as in Definition 2.2). Suppose that ϵ_{unif} is a uniform convergence bound for $\mathcal{D}$ and $\mathcal{H}$, that is,*

$$\Pr_{S\sim\mathcal{D}^n}[\sup_{h\in\mathcal{H}} |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| \geq \epsilon_{\text{unif}}] \leq 1/4.$$

Then $\epsilon_{\text{unif}} \geq 1 - e^{-\frac{n}{36d\sigma^2}} - e^{-n/8}$.

Proof. Let $T_{\mathcal{D}} \subset 2^{(\mathbb{R}^d \times \{-1,1\})^n}$ be the set of training sets S on which the conclusion of Lemma C.2 holds for S. Thus $\Pr_{S\sim\mathcal{D}^n}[S \in T_{\mathcal{D}}] \geq 1 - 3e^{-n}$. Let $H \subset 2^{(\mathbb{R}^d \times \{-1,1\})^n}$ be the set of training sets S on which $\mathcal{A}(S) \in \mathcal{H}$. Thus $\Pr_{S\sim\mathcal{D}^n}[S \in H] \geq \frac{3}{4}$.

Let $T'_{\mathcal{D}}$ be the set on which

$$|\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_{\phi(S)}(h)| \leq \epsilon_{\text{unif}} \qquad \forall h \in \mathcal{H}, \tag{C.40}$$

where $\phi := \bar{\psi}_{\mathcal{D}}$. By assumption, $\Pr_{S\sim\mathcal{D}^n}[\phi(S) \in T'_{\mathcal{D}}] = \Pr_{S\sim\mathcal{D}^n}[S \in T'_{\mathcal{D}}] \geq \frac{3}{4}$. By a union bound, for $n \geq 2$,

$$\Pr_{S\sim\mathcal{D}^n}[S \in T'_{\mathcal{D}} \wedge S \in T_{\mathcal{D}} \wedge S \in H] \geq 1 - \left(1 - \frac{3}{4}\right) - \left(1 - \frac{3}{4} + 3e^{-n}\right) = \frac{1}{2} - 3e^{-n} > 0. \tag{C.41}$$

This is because the distribution of $\bar{\psi}(S)$ with $S \sim \mathcal{D}^n$ is the same as the distribution of n samples from $\bar{\psi}(\mathcal{D}) = \mathcal{D}$.

Let S be any set for which the three event above hold, ie.,

$$S \in T'_{\mathcal{D}} \wedge S \in T_{\phi(\mathcal{D})} \wedge S \in H. \tag{C.42}$$

With $f_w = \mathcal{A}(S)$, we have by combining the results of Lemmas C.2 and A.1 that $\mathcal{L}_{\mathcal{D}}(f_w) \leq e^{-\frac{n}{36d\sigma^2}} - e^{-n/8}$. Further, by the second conclusion of Lemma C.2, we know that $\mathcal{L}_{\phi(S)}(f_w) = 1$, since f_w misclassifies every point in S. It follows that $\epsilon_{\text{unif}} \geq 1 - e^{-\frac{n}{36d\sigma^2}} - e^{-n/8}$. □

Finally we prove Proposition 3.5 via a similar technique, but using the mapping ψ instead of $\bar{\psi}$.

Proposition 3.5 (Polynomial Margin Bounds Fail for Linear Problem)**.** *Fix $\delta > 0$. For any n, d, σ and $\mathcal{D} \in \Omega^{\text{linear}}_{\sigma,d}$ such that $\kappa^{\text{linear}}_{gen} + \delta < \kappa < \kappa^{\text{linear}}_{uc} - \delta$ and $\frac{d}{n} \geq c$, the following holds. Let $\mathcal{A}$ be any algorithm so that $\mathcal{A}(S)$ outputs a $(1-\epsilon)$-max-margin solution f_w for any $S \in (\mathbb{R}^d \times \{1,-1\})^n$. Let $\mathcal{H}$ be any hypothesis class that is useful for $\mathcal{A}$ (as in Definition 2.2) on both $\mathcal{D}^{\text{linear}}_{\mu,\sigma}$ and $\mathcal{D}^{\text{linear}}_{-\mu,\sigma}$. Suppose that there exists an polynomial margin bound of integer degree p: that is, there is some G that satisfies for $\tilde{D} \in \{\mathcal{D}, \psi(\mathcal{D})\}$,*

$$\Pr_{S\sim\tilde{\mathcal{D}}^n}\left[\sup_{h\in\mathcal{H}} \mathcal{L}_{\tilde{\mathcal{D}}}(h) - \mathcal{L}_S(h) \geq \frac{G}{\gamma(h,S)^p}\right] \leq \frac{1}{4}.$$

Then with probability $\frac{1}{2} - 3e^{-n}$ over $S \sim \mathcal{D}^n$, the margin bound is weak even on the max-margin solution, that is, $\frac{G}{\gamma^*(S)^p} \geq \max\left(\frac{1}{c}, 1 - e^{-\frac{\kappa}{36\sigma^2}} - e^{-n/8} - \frac{3\kappa}{c}\right)^p$, which is more than an absolute constant.

We prove the following slightly stronger result, which implies Proposition 3.5, and gives the conditions depending on the distance between κ and the boundaries κ_{uc} and κ_{gen}.

Proposition C.6. *There exists a universal constant c such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega^{\text{linear}}_{\sigma,d}$ such that $\kappa < \kappa^{\text{linear}}_{uc}$ and $\frac{d}{n} \geq \frac{c}{\kappa^2(\kappa_{uc}-\kappa)^4}$, the following holds. Let $\epsilon = \frac{\kappa(\kappa_{uc}-\kappa)^2}{c}$, and let $\mathcal{A}$ be any algorithm so that $\mathcal{A}(S)$ outputs a $(1-\epsilon)$-max-margin solution f_w for any $S \in (\mathbb{R}^d \times \{1,-1\})^n$. Let $\mathcal{H}$ be any hypothesis class that is useful for $\mathcal{A}$ (as in Definition 2.2) on both $\mathcal{D}^{\text{linear}}_{\mu,\sigma}$ and $\mathcal{D}^{\text{linear}}_{-\mu,\sigma}$. Suppose that there exists an polynomial margin bound of integer degree p: that is, there is some G that satisfies for $\tilde{D} \in \{\mathcal{D}, \psi(\mathcal{D})\}$,*

$$\Pr_{S\sim\tilde{\mathcal{D}}^n}\left[\sup_{h\in\mathcal{H}} \mathcal{L}_{\tilde{\mathcal{D}}}(h) - \mathcal{L}_S(h) \geq \frac{G}{\gamma(h,S)^p}\right] \leq \frac{1}{4}.$$

Then with probability $\frac{1}{2} - 3e^{-n}$ over $S \sim \mathcal{D}^n$, the margin bound is weak even on the max-margin solution, that is, $\frac{G}{\gamma^*(S)^p} \geq \max\left(\frac{1}{c}, 1 - e^{-\frac{\kappa}{36\sigma^2}} - e^{-n/8} - \frac{3\kappa}{c}\right)^p$, which is more than an absolute constant.

Proof. For any $\mathcal{D} \in \Omega$, let $T_{\mathcal{D}} \subset 2^{(\mathbb{R}^d\times\{-1,1\})^n}$ be the set of training sets S on which the conclusion of Lemma C.2 holds for $\mathcal{D}$ and S. Thus for any $\mathcal{D} \in \Omega$, $\Pr_{S\sim\mathcal{D}^n}[S \in T_{\mathcal{D}}] \geq 1 - 3e^{-n}$. Let $H \subset 2^{(\mathbb{R}^d\times\{-1,1\})^n}$ be the set of training sets S on which $\mathcal{A}(S) \in \mathcal{H}$. Thus for any $\mathcal{D} \in \Omega$, $\Pr_{S\sim\mathcal{D}^n}[S \in H] \geq \frac{3}{4}$.

For any $\mathcal{D} \in \Omega$, let $T'_{\mathcal{D}}$ be the set on which

$$\mathcal{L}_{\mathcal{D}}(h) \leq \mathcal{L}_S(h) + \frac{G}{\gamma(h,S)^p} \qquad \forall h \in \mathcal{H}. \tag{C.43}$$

By assumption, for any $\mathcal{D} \in \Omega$, $\Pr_{S\sim\mathcal{D}^n}[S \in T'_{\mathcal{D}}] \geq \frac{3}{4}$.

Now fix any $\mathcal{D} = \mathcal{D}_{\mu,\sigma,d} \in \Omega$. By a union bound, with $\psi = \psi_{\mathcal{D}}$,

$$\Pr_{S\sim\mathcal{D}^n}[S \in T'_{\mathcal{D}} \wedge \psi(S) \in T_{\psi(\mathcal{D})} \wedge \psi(S) \in H] \geq 1 - \left(1 - \frac{3}{4}\right) - \left(1 - \frac{3}{4} + 3e^{-n}\right) = \frac{1}{2} - 3e^{-n}. \tag{C.44}$$

This is because the distribution of $\psi(S)$ with $S \sim \mathcal{D}^n$ is the same as the distribution of n samples from $\psi(\mathcal{D})$.

Let S be any set for which the three events above hold, ie.,

$$S \in T'_{\mathcal{D}} \wedge \psi(S) \in T_{\psi(\mathcal{D})} \wedge \psi(S) \in H. \tag{C.45}$$

With $f_w = \mathcal{A}(\psi(S))$, we have by combining the results of Lemmas C.2 and A.1 that $\mathcal{L}_{\psi(\mathcal{D})}(f_w) \leq e^{-\frac{n}{9d\sigma^2}}$, and thus $\mathcal{L}_{\mathcal{D}}(f_w) \geq 1 - e^{-\frac{n}{9d\sigma^2}}$. Further, by the conclusion of Lemma C.2, we know that for $(x_j, y_j) \in S$, with μ being the direction of the distribution $\psi(\mathcal{D})$, for some $C > 0$,

$$y_j v^T \xi_j \geq \left(1 + \frac{1}{C}, \frac{1}{2\kappa}\right) w^T \mu. \tag{C.46}$$

Observe that $\gamma^*(\psi(S)) \leq \frac{1}{1-\epsilon}(w^T\mu + y_j v^T \xi_j)$, and thus letting $b = w^T\mu$ and $a = \min_j y_j v^T \xi_j$, we have

$$\frac{\gamma(f_w, S)}{\gamma^*(\psi(S))} \geq (1-\epsilon)\frac{a-b}{a+b} > (1-\epsilon)\frac{\max\left(1+\frac{1}{C}, \frac{1}{2\kappa}\right) - 1}{\max\left(1+\frac{1}{C}, \frac{1}{2\kappa}\right) + 1} \geq (1-\epsilon)\max\left(\frac{1}{3C}, 1-4\kappa\right) \geq \max\left(\frac{1}{4C}, 1-2\epsilon-8\kappa\right), \tag{C.47}$$

since C is a constant and ϵ is sufficiently small.

It follows that for any such S, we must have

$$G \geq (1 - e^{-\frac{n}{9d\sigma^4}})\gamma(f_w, S)^p \geq (1 - e^{-\frac{n}{9d\sigma^4}})\max\left(\frac{1}{4C}, 1-2\epsilon-8\kappa\right)^p \gamma^*(\psi(S))^p. \tag{C.48}$$

Thus for the distribution $\psi(\mathcal{D})$, with probability at least $\frac{1}{2} - 3e^{-n}$, the margin bound yields a generalization guarantee no better than

$$(1 - e^{-\frac{n}{9d\sigma^4}})\max\left(\frac{1}{4C}, 1-2\epsilon-8\kappa\right)^p \geq \max\left(\frac{1}{c}, 1 - e^{-\frac{\kappa}{9\sigma^2}} - 9\kappa\right)^p, \tag{C.49}$$

where we have assumed c is a sufficiently large constant, and plugged in the assumption that $\epsilon \leq \frac{\kappa}{c}$.

□

D Proofs for XOR 2-layer Neural Network Problem

Throughout this section, since we are only concerned with the XOR problem, we will abbreviate $\Omega = \Omega_{\sigma,d}^{h,\mathrm{XOR}}$, $\kappa_{\text{uc}} = \kappa_{\text{uc}}^{\mathrm{XOR},h}$, and $\kappa_{\text{gen}} = \kappa_{\text{gen}}^{\mathrm{XOR},h}$.

In subsection D.1, we present a series of technical lemmas. In subsection D.3, we prove our main theorems for the XOR problem, assuming the technical lemmas. In subsection D.4, we prove the technical lemmas.

D.1 Technical Lemmas

Notation. Throughout the following subsection, we assume $\mathcal{D} = \mathcal{D}_{\mu_1,\mu_2,\sigma,d} \in \Omega$ is fixed. For a weight matrix $W \in \mathbb{R}^{m\times d}$, we define $U = W\Pi_{\mathrm{span}(\mu_1,\mu_2)}$ and $V = W\Pi_{\mathrm{span}(\mu_1,\mu_2)^\perp}$, where Π_T is the orthogonal projector onto T. For $i \in [m]$, let $w_i \in \mathbb{R}^d$, $u_i \in \mathbb{R}^d$ and $v_i \in \mathbb{R}^d$ denote the rows of W, U and V respectively. We use $\mathbb{E}$ to denote the expectation over i uniformly in $[m]$. Let $H_+ : \{i : a_i > 0\}$, and $H_- : \{i : a_i < 0\}$

Recall that our samples $x_1, \ldots, x_n$ are of the form $x_j = z_j + \xi_j$, where $z_j \in \{\mu_1, -\mu_1, \mu_2, -\mu_2\}$ and $\xi_j \perp \mathrm{span}(\mu_1, \mu_2)$. Let $\mathcal{P}_1, \mathcal{P}_{-1}, \mathcal{N}_1$ and $\mathcal{N}_{-1}$ denote the four clusters of points, that is,

$$\mathcal{P}_1 = \{j \in [n] : z_j = \mu_1\} \tag{D.1}$$
$$\mathcal{P}_{-1} = \{j \in [n] : z_j = -\mu_1\} \tag{D.2}$$
$$\mathcal{N}_1 = \{j \in [n] : z_j = \mu_2\} \tag{D.3}$$
$$\mathcal{N}_{-1} = \{j \in [n] : z_j = -\mu_2\} \tag{D.4}$$

Let $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_{-1}$, and let $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_{-1}$. Let $\Xi \in \mathbb{R}^{d\times n}$ be the matrix with jth column ξ_j. Let $n_{\min} := \min\left(|\mathcal{N}_1|, |\mathcal{N}_{-1}|, |\mathcal{P}_1|, |\mathcal{P}_{-1}|\right)$, and $n_{\max} := \max\left(|\mathcal{N}_1|, |\mathcal{N}_{-1}|, |\mathcal{P}_1|, |\mathcal{P}_{-1}|\right)$ such that we expect $n_{\min}$ and $n_{\max}$ to be close to $\frac{n}{4}$, as per the Lemma D.1 below.

Assume throughout the following section that $h \in (1, 2)$ is fixed, and recall that we have defined the activation $\phi(z) = \max(0, z)^h$.

Lemma D.1. *For any $\beta > 0$, with probability as least $1 - 8e^{-8n\beta^2}$ over $S \in \mathcal{D}^n$, for all clusters $C \in \mathcal{P}_1, \mathcal{P}_{-1}, \mathcal{N}_1, \mathcal{N}_{-1}$, we have*

$$\left||C| - \frac{n}{4}\right| \leq \beta n, \tag{D.5}$$

and thus for $\beta \leq \frac{1}{8}$,

$$\frac{n_{\max}}{n_{\min}} = \frac{\max\left(|\mathcal{P}_1|, |\mathcal{P}_{-1}|, |\mathcal{N}_1|, |\mathcal{N}_{-1}|\right)}{\min\left(|\mathcal{P}_1|, |\mathcal{P}_{-1}|, |\mathcal{N}_1|, |\mathcal{N}_{-1}|\right)} \leq 1 + 16\beta. \tag{D.6}$$

This lemma follows immediately from Hoeffding's inequality on Bernoulli random variables: To prove it, one can apply Hoeffding's inequality four times (once for each cluster), and take a union bound.

D.1.1 OVERVIEW OF TECHNICAL LEMMAS.

Our goal will be to analyze near-optimal solutions to the following optimization program which is defined using the n training examples.

Definition D.2 (Opt 1)**.**
Parameter: P_1
Variables: $W \in \mathbb{R}^{m \times d}$

$$\max \gamma \tag{D.7}$$
$$\mathbb{E}_i a_i \phi(w_i^T x_j) y_j \geq \gamma \qquad \forall j \tag{D.8}$$
$$\mathbb{E}_i \|w_i\|^2 \leq P_1 \tag{D.9}$$

In this subsection, we define a chain of optimization programs beginning from Opt 1. Each subsequent optimization problem becomes simpler and involves fewer variables. The chaining lemmas in this section typically show two conclusions:

1. If a solution is near-optimal for the ith optimization program in the chain, then that solution can be transformed into a [series of] solutions that are [mostly] near-optimal for the $(i+1)$th optimization program in the chain.

2. An optimal solution to the ith optimization program in the chain can be converted into a near-optimal solution to the $(i-1)$th optimization program in the chain.

Ultimately, in Lemma D.15 we study the optimal solution to the final simplest optimization program, which only includes 3 variables. From this, using the first conclusion of the lemmas, we are able to chain back through the optimization programs to analyze certain properties of any W which is a near-max-margin solution. This analysis ultimately leads to Lemmas D.17 and Lemma D.18, which are our main tools in proving generalization and the impossibility of UC bounds.

The second conclusion of the lemmas allows us to construct a near-max-margin solution $\hat{W}$ which satisfies certain properties allowing us to show the the limitations of margin bounds, and the failure of generalization for near-max-margin solutions when $\kappa < \kappa_{\text{gen}}$. This is captured in Lemmas D.25 and D.19.

Chain of Optimization Programs. We define the chain of Optimization Programs. Unless otherwise specified, all variables and parameters lie in $\mathbb{R}$. Like Opt 1, these programs all assume that the set of training examples $S = \{(x_j, y_j)\}_{j \in [n]}$ is fixed.

Definition D.3 (Opt 2)**.**
Parameter: P_2
Variables: $\{c_{ij}\}_{i \in [m], j \in [n]}$, $\{s_i\}_{i \in H_+}$, $\{t_i\}_{i \in H_-}$

$$\max \gamma \tag{D.10}$$
$$\frac{1}{2}\mathbb{E}_{i \in H_+} \phi(s_i + c_{ij}) \geq \gamma \qquad \forall j \in \mathcal{P}_1 \tag{D.11}$$
$$\frac{1}{2}\mathbb{E}_{i \in H_+} \phi(-s_i + c_{ij}) \geq \gamma \qquad \forall j \in \mathcal{P}_{-1} \tag{D.12}$$
$$\frac{1}{2}\mathbb{E}_{i \in H_-} \phi(t_i + c_{ij}) \geq \gamma \qquad \forall j \in \mathcal{N}_1 \tag{D.13}$$
$$\frac{1}{2}\mathbb{E}_{i \in H_-} \phi(-t_i + c_{ij}) \geq \gamma \qquad \forall j \in \mathcal{N}_{-1} \tag{D.14}$$
$$\frac{1}{2}\mathbb{E}_{i \in H_+}\left[s_i^2 + \frac{1}{d\sigma^2}\sum_{j \in \mathcal{P}} c_{ij}^2\right] + \frac{1}{2}\mathbb{E}_{i \in H_-}\left[t_i^2 + \frac{1}{d\sigma^2}\sum_{j \in \mathcal{N}} c_{ij}^2\right] \leq P_2 \tag{D.15}$$

Definition D.4 (Opt 3)**.**
Parameter: P_3

Variables: $\{c_{ij}\}_{i\in H, j\in S_1\cup S_{-1}}$, $\{b_i\}_{i\in H}$

$$\max\ \gamma \tag{D.16}$$

$$\frac{1}{2}\mathbb{E}_{i\in H}\phi(b_i + c_{ij}) \geq \gamma \qquad \forall j \in S_1 \tag{D.17}$$

$$\frac{1}{2}\mathbb{E}_{i\in H}\phi(-b_i + c_{ij}) \geq \gamma \qquad \forall j \in S_{-1} \tag{D.18}$$

$$\mathbb{E}_{i\in H}\left(b_i^2 + \frac{1}{d\sigma^2}\sum_{j\in S_1\cup S_{-1}} c_{ij}^2\right) \leq P_3, \tag{D.19}$$

where sets $S_1, S_{-1} \subset [n]$ *with* $|S_1| = |S_{-1}| = n_{\min}$, *and* $H \subset [m]$ *with* $|H| = \frac{m}{2}$.

Definition D.5 (Opt 4)**.**
Parameter: P_4
Variables: $\{c_i\}_{i\in H}$, $\{d_i\}_{i\in H}$, $\{b_i\}_{i\in H}$

$$\max\ \gamma \tag{D.20}$$

$$\frac{1}{2}\mathbb{E}_{i\in H}\phi(b_i + c_i) \geq \gamma \tag{D.21}$$

$$\frac{1}{2}\mathbb{E}_{i\in H}\phi(-b_i + d_i) \geq \gamma \tag{D.22}$$

$$\mathbb{E}_{i\in H}\left(b_i^2 + \frac{n_{\min}}{d\sigma^2}(c_i^2 + d_i^2)\right) \leq P_4, \tag{D.23}$$

where $H \subset [m]$ *with* $|H| = \frac{m}{2}$.

Definition D.6 (Opt 5: Trivariate Simplification)**.**
Parameter: P_5
Variables: c, d, b

$$\max \frac{1}{4}\left(\phi(b + c) + \phi(-b + d)\right) \tag{D.24}$$

$$b^2 + \frac{\hat{\kappa}}{4}(c^2 + d^2) \leq P_5, \tag{D.25}$$

where $\hat{\kappa} = \frac{4n_{\min}}{d\sigma^2}$.

For $i \in \{1, 2, 3, 4, 5\}$, let D_i denote the domain of parameters and variables in the program Opt i. For an instance of program Opt i in D_i, we say it is $(1-\epsilon)$-optimal if the objective value given by the variables is at least $(1-\epsilon)$ times the maximum objective value for the parameters in the instance. We use $\mathcal{I}_i \in D_i$ to denote an instance of the program Opt i. When such an instance is fixed, we will freely use the names of the parameters and the variables associated with Opt i to refer to the variables and parameters of $\mathcal{I}_i$. For instance, if $\mathcal{I}_1 \in D_1$, then W is the variable associated with $\mathcal{I}_1$.

D.1.2 Chaining Lemmas.

Throughout the following section, we globally assume that $\frac{d}{n} \geq 4C_{B.1}^2$ and $\epsilon \leq \frac{1}{4}$, such that on the condition that Lemma B.1 holds, we have $C_{B.1}\sqrt{\frac{n}{d}} \leq \frac{1}{2}$. Such conditions are assumed in the theorems that follow from these lemmas, so there is no harm in making the assumption now. All of the lemmas in this section are proved in Section D.4.2.

Lemma D.7 (Opt 1 $\leftrightarrow$ Opt 2)**.** *Assume the conclusion of Lemma B.1 holds for* Ξ.

Define the mapping $\psi_{12} : D_1 \to D_2$ *as follows. Given input* $\mathcal{I}_1$ *with variable* $W = U + V$, *output:*

- $P_2 = P_1\left(1 + C_{B.1}\sqrt{\frac{n}{d}}\right)$
- $c_{ij} = v_i^T\xi_j$ *for all* $i \in [m]$, $j \in [n]$
- $s_i = \mu_1^T w_i$ *for all* $i \in H_+$
- $t_i = \mu_2^T w_i$ *for all* $i \in H_-$

Define the mapping $\psi_{21} : D_1 \to D_2$ *as follows: Given input* $\mathcal{I}_2$, *output* $\mathcal{I}_1$ *as follows:*

- $P_1 = \frac{P_2}{1 - C_{B.1}\sqrt{\frac{n}{d}}}$.

- *For all $i \in H_+$, let $u_i = s_i\mu_1$, and choose v_i to be the min-norm vector such that $v_i^T\xi_j = c_{ij}$ for all $j \in \mathcal{P}$ and $v_i^T\xi_j = 0$ for all $j \in \mathcal{N}$.*
- *For all $i \in H_-$, let $u_i = t_i\mu_2$, and choose v_i to be the min-norm vector such that $v_i^T\xi_j = c_{ij}$ for all $j \in \mathcal{N}$ and $v_i^T\xi_j = 0$ for all $j \in \mathcal{P}$.*

Then with $\epsilon' = \sqrt{1-(1-\epsilon)\left(1-C_{B.1}\sqrt{\frac{n}{d}}\right)^h}$,

1. *If $\mathcal{I}_1 \in D_1$ is $(1-\epsilon)$-optimal, then $\psi_{12}(\mathcal{I}_1)$ is $1-\epsilon'$-optimal on Opt 1 and has objective value at most $\frac{1}{1-\epsilon'}$ larger than the objective of $\mathcal{I}_1$.*
2. *If $\mathcal{I}_2 \in D_2$ is $(1-\epsilon)$-optimal, then $\psi_{21}(\mathcal{I}_2)$ is $1-\epsilon'$-optimal on Opt 1.*

The following lemma states that in a near-optimal solution to Opt 1, most of the contribution to the norm constraint comes from the the variables that get used in the mapping ψ_{12} to Opt 2.

Lemma D.8. *Assume the conclusion of Lemma B.1 holds for Ξ. Then any solution $W = U + V$ to Opt 1 with $\|W\| = 1$ that is $(1-\epsilon)$-optimal must satisfy:*

1. $\frac{1}{2}\mathbb{E}_{i\in H_+}\left[\|\mu_2^T w_i\|^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{N}}(v_i^T\xi_j)^2\right] + \frac{1}{2}\mathbb{E}_{i\in H_-}\left[\|\mu_1^T w_i\|^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}}(v_i^T\xi_j)^2\right] < (\epsilon'_{D.8})^2$
2. *For at least a $1-\epsilon'_{D.8}$ fraction of the data points j, we have $\frac{1}{2}\mathbb{E}_{i:\mathrm{sign}(a_i)=-y_j}\left[(v_i^T\xi_j)^2\right] \le \frac{1}{\kappa}\cdot\epsilon'_{D.8}$.*

where $\epsilon'_{D.8} = \epsilon'_{D.8}(\epsilon) := \sqrt{2C_{B.1}\sqrt{\frac{n}{d}} + 2\epsilon}$.

Lemma D.9 (Opt 2 $\leftrightarrow$ Opt 3)**.** *Define the mapping $\psi_{23} : D_2 \to D_3 \times D_3$ as follows. Given input $\mathcal{I}_2$, output $\mathcal{I}_3^{(1)}$, $\mathcal{I}_3^{(2)}$, where for $\mathcal{I}_3^{(1)}$:*

- $H := H_+$
- $P_3 := \frac{1}{2}\mathbb{E}_{i\in H_+}\left(s_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}} c_{ij}^2\right)$
- *Let S_1 be an arbitrary set of $n_{\min}$ elements of $\mathcal{P}_1$, and let S_{-1} be an arbitrary set of $n_{\min}$ elements of $\mathcal{P}_{-1}$. Define c_{ij} to be the same as in $\mathcal{I}_2$ for all $j \in S_1 \cup S_{-1}$, $i \in H_+$.*
- $b_i = s_i$ *for $i \in H_+$,*

and for $\mathcal{I}_3^{(2)}$:

- $H := H_-$
- $P_3 := \mathbb{E}_{i\in H_-}\left(t_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{N}} c_{ij}^2\right)$
- *Let S_1 be an arbitrary set of $n_{\min}$ elements of $\mathcal{N}_1$, and let S_{-1} be an arbitrary set of $n_{\min}$ elements of $\mathcal{N}_{-1}$. Define c_{ij} to be the same as in $\mathcal{I}_2$ for all $j \in S_1 \cup S_{-1}$, $i \in H_-$.*
- $b_i = t_i$ *for $i \in H_-$,*

Define the mapping $\psi_{32} : D_3 \to D_2$ as follows. Given input $\mathcal{I}_3$, output $\mathcal{I}_2$, where

- $P_2 := P_3\left(\frac{n_{\max}}{n_{\min}}\right)$
- *Define an arbitrary bisections π_+ and π_- from H_+ and H_- respectively to H. Similarly define surjections $\rho_{+,1} : \mathcal{P}_1 \to S_1$, $\rho_{-,1} : \mathcal{P}_{-1} \to S_{-1}$, $\rho_{+,1} : \mathcal{N}_1 \to S_1$, $\rho_{-,1} : \mathcal{N}_{-1} \to S_{-1}$, such that for $x \in \{1,-1\}$, $\mathbb{E}_{j\in\mathcal{P}_x}\mathbb{E}_{i\in H} c^2_{i\rho_{+,x}(j)} \le \mathbb{E}_{j\in S_x}\mathbb{E}_{i\in H} c_{ij}^2$, and similarly, $\mathbb{E}_{j\in\mathcal{N}_x}\mathbb{E}_{i\in H} c^2_{i\rho_{-,x}(j)} \le \mathbb{E}_{j\in S_x}\mathbb{E}_{i\in H} c_{ij}^2$. For $i \in H_+$, define $s_i := b_{\pi_+(i)}$ and $c_{ij} := c_{\pi_+(i)\rho_{+,x}(j)}$ for all $x \in \{1,-1\}$ and $j \in \mathcal{P}_x$. For $i \in H_-$, define $t_i := b_{\pi_-(i)}$, and $c_{ij} := c_{\pi_-(i)\rho_{+,x}(j)}$ for all $x \in \{1,-1\}$ and $j \in \mathcal{N}_x$. Note that we here the c_{ij} variables we are* defining *come belong to $\mathcal{I}_2$, and they are define in terms of the c_{ij} variables from $\mathcal{I}_3$.*

Then with $\epsilon' = \sqrt{1-(1-\epsilon)\left(\frac{n_{\max}}{n_{\min}}\right)^{-h}}$

1. *If $\mathcal{I}_2 \in D_2$ is $(1-\epsilon)$-optimal, then each instance of $\psi_{23}(\mathcal{I}_2)$ is $(1-\epsilon')$-optimal on Opt 3, and has objective at most $\frac{1}{1-\epsilon'}$ times the objective of $\mathcal{I}_2$.*

2. *If $\mathcal{I}_3 \in D_3$ is $(1-\epsilon)$-optimal, then $\psi_{32}(\mathcal{I}_3)$ is $(1-\epsilon')$-optimal on Opt 2.*

Lemma D.10 (Opt 3 $\leftrightarrow$ Opt 4)**.** *Define the mapping $\psi_{34} : D_3 \to D_4^{n_{\min}}$ as follows. Arbitrarily choose a list of $n_{\min}$ pairs $p = (j, j') \in S_1 \times S_{-1}$, such that each $j \in S_1$ appears in one pair, and each $j' \in S_{-1}$ appears in one pair. Given input $\mathcal{I}_3$, output $\mathcal{I}_3^{(p)}$ for each pair p as follows:*

- *$P_4 := \mathbb{E}_{i \in H}\left(b_i^2 + \frac{n_{\min}}{d\sigma^2}(c_{ij}^2 + c_{ij'}^2)\right)$*
- *Keep H and all the b_i the same as in $\mathcal{I}_3$.*
- *Put $c_i := c_{ij}$ and $d_i := c_{ij'}$ for all $i \in H$.*

In reverse, define the mapping $\psi_{43} : D_4 \to D_3$ as follows:

- *Put $P_3 := P_4$.*
- *Keep H and all the b_i the same as in $\mathcal{I}_3$.*
- *For all $i \in H$, put $c_{ij} := c_i$ for all $j \in S_1$ and $c_{ij} := d_i$ for all $j \in S_{-1}$.*

Then:

1. *If $\mathcal{I}_3 \in D_3$ is $(1-\epsilon)$-optimal, then on at least a $1-\sqrt{\epsilon}$ fraction of the $\frac{n}{4}$ instances of $\psi_{34}(\mathcal{I}_4)$, the instance is $(1-\sqrt{\epsilon})$-optimal on Opt 4 and has objective value at most $\frac{1}{1-\sqrt{\epsilon}}$ larger than the objective of $\mathcal{I}_3$.*

2. *If $\mathcal{I}_4 \in D_4$ is $(1-\epsilon)$-optimal, then $\psi_{43}(\mathcal{I}_4)$ is $(1-\sqrt{\epsilon})$-optimal on Opt 3.*

Lemma D.11 (Opt 5 $\leftrightarrow$ Opt 4)**.** *define the mapping $\psi_{45} : D_4 \to D_5^{\frac{m}{2}}$ as follows. Given input $\mathcal{I}_4$, output $\frac{m}{2}$ instances $\mathcal{I}_5^{(i)}$ for $i \in H$ with*

- *$P_5^{(i)} := b_i^2 + \frac{\hat{\kappa}}{4}(c_i^2 + d_i^2)$.*
- *$(b, c, d)^{(i)} = (b_i, c_i, d_i)$.*

Define the mapping $\psi_{54} : D_5 \to D_4$ as follows. Given input $\mathcal{I}_5$, output:

- *$P_4 := P_5$*
- *For half of the $i \in H$, put $(b_i, c_i, d_i) = (b, c, d)$.*
- *For the other half of the $i \in H$, put $(b_i, c_i, d_i) = (-b, d, c)$.*

Then:

1. *If $\mathcal{I}_4 \in D_4$ is $(1-\epsilon)$-optimal, then the average objective value of the $\frac{m}{2}$ instances of $\psi_{45}(\mathcal{I}_4)$ is at most $\frac{1}{1-\epsilon}$ times larger than the objective of $\mathcal{I}_4$.*

2. *If $\mathcal{I}_5 \in D_5$ is $(1-\epsilon)$-optimal, then $\psi_{54}(\mathcal{I}_5)$ is $(1-\sqrt{\epsilon})$-optimal on Opt 4.*

Our main tool in proving these chaining lemmas is the following analysis lemma. We first state a definition.

Definition D.12. *An optimization program is q-homogeneous with with respect to a parameter P if the optimal objective equals CP^q, for some fixed value C.*

Lemma D.13. *Consider two optimization programs Opt A and Opt B with domains D_A and D_A with are q-homogeneous with respect to parameters P_A and P_B respectively. Suppose for some positive integer k, we have a mapping $\psi_{AB} : D_A \to D_B^k$ and $\psi_{BA} : D_B \to D_A$. Suppose for any feasible instances $\mathcal{I}_A \in D_A$ and $\mathcal{I}_B \in D_B$:*

1. *All k instances of $\psi_{AB}(\mathcal{I}_A)$ are feasible, and have at least the same objective value as $\mathcal{I}_A$. Similarly $\psi_{BA}(\mathcal{I}_B)$ is feasible and has at least the same objective value as $\mathcal{I}_B$.*

2. *The average parameter P_B of the k instances of $\psi_{AB}(\mathcal{I}_A)$ it at most $(1+\delta)$ times the parameter P_A of $\mathcal{I}_A$.*

3. $P_A(\psi_{BA}(\mathcal{I}_B)) \le (1+\delta)P_B(\mathcal{I}_B)$. *(Here the notation $P(\mathcal{I})$ refers to the parameter P in an instance $\mathcal{I}$.)*

Then letting $\epsilon' = \sqrt{1-(1-\epsilon)(1+\delta)^{-2q}}$,

1. *If $\mathcal{I}_A$ is $(1-\epsilon)$-optimal, then for at least a $1-\epsilon'$ fraction of the k instances $\psi_{AB}(\mathcal{I}_B)$ are $(1-\epsilon')$-optimal and have objective value at most $\frac{1}{1-\epsilon'}$ times the objective of $\mathcal{I}_A$.*
2. *If $\mathcal{I}_B$ is $(1-\epsilon)$-optimal, then $\psi_{BA}(\mathcal{I}_B)$ is $(1-\epsilon')$-optimal.*

In particular, if $\delta = 0$, $\epsilon' = \sqrt{\epsilon}$.

D.1.3 Analysis of Opt 5: Trivariate Program.

The lemmas in this section are proved in Section D.4.3.

Define $\gamma_0(\kappa) := 2\phi\left(\sqrt{\frac{2}{\kappa}}\right)$, and let $\gamma_*(\kappa) := \phi\left(\sqrt{\frac{\kappa}{4+\kappa}} + \sqrt{\frac{16}{\kappa(4+\kappa)}}\right)$. It is straightforward to check that $\gamma_0(\hat{\kappa})$ is 4 times the optimum of Opt 5 when $P_5 = 1$, and we impose the additional constraint that $b = 0$. Similarly, $\gamma_*(\hat{\kappa})$ is 4 times the optimum of Opt 5 when we impose the additional constraint that $d = 0$.

Recall that we have defined κ_{gen} to be the threshold at which $\gamma_*(\kappa_{\text{gen}}) = \gamma_0(\kappa_{\text{gen}})$, and κ_{uc} to be the threshold in κ at which $\sqrt{\frac{\kappa}{4+\kappa}} = \sqrt{\frac{16}{\kappa(4+\kappa)}}$. Observe that $\kappa_{\text{uc}} = 4$.

The following lemma yields the optimal solution to the program Opt 5.

Lemma D.14. *Let $k := \frac{\hat{\kappa}}{4}$ and assume $P_5 = 1$. If $k > \kappa_{gen}/4$, then if we impose the additional constraint that $b \ge 0$, the supremum of Opt 5 (in Definition D.6) and it is achieved uniquely at the point where $d = 0$, $b = \sqrt{\frac{k}{1+k}}$, and $c = \sqrt{\frac{1}{k(1+k)}}$. Outside of any neighborhood of this point, the supremum is bounded away from from the supremum of Opt 5.*

If $k < \kappa_{gen}/4$, then supremum is achieved by some optimal point with $b = 0$.

Lemma D.15. *There exists strictly positive constants $\epsilon = \epsilon(\hat{\kappa})$, $\eta = \eta(\hat{\kappa})$, and $q = q(\hat{\kappa})$, such that any $(1-\epsilon)$-optimal solution to Opt 5 (in Definition D.6), the following holds. If $\hat{\kappa} > \kappa_{gen}$, then*

1. *$\phi(b) \ge \eta\phi(b+c)$; and*
2. *$\phi(-b) \ge \eta\phi(-b+d)$.*

If additionally $\hat{\kappa} < \kappa_{uc}$, then any such solution also satisfies

1. *$\phi(b) \le \frac{1-q}{2^h}\phi(b+c)$; and*
2. *$\phi(-b) \le \frac{1-q}{2^h}\phi(-b+d)$.*

Finally, if $\hat{\kappa} < \kappa_{gen}$, then at the optimum, $\phi(b) = \phi(-b) = 0$.

The following lemma is a more tailored version of a chaining lemma between Opt 4 and Opt 5, which explicitly leverages the previous lemmas on the solution of Opt 5.

Lemma D.16 (Opt 4 $\to$ Opt 5)**.** *Suppose $\hat{\kappa} > \kappa_{gen}$. There exists positive constants $\epsilon = \epsilon(\hat{\kappa})$, $\eta = \eta(\hat{\kappa})$, and $q = q(\hat{\kappa})$ such that for any $(1-\epsilon)$-optimal solution to the program Opt 4 in Definition D.5,*

$$\mathbb{E}_{i\in H}\phi(b_i) \ge \frac{\eta}{2}\mathbb{E}_{i\in H}\phi(b_i + c_i); \tag{D.26}$$

$$\mathbb{E}_{i\in H}\phi(-b_i) \ge \frac{\eta}{2}\mathbb{E}_{i\in H}\phi(-b_i + d_i). \tag{D.27}$$

$$\tag{D.28}$$

If additionally $\hat{\kappa} < \kappa_{uc}$,

$$\mathbb{E}_{i\in H}[\phi(b_i)] \le \left(\frac{1-q/2}{2^h}\right)\mathbb{E}_{i\in H}[\phi(b_i + c_i)]; \tag{D.29}$$

$$\mathbb{E}_{i\in H}[\phi(-b_i)] \le \left(\frac{1-q/2}{2^h}\right)\mathbb{E}_{i\in H}[\phi(-b_i + d_i)]. \tag{D.30}$$

Here $\eta(\hat{\kappa}), q(\hat{\kappa}) > 0$ are the constants from Lemma D.15.

D.2 KEY LEMMAS FOR MAIN RESULTS.

Putting together the results of Lemmas D.7-D.16, we carry out the chain of reductions from Opt 1 though Opt 5 to achieve the following results.

Lemma D.17 (Generalization Lemma)**.** *Assume $\kappa > \kappa_{gen}$. There exists strictly positive constants $\epsilon = \epsilon(\kappa)$ and $c = c(\kappa)$ and $\eta = \eta(\kappa)$ such that for any solution to Opt 1 which achieves a margin γ that is at least $(1-\epsilon)$-optimal, the following holds. For $\frac{d}{n} \geq c$, with probability $1 - 3e^{-n/c}$ over the training data, we have*

1. $\frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(\mu_1^T w_i)] \geq \frac{\gamma\eta}{2}$
2. $\frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(-\mu_1^T w_i)] \geq \frac{\gamma\eta}{2}$
3. $\frac{1}{2}\mathbb{E}_{i\in H_-}[\phi(\mu_2^T w_i)] \geq \frac{\gamma\eta}{2}$
4. $\frac{1}{2}\mathbb{E}_{i\in H_-}[\phi(-\mu_2^T w_i)] \geq \frac{\gamma\eta}{2}$.

Proof. Condition on the event in Lemma B.1 holding for Ξ and the event in Lemma D.1 holding for $\beta = \frac{1}{c_1}$, for some constant $c_1 = c_1(\kappa) > 8$ to be chosen later. These events occur with probability at least $\min(0, 1 - 3e^{-n} - 8e^{-8n/c_1^2}) \geq 1 - 3e^{-n/(c_1^2+1)}$ for $n \geq 1$. If we begin with an instance $\mathcal{I}_1$ which is $(1-\epsilon)$-optimal on Opt 1, then:

1. $\mathcal{I}_2 := \psi_{12}(\mathcal{I}_1)$ is $1 - \epsilon_2 := \sqrt{\left((1-\epsilon)\left(1 - C_{B.1}\sqrt{\frac{n}{d}}\right)^h\right)} \geq \sqrt{\left((1-\epsilon)\left(1 - C_{B.1}\sqrt{\frac{1}{c}}\right)^h\right)}$-optimal on Opt 2 (Lemma D.7)
2. Both instances $(\mathcal{I}_3^{(1)}, \mathcal{I}_3^{(2)}) := \psi_{23}(\mathcal{I}_2)$ are $1 - \epsilon_3 := 1 - \sqrt{1 - (1-\epsilon)\left(\frac{n_{\max}}{n_{\min}}\right)^{-h}} \geq 1 - \sqrt{1 - (1-\epsilon)\left(1 + \frac{16}{c_1}\right)^{-h}}$-optimal on Opt 3 (Lemma D.9)
3. For at least one pair $p = (j, j') \in \mathcal{P}_1 \times \mathcal{P}_{-1}$, the instance $\mathcal{I}_4^{(p)}$ of $\psi_{34}(\mathcal{I}_3^{(1)})$ indexed by that pair is $1 - \epsilon_4 := 1 - \sqrt{\epsilon_3}$-optimal on Opt 4. (Lemma D.10). The same holds for some $(j, j') \in \mathcal{N}_1 \times \mathcal{N}_{-1}$ on $\psi_{34}(\mathcal{I}_3^{(2)})$.

Now choose c_1 large enough and ϵ small enough constants such that for $c \geq c_1$, we have $\hat{\kappa} = \left(\frac{n_{\min}}{n_{\max}}\right)\kappa \geq \frac{\kappa_{\text{gen}}+\kappa}{2}$, and ϵ_4 is less than the value $\min_{\kappa' \in \left[\frac{\kappa_{\text{gen}}+\kappa}{2}, \kappa\right]} \epsilon(\kappa')$ from Lemma D.16. Thus applying Lemma D.16, we observe that on the pair $(j, j') \in \mathcal{P}_1 \times \mathcal{P}_{-1}$, we have

$$\mathbb{E}_{i\in H}\phi(b_i) \geq \frac{\eta}{2}\mathbb{E}_{i\in H}\phi(b_i + c_i) \geq \frac{\eta\gamma}{2}; \tag{D.31}$$

$$\mathbb{E}_{i\in H}\phi(-b_i) \geq \frac{\eta}{2}\mathbb{E}_{i\in H}\phi(-b_i + d_i) \geq \frac{\eta\gamma}{2}, \tag{D.32}$$

$$\tag{D.33}$$

where γ is the objective value of $\mathcal{I}_4^{((j,j'))}$, and $\eta = \eta(\kappa) := \min_{\kappa' \in \left[\frac{\kappa_{\text{gen}}+\kappa}{2}, \kappa\right]} \eta_{D.16}(\kappa')$ where $\eta_{D.16}(\cdot)$ is the positive constant called η from Lemma D.16. By definition of the mapping ψ_{34}, this means that in $\mathcal{I}_3^{(1)}$,

$$\mathbb{E}_{i\in H}\phi(b_i) \geq \frac{\eta\gamma}{2}; \tag{D.34}$$

$$\mathbb{E}_{i\in H}\phi(-b_i) \geq \frac{\eta\gamma}{2}, \tag{D.35}$$

$$\tag{D.36}$$

where γ is the objective value of $\mathcal{I}_3^{(1)}$. By definition of the mapping ψ_{23}, this means that in $\mathcal{I}_2$,

$$\mathbb{E}_{i\in H_+}\phi(s_i) \geq \frac{\eta\gamma}{2}; \tag{D.37}$$

$$\mathbb{E}_{i\in H_+}\phi(-s_i) \geq \frac{\eta\gamma}{2}, \tag{D.38}$$

$$\tag{D.39}$$

where γ is the objective value of $\mathcal{I}_2$. Finally by definition of ψ_{12}, in $\mathcal{I}_1$,

$$\frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(\mu_1^T w_i)] \geq \frac{\gamma\eta}{2} \tag{D.40}$$

$$\frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(-\mu_1^T w_i)] \geq \frac{\gamma\eta}{2}, \tag{D.41}$$

where γ is the objective value of Opt 1. This yields the first two conclusions of the lemma. The second follows via an identical argument on the pair $(j, j') \in \mathcal{N}_1 \times \mathcal{N}_{-1}$. Choosing $c = c_1^2 + 1$ yields the result with probability at least $1 - 3e^{-n/c}$. □

We can now put together the results of the chain of reductions to prove the following:

Lemma D.18 (Phenomenon Lemma)**.** *Assume $\kappa_{gen} < \kappa < \kappa_{uc}$. For any constant $\delta > 0$, there exists strictly positive constants $\epsilon = \epsilon(\kappa, \delta)$ and $c = c(\kappa, \delta)$ and $q = q(\kappa)$ such that for any solution to Opt 1 which achieves a margin γ that is at least $(1-\epsilon)$-optimal, the following holds. For $\frac{d}{n} \geq c$, with probability $1 - 3e^{-n/c}$ over the training data,*

$$\frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(u_i^T \mu_1)] \leq \frac{\gamma(1-q/4)}{2^h}; \tag{D.42}$$

$$\frac{1}{2}\mathbb{E}_{i\in H_-}[\phi(u_i^T \mu_2)] \leq \frac{\gamma(1-q/4)}{2^h}, \tag{D.43}$$

and for at least a $1-\delta$ fraction of $j \in [n]$,

$$\frac{1}{2}\mathbb{E}_{i:\mathrm{sign}(a_i)=y_j}[\phi(v_i^T x_j)] \geq \frac{\gamma(1+q/4)}{2^h}. \tag{D.44}$$

Proof. Condition on the event in Lemma B.1 holding for Ξ and the event in Lemma D.1 holding for $\beta = \frac{1}{c_1}$, for some constant $c_1 = c_1(\delta, \kappa) > 8$ to be chosen later. These events occur with probability at least $\min(0, 1 - 3e^{-n} - 8e^{-8n/c_1^2}) \geq 1 - 3e^{-n/(c_1^2+1)}$. If we begin with an instance $\mathcal{I}_1$ which is $(1-\epsilon)$-optimal on Opt 1, then:

1. $\mathcal{I}_2 := \psi_{12}(\mathcal{I}_1)$ is $1 - \epsilon_2 := \sqrt{\left((1-\epsilon)\left(1 - C_{B.1}\sqrt{\frac{n}{d}}\right)^h\right)} \geq \sqrt{\left((1-\epsilon)\left(1 - C_{B.1}\sqrt{\frac{1}{c}}\right)^h\right)}$-optimal on Opt 2 (Lemma D.7)
2. Both instances $(\mathcal{I}_3^{(1)}, \mathcal{I}_3^{(2)}) := \psi_{23}(\mathcal{I}_2)$ are $1 - \epsilon_3 := 1 - \sqrt{1 - (1-\epsilon)\left(\frac{n_{\max}}{n_{\min}}\right)^{-h}} \geq 1 - \sqrt{1 - (1-\epsilon)\left(1 + \frac{16}{c_1}\right)^{-h}}$-optimal on Opt 3 (Lemma D.9)
3. Let $\{\mathcal{I}_4^{((p)}\}_{p\in L_1} := \psi_{34}(\mathcal{I}_3^{(1)})$ be the $n_{\min}$ instances of Opt 4 indexed by some list L_1 of $n_{\min}$ pairs $p \in \mathcal{P}_1 \times \mathcal{P}_{-1}$. Similarly let $\{\mathcal{I}_4^{((p)}\}_{p\in L_2} := \psi_{34}(\mathcal{I}_3^{(2)})$ for some list of $n_{\min}$ pairs in $\mathcal{N}_1 \times \mathcal{N}_{-1}$. For at least and $1 - \epsilon_4 := 1 - \sqrt{\epsilon_3}$ fraction of pairs (j, j') in L_1, $\mathcal{I}_4^{((j,j'))}$ is $1-\epsilon_4$-optimal on Opt 4. The same hols for a $1-\epsilon_4$ fraction of the pairs in L_2. (Lemma D.10). For each cluster in $\mathcal{P}_1, \mathcal{P}_{-1}, \mathcal{N}_1, \mathcal{N}_{-1}$, at least a $\frac{n_{\min}}{n_{\max}} \geq \frac{1}{1+\frac{16}{c_1}}$ fraction of the points j in that cluster appear in a pair in one of the lists L_1 or L_2.

Now choose c_1 large enough and ϵ small enough constants such that for $c \geq c_1$, we have $\hat{\kappa} = \left(\frac{n_{\min}}{n_{\max}}\right)\kappa \geq \frac{\kappa_{\text{gen}}+\kappa}{2}$, and $\frac{\epsilon_4}{1+\frac{16}{c_1}}$ is both less than δ and than $\min_{\kappa' \in \left[\frac{\kappa_{\text{gen}}+\kappa}{2}, \kappa\right]} \epsilon(\kappa')$, where $\epsilon(\cdot)$ is the function from Lemma D.16.

Applying this Lemma D.16, we observe that on at least a $1-\epsilon_4$ fraction of pairs (j, j') in the list L_1, we have

$$\mathbb{E}_{i\in H}[\phi(b_i)] \leq \left(\frac{1-q/2}{2^h}\right)\mathbb{E}_{i\in H}[\phi(b_i + c_i)]; \tag{D.45}$$

$$\mathbb{E}_{i\in H}[\phi(-b_i)] \leq \left(\frac{1-q/2}{2^h}\right)\mathbb{E}_{i\in H}[\phi(-b_i + d_i)], \tag{D.46}$$

where $q = q(\kappa) := \min_{\kappa' \in \left[\frac{\kappa_{\text{gen}}+\kappa}{2}, \kappa\right]} q_{D.16}(\kappa')$ where $q_{D.16}(\cdot)$ is the positive constant called q from Lemma D.16. We first use this to prove the first conclusion the lemma for j (the argument is analogous for j'). By definition of the mapping ψ_{34}, this means that in $\mathcal{I}_3^{(1)}$,

$$\mathbb{E}_{i \in H}[\phi(b_i)] \leq \left(\frac{1 - q/2}{2^h}\right) \mathbb{E}_{i \in H}[\phi(b_i + c_{ij})]; \tag{D.47}$$

By definition of the mapping ψ_{23}, this means that in $\mathcal{I}_2$,

$$\mathbb{E}_{i \in H_+}[\phi(s_i)] \leq \left(\frac{1 - q/2}{2^h}\right) \mathbb{E}_{i \in H}[\phi(s_i + c_{ij})]; \tag{D.48}$$

Finally, by definition of the mapping ψ_{12}, this means that in $\mathcal{I}_1$,

$$\mathbb{E}_{i \in H_+}[\phi(\mu_1^T x_j)] \leq \left(\frac{1 - q/2}{2^h}\right) \mathbb{E}_{i \in H_+}[\phi(\mu_1^T x_j + v_i^T \xi_j)] = \left(\frac{1 - q/2}{2^h}\right) \mathbb{E}_{i \in H_+}[\phi(w_i^T x_j)]. \tag{D.49}$$

Now by Lemma B.2, for any values s, t, we have $\phi(s + t) \leq (\phi(s) + \phi(t))2^{h-1}$, and thus

$$\mathbb{E}_{i \in H_+}[\phi(u_i^T x_j) + \phi(v_i^T \xi_j)] \geq 2^{-h+1} \mathbb{E}_{i \in H_+}[\phi(w_i^T x_j)] \tag{D.50}$$

So by Equation D.49, we have

$$\mathbb{E}_{i \in H_+}[\phi(v_i^T \xi_j)] \geq \left(\frac{1 + q/2}{2^h}\right) \mathbb{E}_{i \in H_+}[\phi(w_i^T x_j)]. \tag{D.51}$$

Now to relate $\mathbb{E}_{i \in H_+}[\phi(w_i^T x_j)]$ to γ, observe that by Lemma D.7, D.9, and D.10, the objective value of $\mathcal{I}_4^{(j,j')}$, which equals $\min\left(\frac{1}{2}\mathbb{E}_{i \in H_+}\phi(w_i^T x_j), \frac{1}{2}\mathbb{E}_{i \in H_+}\phi(w_i^T x_{j'})\right)$, is at least γ and at most $\frac{\gamma}{(1-\epsilon_2)(1-\epsilon_3)(1-\epsilon_4)}$. Now we also know by the first conclusion of Lemma D.11 that $\psi_{45}(\mathcal{I}_4^{(j,j')})$ produces some instances $\mathcal{I}_5^{(i)}$ for $i \in H_+$ with objective values $\gamma^{(i)}$, for which

$$\mathbb{E}_{i \in H^+}\gamma^{(i)} \leq \frac{1}{1 - \epsilon_4} \min\left(\frac{1}{2}\mathbb{E}_{i \in H_+}\phi(w_i^T x_j), \frac{1}{2}\mathbb{E}_{i \in H_+}\phi(w_i^T x_{j'})\right). \tag{D.52}$$

Plugging in the fact that $\gamma^{(i)} = \frac{1}{4}\left(\phi(b_i + c_i) + \phi(-b_i + d_i)\right) = \frac{1}{4}\left(\phi(w_i^T x_j) + \phi(w_i^T x_{j'})\right)$, we have

$$\frac{1}{4}\mathbb{E}_{i \in H^+}\left(\phi(w_i^T x_j) + \phi(w_i^T x_{j'})\right) \leq \frac{1}{1 - \epsilon_4} \min\left(\frac{1}{2}\mathbb{E}_{i \in H_+}\phi(w_i^T x_j), \frac{1}{2}\mathbb{E}_{i \in H_+}\phi(w_i^T x_{j'})\right). \tag{D.53}$$

Since $\min(\alpha, \beta) \geq (1 - \rho)\frac{a+b}{2}$ implies that $a, b \in \max(a, b) \leq \frac{1+\rho}{1-\rho}$, we must have that

$$\frac{1}{2}\mathbb{E}_{i \in H_+}\phi(w_i^T x_j) \in \gamma\left[1, \frac{1 + \epsilon_4}{(1 - \epsilon_2)(1 - \epsilon_3)(1 - \epsilon_4)^2}\right]. \tag{D.54}$$

Thus for ϵ small enough in terms of q, we have ϵ_2, ϵ_3, and ϵ_4 all small enough that from Equations D.49 and D.51, we have

$$\mathbb{E}_{i \in H_+}[\phi(\mu_1^T x_j)] \leq \left(\frac{1 - q/2}{2^h}\right) \mathbb{E}_{i \in H_+}[\phi(w_i^T x_j)] \leq \left(\frac{1 - q/4}{2^{h-1}}\right) \gamma \tag{D.55}$$

$$\mathbb{E}_{i \in H_+}[\phi(v_i^T \xi_j)] \geq \left(\frac{1 + q/2}{2^h}\right) \mathbb{E}_{i \in H_+}[\phi(w_i^T x_j)] \geq \left(\frac{1 + q/4}{2^{h-1}}\right) \gamma. \tag{D.56}$$

The argument holds for j' in the pair with j, and we can also repeat an analogous argument for the $1 - \epsilon_4$ fraction of pairs in the list of pairs L_2 from $\mathcal{N}_1 \times \mathcal{N}_{-1}$. Now at least a $1 - \delta$ fraction of examples j lie in a pair p from L_1 or L_2 for which $\mathcal{I}_4^{(p)}$ is $(1 - \epsilon_4)$-optimal. This yields the second part of the lemma which involves specific data points. The first point follows for the fact that EquationD.55 only needs to hold for a single example j in each of the four clusters. Choosing $c = c_1^2 + 1$ yields the result with probability at least $1 - 3e^{-n/c}$.

□

Lemma D.19 (No Generalization Lemma)**.** *For any $\kappa < \kappa_{gen}$ and $\epsilon > 0$, there exists some positive constant $c(\epsilon)$, such that if $\frac{d}{n} \geq c$, with probability at least $1 - 3e^{-n/c}$ over $S \sim \mathcal{D}^n$, there exists a classifier W with $\|W\| = 1$ such that*

1. $\gamma(f_W, S) \geq (1 - \epsilon)\gamma^*(S)$

2. $U = 0$.

Proof. We work backwards from Opt 5 through Opt 1. Condition on the event in Lemma B.1 holding for Ξ and the event in Lemma D.1 holding for $\beta = \frac{1}{c_1}$, for some constant $c_1(\epsilon) > 8$ to be chosen later. Given an optimal solution $\mathcal{I}_5^*$ to Opt 5, we can construct an instance $\mathcal{I}_1 = \psi_{21}(\psi_{32}(\psi_{43}(\psi_{54}(\mathcal{I}_5))))$, which is ϵ'-optimal over all solutions W' with the same norm for $\epsilon' = \sqrt{1-(1-\hat{\epsilon})(1-C_{B.1}\sqrt{\frac{n}{d}})^h}$, where $\hat{\epsilon} = \sqrt{1-\left(\frac{n_{\max}}{n_{\min}}\right)^{-h}} \leq \sqrt{1-\left(1+\frac{16}{c_1}\right)^{-h}}$. This can be seen via Lemmas D.7, D.9, D.10, D.11, which show that at each step of the chain, we do not lose any optimality expect from from Opt 3 to Opt 2 and from Opt 2 to Opt 1.

Now recall from Lemma D.15 that since $\kappa < \kappa_{\text{gen}}$, for large enough c_1, we have $\hat{\kappa} < \kappa_{\text{gen}}$, and thus the optimal solution to Opt 5 has $b = 0$. Applying the four mappings above, in the instance $\mathcal{I}_1$, the variable $W = U + V$ has $U = 0$. Taking c_1 large enough such that for $c \geq c_1$, we have $\epsilon' \leq \epsilon$. If we choose $c(\epsilon) = c_1^2 + 1$, then the desired events hold with probability at least $1 - 3e^{-n/c}$ (see eg. Lemma D.17 for the computation). Scaling W to have $\|W\| = 1$ concludes the lemma. □

D.3 Proofs of Main Results

Using Lemma D.17, Lemma D.8, and Lemma A.2, we can prove Theorem 3.2. We restate the theorem for the reader's convenience.

Theorem 3.2 (Extremal-Margin Generalization for XOR on Neural Network)**.** *Let $h \in (1,2)$, and let $\delta > 0$. There exist constants $\epsilon = \epsilon(\delta)$ and $c = c(\delta)$ such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega_{\sigma,d}^{\text{XOR}}$ satisfying $\kappa = \frac{n}{d\sigma^2} \geq \kappa_{gen}^{\text{XOR},h} + \delta$ and $\frac{d}{n} \geq c$, then with probability $1 - 3e^{-n/c}$ over the training set $S \sim \mathcal{D}^n$, for any two-layer neural network with activation function relu^h and weight matrix W that is a $(1-\epsilon)$-max-margin solution (as in Definition 2.5), we have $\mathcal{L}_{\mathcal{D}}(f_W) \leq e^{-\frac{1}{c\sigma^2}}$.*

Proof of Theorem 3.2. First note that $\kappa_{\text{gen}} = \kappa_{\text{gen}}^{\text{XOR},h}$. Let $W = U + V$ be the decomposition of W into the signal space and the orthogonal space, such that $V \perp \text{span}(\mu_1, \mu_2)$. It suffices to consider W with $\|W\| = 1$. Let γ be the margin achieved by f_W, and observe that γ is at least a positive constant since we can achieve a margin of $\frac{1}{4}$ by choosing a solution that only uses components in the signal subspace.

Recall that $\|U\| < 1$ and $\|V\| < 1$, and consider a random $x \sim \mathcal{D}$.

Choosing to $c_0 = c(\kappa)$ and $\epsilon(\kappa)$ to be the values from Lemma D.17, if $\frac{d}{n} \geq c_0$, with probability $1 - 3e^{-n/c_0}$ over the training data (and not x), the conclusion of Lemma D.17 and Lemma D.8 hold, and thus we have for such a W:

1. $\frac{1}{2}\mathbb{E}_{i:\text{sign}(a_i)=y}[\phi(u_i^T x)] \geq \frac{\gamma\eta(\kappa)}{2}$ by Lemma D.17, since $\kappa > \kappa_{\text{gen}}$.
2. $\frac{1}{2}\mathbb{E}_{i:\text{sign}(a_i)=-y}[\phi(u_i^T x)] \leq \left(2\epsilon + 2C_{B.1}\sqrt{\frac{n}{d}}\right)^{\frac{h}{2}}$. This is by Lemma D.8, we have $\frac{1}{2}\mathbb{E}_{i:\text{sign}(a_i)=-y}[\|u_i^T z\|^2] \leq 2\epsilon + 2C_{B.1}\sqrt{\frac{n}{d}}$, and thus by he homogeneity of the activation, $\frac{1}{2}\mathbb{E}_{i:\text{sign}(a_i)=-y}[\phi(u_i^T x)] \leq \left(2\epsilon + 2C_{B.1}\sqrt{\frac{n}{d}}\right)^{\frac{h}{2}} \max_{X:\mathbb{E}[X^2]=1} \mathbb{E}[\phi(X)]$.By Jenson's inequality $\max_{X:\mathbb{E}[X^2]=1} \mathbb{E}[\phi(X)] \leq \max_{X:\mathbb{E}[X^2]=1} \left(\mathbb{E}[X^2]\right)^{\frac{h}{2}} = 1$.

Thus with probability $1 - 3e^{-n/c_0}$ over the training data,

$$y f_U(x) = \mathbb{E}_{i:\text{sign}(a_i)=y}[\phi(u_i^T x)] - \mathbb{E}_{i:\text{sign}(a_i)=-y}[\phi(u_i^T x)] \geq \frac{\gamma\eta(\kappa)}{2} - \left(2\epsilon + 2C_{B.1}\sqrt{\frac{n}{d}}\right)^{\frac{h}{2}}. \quad \text{(D.57)}$$

For some constant $c_1 = c_1(\kappa)$, by Lemma A.2, with probability $1 - e^{-\frac{1}{c_1\sigma^2}}$ over x, $|f_W(x) - f_U(x)| < \frac{\gamma\eta(\kappa)}{4}$. Here we plugged in $t = \frac{1}{100(\gamma^{\frac{2}{h}}+\gamma)\left(\eta(\kappa)^{\frac{2}{h}}+\eta(\kappa)\right)\sigma^2}$ to Lemma A.2, and note that γ is at least a constant.

Thus for $\frac{d}{n} \geq c_2$ for some $c_2 = c_2(\kappa)$, we have with probability at least $1 - 3e^{-n/c_0}$,

$$\frac{\gamma\eta(\kappa)}{2} - \left(2\epsilon + 2C_{B.1}\sqrt{\frac{n}{d}}\right)^{\frac{h}{2}} \geq \frac{\gamma\eta(\kappa)}{4},$$

and thus the loss is at most $1 - e^{-\frac{1}{c_1\sigma^2}} = 1 - e^{-\frac{\kappa d}{c_1 n}}$. Choosing $c = \max(c_0, c_1, c_2)$ yields the theorem. □

We now use Lemma D.18, Lemma D.8, and Lemma A.2 to prove Proposition 3.4(the XOR part) and Theorem 3.6 on the limitations of uniform convergence and inverse margin bounds.

To prove these results, we will demonstrate two phenomenons:

1. Given a near max-margin classifier f_W for a a certain "opposited" dataset $\bar{\psi}(S)$, the classifier f_W completely misclassifies a certain the data set S while still achieving good margin on the distribution from which S and $\bar{\psi}(S)$ are drawn.
2. Given a near max-margin classifier f_W for a set S, the classifier f_W correctly classifies a certain "opposite" dataset $\psi(S)$ while still achieving good margin on this opposite dataset.

We will use the first phenomenon to prove Proposition 3.4, and we will define the mapping $\bar{\psi}$ as follows:

Definition D.20. *For $\mathcal{D} = \mathcal{D}_{\mu_1,\mu_2,\sigma,d} \in \Omega$, define the map $\bar{\psi} : \mathbb{R}^d \to \mathbb{R}^d$ to keep ξ the same, but map z to be in an orthogonal direction, and reverse y as follows:*

$$\bar{\psi}((x,y)) = \begin{cases} (\mu_2 + \xi, 1) & (x,y) = (\mu_1 + \xi, -1) \\ (-\mu_2 + \xi, 1) & (x,y) = (-\mu_1 + \xi, -1) \\ (\mu_1 + \xi, -1) & (x,y) = (\mu_2 + \xi, 1) \\ (-\mu_1 + \xi, -1) & (x,y) = (-\mu_2 + \xi, 1). \end{cases} \tag{D.58}$$

We will use the second phenomenon to prove Proposition 3.6, and we will define the mapping ψ as follows:

Definition D.21. *For $\mathcal{D} = \mathcal{D}_{\mu_1,\mu_2,\sigma,d} \in \Omega$, define the map $\psi : \mathbb{R}^d \to \mathbb{R}^d$ to keep ξ the same, but map z to be in an orthogonal direction, as follows:*

$$\psi((x,y)) = \begin{cases} (\mu_2 + \xi, 1) & (x,y) = (\mu_1 + \xi, 1) \\ (-\mu_2 + \xi, 1) & (x,y) = (-\mu_1 + \xi, 1) \\ (\mu_1 + \xi, -1) & (x,y) = (\mu_2 + \xi, -1) \\ (-\mu_1 + \xi, -1) & (x,y) = (-\mu_2 + \xi, -1). \end{cases} \tag{D.59}$$

When it is clear that we have fixed $\mathcal{D}$, we will just use ψ or $\bar{\psi}$ to denote this mapping. Otherwise, we will specify that we mean the mapping associated with $\mathcal{D}$ by denoting it $\psi_{\mathcal{D}}$ (or $\bar{\psi}_{\mathcal{D}}$).

For $\phi \in \{\psi, \bar{\psi}\}$, we abuse notation and denote $(\phi(x), \psi(y)) := \phi(x,y)$, and for a set S, use $\phi(S)$ to denote the element-wise application of ϕ. For $\mathcal{D} = \mathcal{D}_{\mu_1,\mu_2,\sigma,d}$, we also denote $\psi(\mathcal{D}) = \mathcal{D}_{\mu_2,\mu_1,\sigma,d}$ to be the distribution with the opposite labeling ground truth. Thus the following claim is immediate:

Claim D.22. *Fix $\mathcal{D} \in \Omega$. For any W, we have $\mathcal{L}_{\mathcal{D}}(f_W) = 1 - \mathcal{L}_{\psi(\mathcal{D})}(f_W)$.*

Observe also that ψ is a measure preserving bijection from $\mathcal{D}$ to $\psi(\mathcal{D})$, and that $\bar{\phi}(\mathcal{D}) = \mathcal{D}$.

To prove Proposition 3.4 we will use the following lemma, which shows that with high probability, any near-max-margin classifier does well on the "opposite" dataset, but has poor test loss on the opposite distribution.

Lemma D.23. *Suppose $\kappa_{gen} \le \kappa \le \kappa_{uc}$. Let $\mathcal{A}$ be any algorithm which returns a $(1-\epsilon)$-max margin solution. For a dataset $S \sim \mathcal{D}^n = \mathcal{D}^n_{\mu_1,\mu_2,\sigma,d}$, consider the classifier $W = \mathcal{A}(S)$. For any constant $\delta > 0$, there exists constants $\epsilon(\delta,\kappa)$ and $c = c(\kappa,\delta)$ such that if $\epsilon \le \epsilon(\delta,\kappa)$ and $\frac{d}{n} \ge c$, with probability at least $1 - 3e^{-n/c}$ over $S \sim \mathcal{D}^n$, we have*

1. $\mathcal{L}_{\psi(\mathcal{D})}(f_W) \ge 1 - e^{-\frac{\kappa d}{cn}}$.
2. $\mathcal{L}_{\psi(S)}(f_W) \le \delta$.
3. $\mathcal{L}_{\bar{\psi}(S)}(f_W) \ge 1 - \delta$.

Proof. The first statement follows from Theorem 3.2, since the probability of classifying a example from $\psi(\mathcal{D})$ correctly is the same as the probability of misclassifying a example from $\mathcal{D}$ (Claim D.22).

We now prove the second statement. We expand the margin on the examples in $\psi(S)$. For clarity, we will assume we are expanding on a example from $\psi(x_j)$ where $j \in \mathcal{P}_1$, such that by definition of ψ, we have $\psi(x_j) = \mu_2 + \xi_j$. The same argument will apply to examples mapped from any other cluster by interchanging the roles of the four vectors $\mu_1, -\mu_1, \mu_2, -\mu_2$ and the two sets H_+ and H_- accordingly.

$$\psi(y_j) f_W(\psi(x_j)) = \mathbb{E}_i[y_j \phi(w_i^T \psi(x_j))] \tag{D.60}$$

$$= \frac{1}{2}\mathbb{E}_{i \in H_+ = y_j}[\phi(w_i^T \mu_2 + v_i^T \xi_j)] - \frac{1}{2}\mathbb{E}_{i \in H_-}[\phi(w_i^T \mu_2 + v_i^T \xi_j)] \tag{D.61}$$

By Lemma D.26 (second statement), we have the following:

$$|\mathbb{E}_{i\in H_+}[\phi(w_i^T\mu_2 + v_i^T\xi_j)] - \mathbb{E}_{i\in H_+}[\phi(v_i^T\xi_j)]| \tag{D.62}$$

$$\leq 2\mathbb{E}_{i\in H_+}\left[\phi(v_i^T\xi_j)^2 + 1\right]\sqrt{\mathbb{E}_{i\in H_+}\left[(w_i^T\mu_2)\right]} + h\left(\mathbb{E}_{i\in H_+}[(w_i^T\mu_2)^2]\right)^{\frac{h}{2}}. \tag{D.63}$$

Similarly appealing to Lemma D.26 (first statement), we have

$$|\mathbb{E}_{i\in H_-}[\phi(w_i^T\mu_2 + v_i^T\xi_j)] - \mathbb{E}_{i\in H_-}[\phi(w_i^T\mu_2)]| \tag{D.64}$$

$$\leq \sqrt{\mathbb{E}_{i\in H_-}\left[4(w_i^T\mu_2)^2 + 2\right]}\sqrt{\mathbb{E}_{i\in H_-}\left[(v_i^T\xi_j)^2\right]} + h\left(\mathbb{E}_{i\in H_-}[(v_i^T\xi_j)^2]\right)^{\frac{h}{2}} \tag{D.65}$$

$$\leq \left(2\|U\| + \sqrt{2}\right)\sqrt{\mathbb{E}_{i\in H_-}\left[(v_i^T\xi_j)^2\right]} + h\left(\mathbb{E}_{i\in H_-}[(v_i^T\xi_j)^2]\right)^{\frac{h}{2}} \tag{D.66}$$

$$\tag{D.67}$$

Thus

$$\psi(y_j)f_W(\psi(x_j)) = \frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(w_i^T\mu_2 + v_i^T\xi_j)] - \frac{1}{2}\mathbb{E}_{i\in H_-}[\phi(w_i^T\mu_2 + v_i^T\xi_j)] \tag{D.68}$$

$$\geq \frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(v_i^T\xi_j)] - \frac{1}{2}\mathbb{E}_{i\in H_-}[\phi(w_i^T\mu_2)] - \frac{1}{2}\mathcal{E}_j, \tag{D.69}$$

where

$$\mathcal{E}_j := 2\mathbb{E}_{i\in H_+}\left[\phi(v_i^T\xi_j)^2 + 1\right]\sqrt{\mathbb{E}_{i\in H_+}\left[(w_i^T\mu_2)^2\right]} + h\left(\mathbb{E}_{i\in H_+}[(w_i^T\mu_2)^2]\right)^{\frac{h}{2}} \tag{D.70}$$

$$+ 4\sqrt{\mathbb{E}_{i\in H_-}\left[(v_i^T\xi_j)^2\right]} + h\left(\mathbb{E}_{i\in H_-}[(v_i^T\xi_j)^2]\right)^{\frac{h}{2}}, \tag{D.71}$$

$$\tag{D.72}$$

where we have plugged in the fact that $\|U\| \leq \|W\| \leq 1$.

By the first and second conclusions of Lemma D.8, for $\kappa_{\text{gen}} < \kappa < \kappa_{\text{uc}}$, for at least a $1 - \epsilon_{D.8}$ a set of examples $T \subset S$ of size at least $(1 - \epsilon'_{D.8}n)$, we have if $j \in T$, $\frac{1}{2}\mathbb{E}_{i\in H_+}\left[(w_i^T\mu_2)^2\right] \leq (\epsilon'_{D.8})^2$ and $\frac{1}{2}\mathbb{E}_{i\in H_-}\left[(v_i^T\xi_j)^2\right] \leq \frac{1}{\kappa}\cdot\epsilon'_{D.8}$, where $\epsilon'_{D.8} = \sqrt{2C_{B.1}\sqrt{\frac{n}{d}} + 2\epsilon}$. Thus if $j \in T$,

$$\mathcal{E}_j \leq 2\mathbb{E}_{i\in H_+}\left[\phi(v_i^T\xi_j)^2 + 1\right]\sqrt{\frac{2\epsilon'_{D.8}}{\kappa}} + h\left(\sqrt{2}\epsilon'_{D.8}\right)^h + 4\sqrt{\frac{2\epsilon'_{D.8}}{\kappa}} + h\left(\frac{2\epsilon'_{D.8}}{\kappa}\right)^{\frac{h}{2}} \tag{D.73}$$

$$\leq 8\mathbb{E}_{i\in H_+}\left[\phi(v_i^T\xi_j)^2 + 1\right]\sqrt{\frac{2\epsilon'_{D.8}}{\kappa}} + 4\epsilon'_{D.8} \tag{D.74}$$

for $\epsilon'_{D.8}$ small enough.

Now by Lemma D.18 applied to S, if $\epsilon \leq \epsilon(\kappa, \delta/2)$ and $\frac{d}{n} \geq c(\kappa, \delta/2)$, there exists a set $T' \subset S$ size at least $(1 - \frac{\delta}{2})n$ on which the second conclusion of the lemma holds. Thus for the constant $q = q(\kappa)$ in Lemma D.18,

$$\frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(u_i^T\mu_1)] \leq \frac{(1 - q/4)}{2^h}\gamma, \tag{D.75}$$

$$\frac{1}{2}\mathbb{E}_{i\in H_-}[\phi(u_i^T\mu_2)] \leq \frac{(1 - q/4)}{2^h}\gamma, \tag{D.76}$$

and if $j \in T'$,

$$\frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(v_i^T x_j)] \geq \frac{(1 + q/4)}{2^h}\gamma. \tag{D.77}$$

Thus

$$\psi(y_j)f_W(\psi(x_j)) \geq \frac{1}{2}\mathbb{E}_{i\in H_+}[\phi(v_i^T\xi_j)] - \frac{1}{2}\mathbb{E}_{i\in H_-}[\phi(u_i^T\mu_2)] - \frac{1}{2}\mathcal{E}_j \tag{D.78}$$

$$\geq \frac{\gamma(1 + q/4)}{2^h} - \frac{\gamma(1 - q/4)}{2^h} - \frac{1}{2}\mathcal{E}_j \tag{D.79}$$

which is greater than zero for ϵ small enough and c large enough in terms of q and δ. (In particular, we will need that $\mathcal{E}_j \leq \frac{\gamma q}{2^{h+1}}$ and $\epsilon_{D.8} \leq \frac{\delta}{2}$, and note that γ is at least a constant).

Thus f_W correctly classifies each example $\psi(x_j)$ for $j \in T \cap T'$, which is at least a $1-\delta$ fraction of the examples in $\psi(S)$.

The final statement follows from the fact that $\psi(\bar{S})$ has the opposite labels as $\psi(S)$, but is otherwise the same. □

We now prove Proposition 3.4 for the XOR problem. We restate the proposition below, and only include the XOR part.

Proposition D.24 (UC Bounds are Vacuous for XOR Problem)**.** *Fix any $h \in (1,2)$, and $\delta > 0$. For any n, d, σ and $\mathcal{D} \in \Omega^{\mathrm{XOR}}_{\sigma,d}$, if $\kappa^{\mathrm{XOR},h}_{gen} + \delta \le \kappa \le \kappa^{\mathrm{XOR},h}_{uc} - \delta$, there exist strictly positive constants $\epsilon = \epsilon(\delta)$ and $c = c(\delta)$ such that the following holds. Let $\mathcal{A}$ be any algorithm that outputs a $(1-\epsilon)$-max-margin two-layer neural network f_W for any $S \in (\mathbb{R}^d \times \{1,-1\})^n$. Let $\mathcal{H}$ be any concept class that is useful for $\mathcal{D}$ (as in Definition 2.2). Suppose that ϵ_{unif} is a uniform convergence bound for $\mathcal{D}$ and $\mathcal{H}$ that is,*

$$\Pr_{S\sim\mathcal{D}^n}[\sup_{h\in\mathcal{H}} |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| \ge \epsilon_{\mathrm{unif}}] \le 1/4.$$

Then if $\frac{d}{n} \ge c$ and $n > c$, we must have $\epsilon_{\mathrm{unif}} \ge 1-\delta$.

Proof. Let $c = 3c_0$ and $\epsilon = 3\epsilon_0$ where c_0 and ϵ_0 are the constants from Lemma D.23 for κ and δ.

Let $T_{\mathcal{D}} \subset 2^{(\mathbb{R}^d\times\{-1,1\})^n}$ be the set of training sets S on which the conclusion of Lemma D.23 holds for $\mathcal{D}$ and S. Thus $\Pr_{S\sim\mathcal{D}^n}[S \in T_{\mathcal{D}}] \ge 1 - 3e^{-n/c_0}$ for some $c_0 = c(\kappa,\delta)$. Let $H \subset 2^{(\mathbb{R}^d\times\{-1,1\})^n}$ be the set of training sets S on which $\mathcal{A}(S) \in \mathcal{H}$. Thus $\Pr_{S\sim\mathcal{D}^n}[S \in H] \ge \frac{3}{4}$.

Let $T'_{\mathcal{D}}$ be the set on which

$$|\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_{\phi(S)}(h)| \le \epsilon_{\mathrm{unif}} \qquad \forall h \in \mathcal{H}, \tag{D.80}$$

where $\phi := \bar{\psi}$. By assumption, $\Pr_{S\sim\mathcal{D}^n}[\phi(S) \in T'_{\mathcal{D}}] = \Pr_{S\sim\mathcal{D}^n}[S \in T'_{\mathcal{D}}] \ge \frac{3}{4}$. By a union bound, for any $\mathcal{D} \in \Omega$, for $n \ge c = 3c_0$, with $\phi = \bar{\psi}_{\mathcal{D}}$,

$$\Pr_{S\sim\mathcal{D}^n}[S \in T'_{\mathcal{D}} \wedge S \in T_{\phi(\mathcal{D})} \wedge S \in H] \ge 1 - \left(1 - \frac{3}{4}\right) - \left(1 - \frac{3}{4} + 3e^{-n}\right) = \frac{1}{2} - 3e^{-n/c_0} > 0. \tag{D.81}$$

Let S be any set for which the three events above hold, ie.,

$$S \in T'_{\mathcal{D}} \wedge S \in T_{\psi(\mathcal{D})} \wedge S \in H. \tag{D.82}$$

With $f_W = \mathcal{A}(S)$, by the first conclusion of Lemma D.23, we have $\mathcal{L}_{\mathcal{D}}(f_W) \le e^{-\frac{1}{c_0\sigma^2}}$. Further, by the third conclusion of Lemma D.23, we know that $\mathcal{L}_{\phi(S)}(f_W) \ge 1-\delta$. It follows that $\epsilon_{\mathrm{unif}} \ge 1 - e^{-\frac{1}{c_0\sigma^2}}$. Since $c > c_0$, this yields the proposition. □

Lemma D.25 (Margin Lower Bound Lemma)**.** *For any $\kappa_{gen} < \kappa < \kappa_{uc}$ and $\epsilon > 0$, there exists some positive constants $q(\kappa) > 0$, $c = c(\kappa,\varepsilon)$ such that if $\frac{d}{n} \ge c$ with probability at least $1 - 3e^{-n/c}$ over $S \sim \mathcal{D}^n = \mathcal{D}^n_{\mu_1,\mu_2,\sigma,d}$, there exists a classifier W with $\|W\| = 1$ such that*

1. $\gamma(f_W, S) \ge (1-\epsilon)\gamma^*(S)$
2. $\mathcal{L}_{\psi(\mathcal{D})}(f_W) \ge 1 - e^{-\frac{1}{c\sigma^2}}$.
3. $\gamma(f_W, \psi(S)) \ge q(\kappa)\gamma^*(S)$

Proof. We work backwards from Opt 5 through Opt 1. Condition on the event in Lemma B.1 holding for Ξ and the event in Lemma D.1 holding for $\beta = \frac{1}{c_0}$, for some constant $c_0(\kappa,\epsilon) > 8$ to be chosen later. We will eventually choose $c(\kappa,\epsilon) \ge c_0^2 + 1$, such that these events hold with probability at least $1 - 3e^{-n/c}$ (see eg. Lemma D.17 for the computation).

Given an optimal solution $\mathcal{I}^*_5$ to Opt 5, we can construct an instance $\mathcal{I}_1 = \psi_{21}(\psi_{32}(\psi_{43}(\psi_{54}(\mathcal{I}_5))))$, which is ϵ'-optimal over all solutions W' with the same norm for $\epsilon' = \sqrt{1-(1-\hat{\epsilon})(1-C_{B.1}\sqrt{\frac{n}{d}})^h}$, where $\hat{\epsilon} = \sqrt{1-\left(\frac{n_{\max}}{n_{\min}}\right)^{-h}} \le \sqrt{1-\left(1+\frac{16}{c_0}\right)^{-h}}$. This can be seen via Lemmas D.7, D.9, D.10, D.11, which show that at each step of the chain, we do not lose any optimality expect from from Opt 3 to Opt 2 and from Opt 2 to Opt 1.

This will yield the first statement in the lemma for c_0 large enough in terms of ϵ and $\frac{d}{n} \geq c_0$. If we make ϵ' small enough (in terms of κ), then we know from Theorem 3.2 that $\mathcal{L}_{\mathcal{D}}(f_W) \leq e^{-\frac{1}{c_1\sigma^2}}$ for some $c_1 = c_1(\kappa)$. This yields the second conclusion (as long as $c \geq c_1$), since the probability of classifying an example from $\psi(\mathcal{D})$ correctly is equal to the probability of classifying an example from $\mathcal{D}$ incorrectly (Claim D.22).

We proceed to analyze the properties of f_W to obtain the final conclusion.

Recall from Lemma D.15 that Since $\hat{\kappa} \leq \kappa < \kappa_{\text{uc}}$, the optimal solution to Opt 5 has $\phi(b) \leq \frac{1-q_1}{2^h}\phi(b+c)$ and $\phi(-b) \leq \frac{1-q_1}{2^h}\phi(-b+d)$ for some constant $q_1 = q_1(\kappa)$. Let γ_j be the margin $y_j f_W(x_j)$, and observe that by the symmetry of the backwards mapping γ_j is the same for all points j. We call this value γ.

Applying the four mappings above, in the instance $\mathcal{I}_1$, the variable $W = U + V$ satisfies for all $j \in \mathcal{P}$ and $i \in H_+$,

$$\phi(y_j w_i^T \mu_1) \leq \frac{1-q_1}{2^h}\phi(y_j w_i^T \mu_1 + w_i^T \xi_j), \tag{D.83}$$

and for all $j \in \mathcal{N}$ and $i \in H_-$,

$$\phi(y_j w_i^T \mu_2) \leq \frac{1-q_1}{2^h}\phi(y_j w_i^T \mu_2 + w_i^T \xi_j), \tag{D.84}$$

Further, by definition of the mapping ψ_{21}, for $i \in H_+$, we have $w_i^T \mu_2 = 0$ and $w_i^T \xi_j = 0$ for all $j \in \mathcal{N}$. Similarly, for $i \in H_-$, we have $w_i^T \mu_1 = 0$, and $w_i^T \xi_j = 0$ for all $j \in \mathcal{P}$.

Now we appeal to the fact that by Lemma B.2, for any values s, t, we have $\phi(s+t) \leq (\phi(s) + \phi(t))2^{h-1}$, and thus (repeating the argument in Equations D.50 and D.51 of Lemma D.18, which we omit the details of here) for all $j \in \mathcal{P}$ and taking expectation over $i \in H_+$,

$$\mathbb{E}_{i \in H_+}[\phi(v_i^T \xi_j)] \geq \left(\frac{1+q_1}{2^h}\right)\mathbb{E}_{i \in H_+}[\phi(w_i^T x_j)] = \left(\frac{1+q_1}{2^h}\right)(2\gamma). \tag{D.85}$$

Similarly for all $j \in \mathcal{N}$ and $i \in H_-$,

$$\mathbb{E}_{i \in H_+}[\phi(v_i^T \xi_j)] \geq \left(\frac{1+q_1}{2^h}\right)\mathbb{E}_{i \in H_+}[\phi(w_i^T x_j)] = \left(\frac{1+q_1}{2^h}\right)(2\gamma). \tag{D.86}$$

Finally, by inspecting the mapping in Lemma D.11, and the fact that all of the backwards mapping duplicate solutions to the simpler problems, we have the following symmetry property of W:

$$\mathbb{E}_{i \in H_+}[\phi(w_i^T \mu_1)] = \mathbb{E}_{i \in H_+}[\phi(-w_i^T \mu_1)] = \mathbb{E}_{i \in H_-}[\phi(w_i^T \mu_2)] = \mathbb{E}_{i \in H_-}[\phi(-w_i^T \mu_2)]. \tag{D.87}$$

We can now examine the margin on the flipped dataset $\psi(S)$. Without loss of generality, consider an example $\psi(x_j)$ where $j \in \mathcal{P}_1$, such that $\psi(x_j) = \mu_2 + \xi_j$.

$$\begin{aligned}
\psi(y_j) f_W(\psi(x_j)) &= \mathbb{E}_i[y_j a_i \phi(w_i^T \psi(x_j))] && \text{(D.88)}\\
&= \frac{1}{2}\mathbb{E}_{i \in H_+}[\phi(w_i^T \mu_2 + v_i^T \xi_j)] - \frac{1}{2}\mathbb{E}_{i \in H_-}[\phi(w_i^T \mu_2 + v_i^T \xi_j)] && \text{(D.89)}\\
&= \frac{1}{2}\mathbb{E}_{i \in H_+}[\phi(v_i^T \xi_j)] - \frac{1}{2}\mathbb{E}_{i \in H_-}[\phi(w_i^T \mu_2)] && \text{(D.90)}\\
&\geq \left(\frac{1+q_1}{2^h}\right)(\gamma) - \left(\frac{1-q_1}{2^h}\right)(\gamma) && \text{(D.91)}\\
&= \frac{q_1 \gamma}{2^{h-1}}. && \text{(D.92)}
\end{aligned}$$

Thus $\psi(y_j) f_W(\psi(x_j)) \geq \frac{q_1\gamma}{2^{h-1}} \geq (1-\epsilon')\frac{q_1 \gamma^*(S)}{2^{h-1}}$, and the conclusion follows by choosing $q = \frac{q_1}{2^h}$ since we have $\epsilon' \leq \frac{1}{2}$ for c_0 large enough. □

Proposition 3.6 (Polynomial Margin Bounds Fail for XOR on Neural Network)**.** *Fix an integer $p \geq 1$, and any $\epsilon > 0$. There exists $c = c(p, \epsilon)$ such that the following holds for any n, d, σ and $\mathcal{D} \in \Omega_{\sigma,d}^{\text{XOR}}$ with $\kappa_{gen}^{\text{XOR},h} + \epsilon < \kappa < \kappa_{uc}^{\text{XOR},h} - \epsilon$, $\frac{d}{n} \geq c$ and $n \geq c$. Let $\mathcal{H}$ be any hypothesis class such that for $\tilde{\mathcal{D}} \in \{\mathcal{D}, \psi(\mathcal{D})\}$,*

$$\Pr_{S \sim \tilde{\mathcal{D}}^n}[\textit{all } (1-\epsilon)\textit{-max-margin two-layer neural networks } f_W \textit{ for } S \textit{ lie in } \mathcal{H}] \geq 3/4.$$

Suppose that there exists an polynomial margin bound of degree p: that is, there is some G that satisfies for $\tilde{\mathcal{D}} \in \{\mathcal{D}, \psi(\mathcal{D})\}$,

$$\Pr_{S \sim \tilde{\mathcal{D}}^n}\left[\sup_{h \in \mathcal{H}} \mathcal{L}_{\tilde{\mathcal{D}}}(h) - \mathcal{L}_S(h) \geq \frac{G}{\gamma(h,S)^p}\right] \leq \frac{1}{4}.$$

Then with probability $\frac{1}{2} - 3e^{-n/c}$ over $S \sim \mathcal{D}^n$, on the max-margin solution, the generalization guarantee is no better than $\frac{1}{c}$, that is, $\frac{G}{\gamma^*(S)^p} \geq \frac{1}{c}$.

To prove Proposition 3.6 we use Lemma D.25.

Proof of Proposition 3.6. Let $c = 3c_0$, where $c_0 = c(\kappa, \epsilon)$ is the constant from Lemma D.25.

For any $\mathcal{D} \in \Omega$, let $T_{\mathcal{D}} \subset 2^{(\mathbb{R}^d \times \{-1,1\})^n}$ be the set of training sets S on which the conclusion of Lemma D.25 holds for $\mathcal{D}$ and S. Thus for any $\mathcal{D} \in \Omega$, $\Pr_{S\sim\mathcal{D}^n}[S \in T_{\mathcal{D}}] \geq 1 - 3e^{-n/c_0}$ for some constant c. Let $H \subset 2^{(\mathbb{R}^d \times \{-1,1\})^n}$ be the set of training sets S on which all $(1-\epsilon)$-max-margin two-layer neural networks f_W for S lie in $\mathcal{H}$. Thus for any $\mathcal{D} \in \Omega$, $\Pr_{S\sim\mathcal{D}^n}[S \in H] \geq \frac{3}{4}$.

For any $\mathcal{D} \in \Omega$, let $T'_{\mathcal{D}}$ be the set on which

$$\mathcal{L}_{\mathcal{D}}(h) \leq \mathcal{L}_S(h) + \frac{G}{\gamma(h,S)^p} \qquad \forall h \in \mathcal{H}. \tag{D.93}$$

By assumption, for any $\mathcal{D} \in \Omega$, $\Pr_{S\sim\mathcal{D}^n}[S \in T'_{\mathcal{D}}] \geq \frac{3}{4}$.

Now fix any $\mathcal{D} = \mathcal{D}_{\mu,\sigma,d} \in \Omega$. By a union bound, with $\psi = \psi_{\mathcal{D}}$,

$$\Pr_{S\sim\mathcal{D}^n}[S \in T'_{\mathcal{D}} \wedge \psi(S) \in T_{\psi(\mathcal{D})} \wedge \psi(S) \in H] \geq 1 - \left(1 - \frac{3}{4}\right) - 3e^{-n/c_0} - \left(1 - \frac{3}{4}\right) = \frac{1}{2} - 3e^{-n/c_0}. \tag{D.94}$$

This is because the distribution of $\psi(S)$ with $S \sim \mathcal{D}^n$ is the same as the distribution of n samples from $\psi(\mathcal{D})$.

Let S be any set for which the three events above hold, ie.,

$$S \in T'_{\mathcal{D}} \wedge \psi(S) \in T_{\psi(\mathcal{D})} \wedge \psi(S) \in H. \tag{D.95}$$

Let f_W be the classifier produced by Lemma D.25 on input $\psi(S)$ and distribution $\psi(\mathcal{D})$, such that:

1. $\gamma(f_W, \psi(S)) \geq (1-\epsilon)\gamma^*(\psi(S))$, and thus since $\psi(S) \in H$, we have $f_W \in \mathcal{H}$.
2. $\mathcal{L}_{\mathcal{D}}(f_W) \geq 1 - e^{-\frac{1}{c_0\sigma^2}}$.
3. $\gamma(f_W, S) \geq q\gamma^*(\psi(S))$ for some constant $q(\kappa)$.

It follows that for any such S, we must have

$$G \geq \left(1 - e^{-\frac{1}{c_0\sigma^2}}\right)\gamma(f_W, S)^p \geq \left(1 - e^{-\frac{1}{c_0\sigma^2}}\right)\gamma^*(\psi(S))^p q^p \tag{D.96}$$

Thus for the distribution $\psi(\mathcal{D})$, with probability at least $\frac{1}{2} - 3e^{-n/c_0}$, the margin bound yields a generalization guarantee no better than

$$\left(1 - e^{-\frac{1}{c_0\sigma^2}}\right) q^p. \tag{D.97}$$

Taking $c = \max(\frac{1}{q^p\left(1-e^{-\frac{1}{c_0\sigma^2}}\right)}, c_0, \frac{c_0}{\kappa})$ yields the proposition. Note that $e^{-\frac{1}{c_0\sigma^2}} = e^{-\frac{\kappa d}{c_0 n}}$, so for $\frac{d}{n} \geq \frac{c_0}{\kappa}$, $1 - e^{-\frac{1}{c_0\sigma^2}}$ is bounded away from 0 and thus c only depends on κ and δ (since c_0 additionally depends on c_0). □

Finally, we prove Proposition 3.3, which we restate.

Proposition 3.3 (Region where Max-Margin Generalization not Guaranteed)**.** *Let $h \in (1,2)$, and let $\epsilon > 0$. There exists a constant $c = c(\epsilon)$ such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega^{\mathrm{XOR}}_{\sigma,d}$ satisfying $\kappa \leq \kappa^{\mathrm{XOR},h}_{gen} - \epsilon$ and $\frac{d}{n} \geq c$, with probability $1 - 3e^{-n/c}$ over $S \sim \mathcal{D}^n$, there exists some W with $\|W\| = 1$ and $\gamma(f_W, S) \geq (1-\epsilon)\gamma^*(S)$ such that $\mathcal{L}_{\mathcal{D}}(f_W) = \frac{1}{2}$.*

Proof of Proposition 3.3. This follows directly from Lemma D.19, since for any $\mathcal{D}_{\mu_1,\mu_2,\sigma,d}$, any classifier f_W with $U = 0$ must have a test loss of exactly $\frac{1}{2}$. □

D.4 Proof of Technical Lemmas

Throughout the following section we assume $\mathcal{D}_{\mu_1,\mu_2,\sigma,d} \in \Omega$ is fixed, $h \in (1,2)$, and we use the same notation defined in the notation section at the beginning of Section D.1.

D.4.1 PROOF OF LEMMA A.2

We begin by proving Lemma A.2, for which we will need the following general analysis claim:

Claim D.26. *For any random variables a and b, with $\phi(x) = \max(0, x)^h$, we have*

$$|\mathbb{E}\left[\phi(a+b) - \phi(b)\right]| \leq \sqrt{\mathbb{E}\left[4a^2 + 2\right]}\sqrt{\mathbb{E}\left[b^2\right]} + h\left(\mathbb{E}[b^2]\right)^{\frac{h}{2}} \tag{D.98}$$

and

$$|\mathbb{E}\left[\phi(a+b) - \phi(b)\right]| \leq 2\mathbb{E}\left[1 + \phi(a)\right]\sqrt{\mathbb{E}\left[b^2\right]} + h\left(\mathbb{E}[b^2]\right)^{\frac{h}{2}}. \tag{D.99}$$

Proof. First note that for any a, b, we have:

$$|\phi(a+b) - \phi(a)| \leq \phi'(a+b)|b| \leq \left(\phi'(a) + \phi'(b)\right)|b| \leq \phi'(a)|b| + h|b|^h. \tag{D.100}$$

and

$$\phi'(a) =\leq 2|a| + 1, \tag{D.101}$$

$$\begin{aligned}
|\mathbb{E}\left[\phi(a+b) - \phi(b)\right]| &\leq \mathbb{E}\left[|\phi'(a)b|\right] + h\mathbb{E}[|b|^h] && \text{(D.102)}\\
&\leq \sqrt{\mathbb{E}\left[(\phi'(a))^2\right]}\sqrt{\mathbb{E}\left[b^2\right]} + h\mathbb{E}[|b|^h] && \text{(D.103)}\\
&\leq \sqrt{\mathbb{E}\left[(2|a|+1)^2\right]}\sqrt{\mathbb{E}\left[b^2\right]} + h\mathbb{E}[|b|^h] && \text{(D.104)}\\
&\leq \sqrt{\mathbb{E}\left[4a^2 + 2\right]}\sqrt{\mathbb{E}\left[b^2\right]} + h\mathbb{E}[|b|^h] && \text{(D.105)}\\
&\leq \sqrt{\mathbb{E}\left[4a^2 + 2\right]}\sqrt{\mathbb{E}\left[b^2\right]} + h\left(\mathbb{E}[b^2]\right)^{\frac{h}{2}}. && \text{(D.106)}
\end{aligned}$$

Here we used Equation D.100 in the first inequality, Cauchy-Schwartz in the second, Equation D.101 in the third, Jenson's in the fourth, and Jensen's again in the fifth inequality.

If instead of Equation D.100, we can obtain an alternative result.

$$\left(\phi'(a)\right)^2 = h^2\max(0, a)^{2h-2} \leq 4(1 + \phi(a)) \tag{D.107}$$

This yields

$$\begin{aligned}
|\mathbb{E}\left[\phi(a+b) - \phi(b)\right]| &\leq \sqrt{\mathbb{E}\left[(\phi'(a))^2\right]}\sqrt{\mathbb{E}\left[b^2\right]} + h\mathbb{E}[|b|^h] && \text{(D.108)}\\
&\leq 2\sqrt{\mathbb{E}\left[1 + \phi(a)\right]}\sqrt{\mathbb{E}\left[b^2\right]} + h\mathbb{E}[|b|^h] && \text{(D.109)}\\
&\leq 2\mathbb{E}\left[1 + \phi(a)\right]\sqrt{\mathbb{E}\left[b^2\right]} + h\left(\mathbb{E}[b^2]\right)^{\frac{h}{2}}. && \text{(D.110)}
\end{aligned}$$

□

We restate Lemma A.2 for the reader's convenience.

Lemma A.2. *Fix a distribution $\mathcal{D}_{\mu_1,\mu_2,\sigma,d} \in \Omega^{h,\mathrm{XOR}}_{\sigma,d}$. For $W \in \mathbb{R}^{m\times d}$, let $W = U + V$ where V is orthogonal to the subspace containing μ_1 and μ_2. Then for some universal constant c, for any $t \geq 1$, with probability at least $1 - e^{-ct}$, on a random sample $x \sim \mathcal{D}_{\mu_1,\mu_2,\sigma,d}$,*

$$|f_W(x) - f_U(x)| \leq (8\|U\| + 3)\,(t+1)\sigma^2\|V\|^2 + 2\left((t+1)\sigma^2\|V\|^2\right)^{\frac{h}{2}}.$$

Proof of Lemma A.2. We can write $x = z + \xi$ for where $z \in \mathrm{Span}(\mu_1, \mu_2)$ and $\xi \perp \mu_1, \mu_2$, such that by Claim D.26 we have

$$\begin{aligned}
|f_W(x) - f_U(x)| &= \left|\mathbb{E}_i\left[\phi(u_i^T z + v_i^T\xi) - \phi(u_i^T z)\right]\right| && \text{(D.111)}\\
&\leq \sqrt{\mathbb{E}_i\left[4(u_i^T z)^2 + 2\right]}\sqrt{\mathbb{E}_i\left[(v_i^T\xi)^2\right]} + h\left(\mathbb{E}_i[(v_i^T\xi)^2]\right)^{\frac{h}{2}} && \text{(D.112)}\\
&\leq \sqrt{\mathbb{E}_i\left[8\|u_i\|^2 + 2\right]}\sqrt{\mathbb{E}_i\left[(v_i^T\xi)^2\right]} + h\left(\mathbb{E}_i[(v_i^T\xi)^2]\right)^{\frac{h}{2}}, && \text{(D.113)}
\end{aligned}$$

where we have plugged $\|z\|_2 \leq \sqrt{2}$.

Now it suffices to get a high probability bound on $\mathbb{E}_i[(v_i^T\xi)^2] = \xi^T\mathbb{E}_i[v_i v_i^T]\xi$ for a random ξ. Let $M := \mathbb{E}_i[v_i v_i^T]$. We know by the Hanson-Wright Inequality that for some universal constant c,

$$\begin{aligned}
\Pr\left[\xi^T\mathbb{E}_i[v_i v_i^T]\xi \geq \sigma^2\,\mathrm{Tr}\,(M) + t\right] &\leq 2\exp\left(-c\min\left(\frac{t^2}{\|M\|_F^2}, \frac{t}{\|M\|_2}\right)\right) && \text{(D.114)}\\
&\leq 2\exp\left(-c\min\left(\frac{\sigma^2 t^2}{\mathrm{Tr}\,(M)^2}, \frac{\sigma t}{\mathrm{Tr}\,(M)}\right)\right), && \text{(D.115)}
\end{aligned}$$

where $\|\|\|_F$ denotes the Frobenius norm, and $\|\|\|_2$ denotes the spectral norm. Thus for $t \geq 1$,

$$\Pr\left[\xi^T \mathbb{E}_i[v_i v_i^T]\xi \geq (t+1)\sigma^2\|V\|^2\right] \leq 2\exp\left(-ct\right). \tag{D.116}$$

If follows that for any $t \geq 1$, with probability $1 - 2\exp(-ct)$,

$$|f_W(x) - f_U(x)| = (8\|U\| + 3)\,(t+1)\sigma^2\|V\|^2 + 2\left((t+1)\sigma^2\|V\|^2\right)^{\frac{h}{2}}. \tag{D.117}$$

□

D.4.2 Proof of Chaining Lemmas

Proof of Lemma D.8. To prove the lemma, we will begin with a $(1-\epsilon)$-solution W to Opt 1. Assuming toward a contradiction that items (1) or (2) in the lemma statement do not hold, we will construct a solution W'' for Opt 1 that is more than a $1/(1-\epsilon)$-factor times better than W, contradicting the $(1-\epsilon)$-optimality of W. We condition on the event that the conclusion of Lemma B.1 holds for Ξ. Given a solution W to Opt 1, construct a solution W' for Opt 1 as follows. First define $c_{ij} := w_i^T \xi_j$. For $i \in H_+$, let $w_i' = \mu_1\mu_1^T u_i + v_i'$, where v_i' is the min-norm vector such that $(v_i')^T\xi_j = c_{ij}$ for all $j \in \mathcal{P}$, and $(v_i')^T\xi_j = 0$ for all $j \in \mathcal{N}$. For $i \in H_-$, let $w_i' = \mu_2\mu_2^T u_i + v_i'$, where v_i' is the min-norm vector such that $(v_i')^T\xi_j = c_{ij}$ for all $j \in \mathcal{N}$, and $(v_i')^T\xi_j = 0$ for all $j \in \mathcal{P}$. Note that all such v_i' are guaranteed to exist since the conclusion of Lemma B.1 holds.

Let $s_i = \|\mu_1\mu_1^T u_i\|$ and $t_i = \|\mu_2\mu_2^T u_i\|$.

Observe that by Lemma B.1, we have:

$$\|w_i'\|^2 \leq s_i^2 + \|v_i'\|^2 \leq s_i^2 + \left(1 + C_{B.1}\sqrt{\frac{n}{d}}\right)\frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}} c_{ij}^2 \qquad \forall i \in H_+ \tag{D.118}$$

$$\|w_i'\|^2 \leq t_i^2 + \|v_i'\|^2 \leq t_i^2 + \left(1 + C_{B.1}\sqrt{\frac{n}{d}}\right)\frac{1}{d\sigma^2}\sum_{j\in\mathcal{N}} c_{ij}^2 \qquad \forall i \in H_- \tag{D.119}$$

$$\|w_i\|^2 \geq s_i^2 + t_i^2 + \|v_i\|^2 \geq s_i^2 + t_i^2 + \left(1 - C_{B.1}\sqrt{\frac{n}{d}}\right)\frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}\cup N} c_{ij}^2. \tag{D.120}$$

(D.121)

$$\mathbb{E}_i[\|w_i\|^2] \geq \left(1 - C_{B.1}\sqrt{\frac{n}{d}}\right)D + \frac{1}{1 + C_{B.1}\sqrt{\frac{n}{d}}}\mathbb{E}_i[\|w_i'\|^2], \tag{D.122}$$

where

$$D := \frac{1}{2}\mathbb{E}_{i\in H_+}\left[t_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{N}} c_{ij}^2\right] + \frac{1}{2}\mathbb{E}_{i\in H_-}\left[s_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}} c_{ij}^2\right]. \tag{D.123}$$

Further observe that:

$$\phi((w_i')^T x_j) = \phi(w_i^T x_j) \qquad \forall i : a_i > 0, j \in \mathcal{P} \tag{D.124}$$

$$\phi((w_i')^T x_j) = 0 \leq \phi(w_i^T x_j) \qquad \forall i : a_i < 0, j \in \mathcal{P} \tag{D.125}$$

$$\phi((w_i')^T x_j) = 0 \leq \phi(w_i^T x_j) \qquad \forall i : a_i > 0, j \in \mathcal{N} \tag{D.126}$$

$$\phi((w_i')^T x_j) = \phi(w_i^T x_j) \qquad \forall i : a_i < 0, j \in \mathcal{N}, \tag{D.127}$$

(D.128)

thus W' satisfies the constraint that $\mathbb{E}_i a_i\phi(w_i^T x_j)y_j \geq \gamma$. Indeed, we have by construction that have for all $j \in \mathcal{P}$ that

$$\sum_i a_i\phi(w_i^T x_i)y_j = \sum_{i\in H_+}\phi(w_i^T x_i) - \sum_{i\in H_-}\phi(w_i^T x_i) \tag{D.129}$$

$$\geq \sum_{i\in H_+}\phi((w_i')^T x_i) - \sum_{i\in H_-}\phi((w_i')^T x_i) \tag{D.130}$$

$$= \sum_i a_i\phi((w_i')^T x_i)y_j \tag{D.131}$$

and similarly for all $j \in \mathcal{N}$. If $D \geq 2C_{B.1}\sqrt{\frac{n}{d}} + 2\epsilon$, then

$$\mathbb{E}[\|w_i'\|^2] \leq \mathbb{E}[\|w_i'\|^2] - \left(1 - C_{B.1}\sqrt{\frac{n}{d}}\right) D \tag{D.132}$$

$$\leq 1 - \left(1 - C_{B.1}\sqrt{\frac{n}{d}}\right)\left(2\epsilon + 2C_{B.1}\sqrt{\frac{n}{d}}\right) \tag{D.133}$$

$$\leq 1 - 2\epsilon, \tag{D.134}$$

where we have used the global assumptions that $\epsilon \leq 1/4$ and $C_{B.1} \leq 1/2$. Thus we can scale W' up by a factor of $\frac{\left(\mathbb{E}_i[\|w_i\|^2]\right)^{\frac{1}{2}}}{\left(\mathbb{E}_i[\|w_i'\|^2]\right)^{\frac{1}{2}}}$ to achieve a feasible solution W'' that has objective value $\frac{1}{(1-2\epsilon)^{\frac{h}{2}}} \geq \frac{1}{(1-\epsilon)}$ times better than the solution given by W. This would contradict the $(1-\epsilon)$-optimality of W, proving the first conclusion of the lemma.

For the second part, suppose for greater than a $\sqrt{2C_{B.1}\sqrt{\frac{n}{d}} + 2\epsilon}$ fraction of data points we have $\frac{1}{2}\mathbb{E}_{i\in H_-}\left[(v_i^T\xi_j)^2\right] \geq \frac{1}{\kappa}\cdot\sqrt{2C_{B.1}\sqrt{\frac{n}{d}} + 2\epsilon}$ (if $j \in \mathcal{P}$) or $\frac{1}{2}\mathbb{E}_{i\in H_+}\left[(v_i^T\xi_j)^2\right] \geq \frac{1}{\kappa}\cdot\sqrt{2C_{B.1}\sqrt{\frac{n}{d}} + 2\epsilon}$ (if $j \in \mathcal{N}$) . This would imply that $D \geq \left(\frac{\sqrt{2C_{B.1}\sqrt{\frac{n}{d}}+2\epsilon}}{\kappa}\right)\frac{1}{d\sigma^2}\left(n\sqrt{2C_{B.1}\sqrt{\frac{n}{d}} + 2\epsilon}\right) = 2C_{B.1}\sqrt{\frac{n}{d}} + 2\epsilon$, which as we saw above contradicts the $(1-\epsilon)$-optimality of W.

□

Proof of Lemma D.13. By homogeneity, there exists some value C_B such that the optimum of any instance of Opt B with parameter P_B equals $C_B P_B^q$. Thus by the properties of ψ_{BA}, given an optimal instance $\mathcal{I}_B^* \in D_B$ with parameter P_B, we can construct and instance of Opt A with parameter at most $(1+\delta)P_B$ and optimum at least $C_B P_B^q$. Thus for some value $C_A \geq C_B(1+\delta)^{-q}$, the optimum of any instance of Opt A with parameter P_A equals $C_A P_A^q$.

Suppose $\mathcal{I}_A$ with parameter P_A is $(1-\epsilon)$-optimal and $\mathcal{I}_B^{(1)}, \cdots, \mathcal{I}_B^{(k)} := \psi_{AB}(\mathcal{I}_A)$. Let γ be the objective value of $\mathcal{I}_A$. Define $P_B^{(1)} \cdots P_B^{(p)}$ to be the parameters P_B of the k instances respectively, and let $\gamma^{(p)}$ be their objective values. For $p \in [k]$, let $s(p)$ be the optimality of each $\mathcal{I}_B^{(p)}$ times $\frac{\gamma}{\gamma^{(p)}}$. Then: $s(p) = \frac{\gamma^{(p)}}{C_B(P_B^{(p)})^q}\frac{\gamma}{\gamma^{(p)}}$, so

$$s(p)C_B(P_B^{(p)})^q \geq (1-\epsilon)C_A(P_A)^q \geq (1-\epsilon)(1+\delta)^{-q}C_B(P_A)^q \quad \forall p \tag{D.135}$$

$$\mathbb{E}_{p\in[k]}[P_B^{(p)}] \leq (1+\delta)P_A \tag{D.136}$$

Here the first inequality in the first line follows from the fact that the objective value achieved by $\mathcal{I}_B^{(p)}$ is at least as large as the objective value of $\mathcal{I}_A$, which by assumption is at lease $(1-\epsilon)$-optimal.

We now proceed by contradiction: Suppose for some set $S \subset [k]$ of size at least $k\epsilon'$, we have $s(p) \leq 1-\epsilon'$. Then

$$\mathbb{E}_{p\in[k]}[P_B^{(p)}] \geq \frac{1}{k}\sum_{p\in S} P_B^{(p)} + \frac{1}{k}\sum_{p\notin S} P_B^{(p)} \tag{D.137}$$

$$\geq \epsilon'\left((1-\epsilon')^{-\frac{1}{q}}(1-\epsilon)^{\frac{1}{q}}(1+\delta)^{-1}P_A\right) + (1-\epsilon')\left((1-\epsilon)^{\frac{1}{q}}(1+\delta)^{-1}P_A\right) \tag{D.138}$$

$$= P_A(1-\epsilon)^{\frac{1}{q}}(1+\delta)^{-1}\left(\epsilon'(1-\epsilon')^{-\frac{1}{q}} + (1-\epsilon')\right) \tag{D.139}$$

Thus if

$$\left(\epsilon'(1-\epsilon')^{-\frac{1}{q}} + (1-\epsilon')\right) > (1+\delta)^2(1-\epsilon)^{-\frac{1}{q}}, \tag{D.140}$$

we will have a contradiction, since the equation above will be strictly greater than $(1+\delta)P_A$.

Choosing $\epsilon' = \sqrt{1-(1-\epsilon)(1+\delta)^{-2q}}$, this produces the desired contradiction. Indeed, on can check that for all $\epsilon' \in (0,1)$, we have

$$\epsilon'(1-\epsilon')^{-\frac{1}{q}} + (1-\epsilon') > (1-(\epsilon')^2)^{-\frac{1}{q}}, \tag{D.141}$$

yielding the desired contradiction. Thus for at least a $1-\epsilon'$ fraction of $p \in [k]$, we have

$$\frac{\gamma}{\gamma^{(p)}} \times (\text{optimality of } \mathcal{I}_B^{(p)}) \geq 1-\epsilon',$$

which implies that each of these two terms are greater that $1 - \epsilon'$.

This proves the first conclusion.

To achieve the second conclusion, consider the mapping $\psi'_{AB} : D_B \to D_A$ which maps $\mathcal{I}_A$ to the instance of $\psi_{AB}(\mathcal{I}_A)$ which has the smallest parameter P_B. Necessarily, this value at most $(1+\delta)P_A$, since the average value of $P_B^{(p)}$ is at most $(1+\delta)P_A$. Thus the pair of mappings ψ_{BA} and ψ'_{AB} and ψ_B satisfy the conditions of the lemma, which we now apply with $k = 1$, and the roles of A and B reversed. The second conclusion follows. □

Proof of Lemma D.7. Recall that we have conditioned on the event that for any $c \in \mathbb{R}^n$, the min-norm vector v satisfying $\Xi^T v = c$ has $\|v\|_2^2 \in \frac{\|c\|_2^2}{\sigma^2 d}\left[\frac{1}{1+C_{B.1}\sqrt{\frac{n}{d}}}, \frac{1}{1-C_{B.1}\sqrt{\frac{n}{d}}}\right]$.

Observe that the mappings $\psi_{12}(\mathcal{I}_1)$ produces a feasible instance, since for all $i \in H_+$,

$$s_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}} c_{ij}^2 \le s_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}\cup\mathcal{N}} c_{ij}^2 \tag{D.142}$$

$$\le \|u_i\|^2 + \frac{1}{d\sigma^2}\left(1 + C_{B.1}\sqrt{\frac{n}{d}}\right)\sigma^2 d\|v_i\|^2 \le \left(1 + C_{B.1}\sqrt{\frac{n}{d}}\right)\|w_i\|^2. \tag{D.143}$$

A similar statement holds for $i \in H_-$, summing over $j \in \mathcal{N}$. Further, the objective value of $\psi_{12}(\mathcal{I}_1)$ is at least the objective value of $\mathcal{I}_1$.

The mapping ψ_{21} always maintains the exact same objective value, and is feasible because for $i \in H_+$, $\|v_i\|^2 \le \frac{1}{1-C_{B.1}\sqrt{\frac{n}{d}}}\frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}} c_{ij}^2$, and a similar statement holds for $i \in H_-$.

Thus applying Lemma D.13 twice (with Opt A = Opt 1 and Opt B = Opt 2, and then in reverse, and with $q = \frac{h}{2}$ and $1 + \delta = \frac{1}{1-C_{B.1}\sqrt{\frac{n}{d}}}$) yields the result.

□

Lemma D.9 (Opt 2 $\leftrightarrow$ Opt 3)**.** *Define the mapping $\psi_{23} : D_2 \to D_3 \times D_3$ as follows. Given input $\mathcal{I}_2$, output $\mathcal{I}_3^{(1)}$, $\mathcal{I}_3^{(2)}$, where for $\mathcal{I}_3^{(1)}$:*

- $H := H_+$
- $P_3 := \frac{1}{2}\mathbb{E}_{i\in H_+}\left(s_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}} c_{ij}^2\right)$
- *Let S_1 be an arbitrary set of $n_{\min}$ elements of $\mathcal{P}_1$, and let S_{-1} be an arbitrary set of $n_{\min}$ elements of $\mathcal{P}_{-1}$. Define c_{ij} to be the same as in $\mathcal{I}_2$ for all $j \in S_1 \cup S_{-1}$, $i \in H_+$.*
- *$b_i = s_i$ for $i \in H_+$,*

and for $\mathcal{I}_3^{(2)}$:

- $H := H_-$
- $P_3 := \mathbb{E}_{i\in H_-}\left(t_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{N}} c_{ij}^2\right)$
- *Let S_1 be an arbitrary set of $n_{\min}$ elements of $\mathcal{N}_1$, and let S_{-1} be an arbitrary set of $n_{\min}$ elements of $\mathcal{N}_{-1}$. Define c_{ij} to be the same as in $\mathcal{I}_2$ for all $j \in S_1 \cup S_{-1}$, $i \in H_-$.*
- *$b_i = t_i$ for $i \in H_-$,*

Define the mapping $\psi_{32} : D_3 \to D_2$ as follows. Given input $\mathcal{I}_3$, output $\mathcal{I}_2$, where

- $P_2 := P_3\left(\frac{n_{\max}}{n_{\min}}\right)$
- *Define an arbitrary bisections π_+ and π_- from H_+ and H_- respectively to H. Similarly define surjections $\rho_{+,1} : \mathcal{P}_1 \to S_1$, $\rho_{-,1} : \mathcal{P}_{-1} \to S_{-1}$, $\rho_{+,1} : \mathcal{N}_1 \to S_1$, $\rho_{-,1} : \mathcal{N}_{-1} \to S_{-1}$, such that for $x \in \{1,-1\}$, $\mathbb{E}_{j\in\mathcal{P}_x}\mathbb{E}_{i\in H} c^2_{i\rho_{+,x}(j)} \le \mathbb{E}_{j\in S_x}\mathbb{E}_{i\in H} c_{ij}^2$, and similarly, $\mathbb{E}_{j\in\mathcal{N}_x}\mathbb{E}_{i\in H} c^2_{i\rho_{-,x}(j)} \le \mathbb{E}_{j\in S_x}\mathbb{E}_{i\in H} c_{ij}^2$. For $i \in H_+$, define $s_i := b_{\pi_+(i)}$ and $c_{ij} := c_{\pi_+(i)\rho_{+,x}(j)}$ for all $x \in \{1,-1\}$ and $j \in \mathcal{P}_x$. For $i \in H_-$, define $t_i := b_{\pi_-(i)}$, and $c_{ij} := c_{\pi_-(i)\rho_{+,x}(j)}$ for all $x \in \{1,-1\}$ and $j \in \mathcal{N}_x$.*

Note that we here the c_{ij} variables we are defining *come belong to $\mathcal{I}_2$, and they are define in terms of the c_{ij} variables from $\mathcal{I}_3$.*

Then with $\epsilon' = \sqrt{1 - (1-\epsilon)\left(\frac{n_{\max}}{n_{\min}}\right)^{-h}}$

1. *If $\mathcal{I}_2 \in D_2$ is $(1-\epsilon)$-optimal, then each instance of $\psi_{23}(\mathcal{I}_2)$ is $(1-\epsilon')$-optimal on Opt 3, and has objective at most $\frac{1}{1-\epsilon'}$ times the objective of $\mathcal{I}_2$.*

2. *If $\mathcal{I}_3 \in D_3$ is $(1-\epsilon)$-optimal, then $\psi_{32}(\mathcal{I}_3)$ is $(1-\epsilon')$-optimal on Opt 2.*

Proof of Lemma D.9. It is easy to check by the definition of the mappings that if an instance $\mathcal{I}_2 \in D_2$ is feasible, then so is $\psi_{23}(\mathcal{I}_2)$. Likewise, if instance $\mathcal{I}_3 \in D_3$ is feasible, then so is $\psi_{32}(\mathcal{I}_3)$. Indeed, in $\psi_{32}(\mathcal{I}_3)$, we have

$$\frac{1}{2}\mathbb{E}_{i\in H_+}\left[s_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}}(c_{ij}^{(2)})^2\right] = \frac{1}{2}\mathbb{E}_{i\in H}\left[b_i^2 + \frac{n_{\max}}{n_{\min}}\frac{1}{d\sigma^2}\sum_{j\in S_1\cup S_{-1}}(c_{ij}^{(3)})^2\right] \leq \frac{n_{\max}}{n_{\min}}\frac{P_3}{2}, \quad \text{(D.144)}$$

where we have supersripted the variables c_{ij} in $\psi_{32}(\mathcal{I}_3)$ by (2), and those in $\mathcal{I}_3$ by (3). A similar statement holds for the sum over $\mathcal{N}$, such that

$$\frac{1}{2}\mathbb{E}_{i\in H_+}\left[s_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{P}}(c_{ij}^{(2)})^2\right] + \frac{1}{2}\mathbb{E}_{i\in H_-}\left[t_i^2 + \frac{1}{d\sigma^2}\sum_{j\in\mathcal{N}}(c_{ij}^{(2)})^2\right] \leq \frac{n_{\max}}{n_{\min}}P_3 = P_2. \quad \text{(D.145)}$$

It is easy to check also that the objective value of $\psi_{23}(\mathcal{I}_2)$ is at least that of $\mathcal{I}_2$, and likewise the objective value of $\psi_{32}(\mathcal{I}_3)$ is at least that of $\mathcal{I}_3$.

We can now apply Lemma D.13 with Opt A = Opt 2 and Opt B = Opt 3, $q = \frac{h}{2}$, $1+\delta = \frac{n_{\max}}{n_{\min}}$, and $k = 2$. This yields the result.

□

Proof of Lemma D.10. The proof is similar to the last lemma. It is straightforward to check the conditions of Lemma D.13 Opt A = Opt 3, Opt B = Opt 4, the mappings ψ_{34} and ψ_{43}, $k = \frac{n}{4}$, $q = \frac{h}{2}$, and $\delta = 0$. The conclusion follows from Lemma D.13.

□

Proof of Lemma D.11. We will eventually appeal to Lemma D.13. First observe that the mapping ψ_{54} yields an program in D_4 with the exact same objective value and parameter. We will construct an alternative mapping $\psi'_{45} : D_4 \to D_5$ that preserves the parameter and maintains or increases the objective. We use ψ_{45}. Let P_4 and γ be the parameter and objective of $\mathcal{I}_4$. First identify the instance $\mathcal{I}_5^{(i)}$ of $\psi_{45}(\mathcal{I}_4)$ which achieves the highest ratio between objective value, which we denote $\gamma^{(i)}$, and $P_5^{\frac{2}{h}}$. By the positivity of the of the $\gamma^{(i)}$ and $P_5^{(i)}$ and Jenson's inequality, for at least one instance i, we have

$$\frac{\gamma^{(i)}}{(P_5^{(i)})^{\frac{h}{2}}} \geq \frac{\mathbb{E}[\gamma^{(i)}]}{\mathbb{E}[(P_5^{(i)})^{\frac{h}{2}}]} \geq \frac{\mathbb{E}[\gamma^{(i)}]}{(\mathbb{E}[P_5^{(i)}])^{\frac{h}{2}}} \geq \frac{\gamma}{P_4^{\frac{h}{2}}}. \quad \text{(D.146)}$$

Then scaling each variable in this instance by a factor of $\sqrt{\frac{P_4}{P_5}}$ to produce an instance feasible instance of Opt 5 with parameter P_4 and objective value γ.

This suffices to apply Lemma D.13 with the mappings ψ_{54} and ψ'_{45} and $\delta = 0$. The second conclusion follows.

Now we prove the first conclusion. Let C_4 and C_5 be such that optimal value of Opt 4 equals $C_4 P^{\frac{h}{2}}$ and the optimal value of Opt 5 equals $C_5 P^{\frac{h}{2}}$. This holds by the homogeneity of the programs. The argument of Lemma D.13 in the paragraph beginning "We now proceed", applied using the mappings ψ_{54} and ψ'_{45} shows that $C_4 = C_5$.

Observe that

$$\mathbb{E}[\gamma^{(i)}] \leq \mathbb{E}[C_5 (P_5^{(i)})^{\frac{h}{2}}] \tag{D.147}$$

$$\leq \left(\mathbb{E}[C_5 (P_5^{(i)})]\right)^{\frac{h}{2}} \tag{D.148}$$

$$= C_5 P_4^{\frac{h}{2}} \tag{D.149}$$

$$= C_4 P_4^{\frac{h}{2}} \leq \frac{\gamma}{1-\epsilon}. \tag{D.150}$$

Here the first line follows from homogeneity of Opt 5, the second follows from Jenson's since $h < 2$, the third line from observing that the mapping ψ_{45} produces instances with an average parameter equal to P_4, and the fourth from the fact that $C_4 = C_5$ and that $\mathcal{I}_4$ is $(1-\epsilon)$-optimal.

This proves the first conclusion of the lemma. □

D.4.3 Proof of Lemmas analyzing Opt 5

We now prove the two lemmas analyze the trivariate program, Opt 5.

Proof of Lemma D.14. We consider two classes of feasible solutions. In the first, S_1, we impose the constraint that $-b + d > 0$. In the second, S_2, we have $-b + d \leq 0$.

For solutions in S_1, it is easy to check that for any (b, c, d), we can increase the objective value via the solution (b', c', d'), where $d' = 0$, and $c' = \sqrt{\frac{1-b^2}{k}} > c$.

The the optimum in S_1 is achieved by setting $d = 0$. It is then easy to check via the KKT conditions of the resulting convex program that the optimum in this set chooses b and c as in the claim.

We now consider the second set, S_2. Our goal will be to re-parameterize the objective in terms of a single variable $\alpha := \frac{c-d}{c+d}$, and then analyze the one-dimensional optimization landscape as a function of α. Recall that $\gamma_0 = \max_{c,d:kc^2+kd^2\leq 1} (\phi(b+c) + \phi(-b+d))$. Further define

$$\gamma_{bd} := \max_{b,c,d:0<b=d,kc^2+kd^2\leq 1} (\phi(b+c) + \phi(-b+d)) \tag{D.151}$$

We proceed in a series of claims.

The first claim reduces this 3 variable program to a 2 variable program.

Claim D.27.

$$\max_{b,c,d:0\leq b\leq d, b^2+k(c^2+d^2)\leq 1} \phi(b+c) + \phi(-b+d) \tag{D.152}$$

$$\leq \max\left(\gamma_0, \gamma_{bd}, \max_{c,d:k(c-d)\leq d\leq c, k^2(c-d)^2+k(c^2+d^2)\leq 1} \phi(k(c-d)+c) + \phi(-k(c-d)+d)\right) \tag{D.153}$$

$$\tag{D.154}$$

Proof. This claim reduces to showing that any locally optimal solution in S_2 that is not at one of the boundaries $b = 0$ or $b = d$ must satisfy $b = k(c-d)$. We proceed by contradiction. Suppose there was a feasible solution in S_2 with $0 < b < d$ which didn't satisfy $b = k(c-d)$. Then we can construct a new solution $b' = b + \Delta$, $c' = c - \Delta$, $d' = d + \Delta$. Then the objective value doesn't change ($\phi(b'+c') + \phi(-b'+d') = \phi(b+c) + \phi(-b+d)$), but for small enough Δ with the correct sign ($\text{sign}(-b + k(c-d))$) the constraint value decrease, since

$$(b')^2 + k\left((c')^2 + (d')^2\right) - b^2 - k(c^2+d^2) = 2\Delta\left(b - k(c-d)\right) + \Theta(\Delta^2) < 0. \tag{D.155}$$

Thus is we make this change and then scale up the solution such that the constraint is satisfied with equality, we will have increases the objective. Further, since b is bounded away from 0 and d, for Δ small enough, we will still have a point in S_2. Note, also introduce a constraint that $c \geq d$, since by the convexity of ϕ, we can switch the values of c and d and increase the objective if $c < d$. □

The next claim reduces the two-variable program to a single variable optimization problem.

Claim D.28.

$$\max_{c,d:k(c-d)\le d\le c, k^2(c-d)^2+k(c^2+d^2)\le 1} \phi(k(c-d)+c)+\phi(-k(c-d)+d) \tag{D.156}$$

$$\le \max\left(\gamma_0, \gamma_{bd}, \max_{0\le\alpha\le\frac{1}{2k+1}; f(\alpha)=0}\left(\frac{1}{\left(k^2+\frac{k}{2}\right)\alpha^2+\frac{k}{2}}\right)^{\frac{h}{2}}\left(\phi\left(\left(k+\frac{1}{2}\right)\alpha+\frac{1}{2}\right)+\phi\left(-\left(k+\frac{1}{2}\right)\alpha+\frac{1}{2}\right)\right)\right), \tag{D.157}$$

where

$$f(\alpha) := (1-\alpha)\,\phi'\left((2k+1)\,\alpha+1\right)-(1+\alpha)\,\phi'\left(-(2k+1)\,\alpha+1\right)=0. \tag{D.158}$$

Proof. First we reparameterize $A=c-d$ and $B=c+d$, such that we can upper bound by the optimum of the following program:

$$\max_{A,B} \quad \phi\left(\left(k+\frac{1}{2}\right)A+\frac{1}{2}B\right)+\phi\left(-\left(k+\frac{1}{2}\right)A+\frac{1}{2}B\right) \tag{D.159}$$

$$s.t. \quad \left(k^2+\frac{k}{2}\right)A^2+\frac{k}{2}B^2\le 1 \tag{D.160}$$

$$0\le A\le\frac{B}{2k+1} \tag{D.161}$$

Now by the KKT conditions, for any stationary point bounded away from the boundary of $A=0$ or $A=\frac{B}{2k+1}$, we must have $\frac{\frac{\partial g}{\partial A}}{\frac{\partial g}{\partial B}}=\frac{\frac{\partial f}{\partial A}}{\frac{\partial f}{\partial B}}$, where f and g represent the objective and the constraint respectively. Thus at these stationary points, we have

$$\frac{(2k^2+k)A}{kB}=\frac{k+\frac{1}{2}}{\frac{1}{2}}\frac{\phi'\left(\left(k+\frac{1}{2}\right)A+\frac{1}{2}B\right)-\phi'\left(-\left(k+\frac{1}{2}\right)A+\frac{1}{2}B\right)}{\phi'\left(\left(k+\frac{1}{2}\right)A+\frac{1}{2}B\right)+\phi'\left(-\left(k+\frac{1}{2}\right)A+\frac{1}{2}B\right)}, \tag{D.162}$$

or equivalently, setting $\alpha:=\frac{A}{B}$,

$$\frac{1+\alpha}{1-\alpha}=\frac{\phi'\left(\left(k+\frac{1}{2}\right)A+\frac{1}{2}B\right)}{\phi'\left(-\left(k+\frac{1}{2}\right)A+\frac{1}{2}B\right)}. \tag{D.163}$$

This manipulation and reparameterization in terms of α is useful for analysis because it allows us to leverage the homogeneity of ϕ without explicitly computing the KKT solution. Indeed, by homogeneity (and plugging in $\alpha=\frac{A}{B}$), we have at stationary point,

$$\frac{1+\alpha}{1-\alpha}=\frac{\phi'\left((2k+1)\,\alpha+1\right)}{\phi'\left(-(2k+1)\,\alpha+1\right)}, \tag{D.164}$$

or

$$f(\alpha) := (1-\alpha)\,\phi'\left((2k+1)\,\alpha+1\right)-(1+\alpha)\,\phi'\left(-(2k+1)\,\alpha+1\right)=0 \tag{D.165}$$

Now we check the boundary points. When $A=0$, this corresponds to the point where $c=d$ and $b=0$, which yields the objective value γ_0. When $A=\frac{B}{2k+1}$, this corresponds to the point when $b=d$, and thus yields the objective value γ_{bd}. □

In the next claim, we will show the single variable optimization program in terms of α achieves its maximum at the boundaries.

Claim D.29.

$$\max_{0\le\alpha\le\frac{1}{2k+1}; f(\alpha)=0}\left(\frac{1}{\left(k^2+\frac{k}{2}\right)\alpha^2+\frac{k}{2}}\right)^{\frac{h}{2}}\left(\phi\left(\left(k+\frac{1}{2}\right)\alpha+\frac{1}{2}\right)+\phi\left(-\left(k+\frac{1}{2}\right)\alpha+\frac{1}{2}\right)\right)\le\max\left(\gamma_0,\gamma_{bd}\right). \tag{D.166}$$

Proof. First we show that $f(\alpha)=0$ has at most one strictly positive solution. To show this, since we know $f(0)=0$, it suffices to check that the second derivative of $f(\alpha)$ is always positive for $0<\alpha\le 1$. Indeed, the second derivative evaluates to

$$(2-h)(h-1)t^2\left(\frac{1+\alpha}{(1-tx)^{3-h}}-\frac{1-\alpha}{(1+tx)^{3-h}}\right)+2(h-1)t\left(\frac{1}{(1-tx)^{2-h}}-\frac{1}{(1+tx)^{2-h}}\right). \tag{D.167}$$

where $t := (2k+1)$. Since $h \in (1,2)$ and $t \geq 0$, this expression is positive for $\alpha > 0$ since $\frac{1+\alpha}{(1-tx)^{3-h}} > \frac{1-\alpha}{(1+tx)^{3-h}}$ and $\frac{1}{(1-tx)^{2-h}} > \frac{1}{(1+tx)^{2-h}}$.

Now, we will show the derivative of the objective, which we will call $g(\alpha)$, is positive at the boundary point $\alpha = \frac{1}{2k+1}$. At this value, the term inside the second ϕ evaluates to 0, and thus so does its derivative. The remaining part of the objective evaluates to

$$\left(\frac{1}{\left(k^2+\frac{k}{2}\right)\alpha^2+\frac{k}{2}}\right)^{\frac{h}{2}} \phi\left(\left(k+\frac{1}{2}\right)\alpha+\frac{1}{2}\right) = \frac{1}{k^{\frac{h}{2}}}\left(\left(\left(k+\frac{1}{2}\right)\alpha+\frac{1}{2}\right)^2\left(\left(k+\frac{1}{2}\right)\alpha^2+\frac{1}{2}\right)^{-1}\right)^{\frac{h}{2}}, \tag{D.168}$$

so it suffices to check that

$$\left(\left(k+\frac{1}{2}\right)\alpha+\frac{1}{2}\right)^2\left(\left(k+\frac{1}{2}\right)\alpha^2+\frac{1}{2}\right)^{-1} \tag{D.169}$$

is increasing as a function of α. Indeed we can take the derivative to confirm this is the case for $\alpha \leq 1$.

Now since $g(\alpha)$ is increasing at the upper boundary $\alpha = \frac{1}{2k+1}$ and has at most one stationary point between 0 and the upper boundary, we conclude that this stationary point cannot be a maximum. Thus the maximum must be obtained at the boundary. Again the boundary points correspond to when $c = d$ and $b = 0$, yielding γ_0 and when $b = d$, yielding γ_{bd}. □

These three claims have shown that the maximum inside S_2 is obtained at one of the boundaries where $b = 0$ or $b = d$. It is easy to check that any solution when $b = d$ is suboptimal if $d > 0$, since we can decrease d and increase c by a small amount which will improve the objective. Now if $\hat{\kappa} > \kappa_{\text{gen}}$, then by definition, $\gamma_0 < \gamma_*$, and thus the optimal solution with $b \geq 0$ is given by choosing c and d as in the lemma.

If $\hat{\kappa} < \kappa_{\text{gen}}$, the greater solution of γ_0 and γ_* is given by γ_0, and thus we choose c and d to be equal as in the lemma. □

Proof of Lemma D.15. Because of the homogeneity of each constraint, it suffices to prove the result for $B_4 = 1$. Consider some ϵ-optimal solution (b, c, d). Without loss of generality, by the symmetry of the problem and the conclusion, we can assume b is non-negative.

Observe that for $\epsilon = \epsilon(\hat{\kappa})$ small enough, by the continuity of the objective, any $(1-\epsilon)$-optimal solution must be arbitrarily close to the solution given in Lemma D.14, which we name (b^*, c^*, d^*).

This means that for $\hat{\kappa} > \kappa_{\text{gen}}$, we must have b arbitrarily close to $\sqrt{\frac{\hat{\kappa}}{4+\hat{\kappa}}}$, c arbitrarily close to $\sqrt{\frac{16}{\hat{\kappa}(4+\hat{\kappa})}}$,and d arbitrarily close to 0. Thus the first conclusion follows by the fact that

$$\lim_{\epsilon \to 0} \frac{\phi(b)}{\phi(b+c)} = \frac{\phi(b^*)}{\phi(b^*+c^*)} = \left(\frac{1}{1+\frac{4}{\hat{\kappa}}}\right)^h. \tag{D.170}$$

The second one follows from the fact that in a neighborhood of b^* and d^*, both sides are 0.

If additionally $\hat{\kappa} < \kappa_{\text{uc}}$, then by definition of κ_{uc}, we have that $b^* < c^*$. So for small enough ϵ, $b < c$. Thus $\phi(b) = \frac{1}{2^h}\phi(2b) < \frac{1}{2^h}\phi(b+c)$, which yields the third conclusion. Again the fourth conclusion follows from the fact that in a neighborhood of b^* and d^*, both sides are 0.

The last conclusion (if $\hat{\kappa} < \kappa_{\text{gen}}$) follows immediately from the Lemma D.14. □

Proof of Lemma D.16. For each $i \in H$, define an instance of Opt 5 by putting $b = b_i$, $c = c_i$, $d = d_i$, and $P_5 = P_5^{(i)} := b_i^2 + \kappa(c_i^2 + d_i^2)$. Let γ_i denote the objective value of this instance of Opt 5, that is, $\phi(b_i + c_i) + \phi(-b_i + d_i)$.

By the homogeneity of Opt 5, the optimum of Opt 5 equals $C_4 P_5^{\frac{h}{2}}$ for some value C_4. Further, ClaimD.11 guarantees that the optimum of Opt 4 is at least $\frac{1}{2}C_4(P_4)^{\frac{h}{2}}$, because it is possible to construct a solution to Opt 4 from an optimal solution (b^*, c^*, d^*) of Opt 5 in the following way: For half the $i \in H$, take $(b_i, c_i, d_i) = (b^*, c^*, d^*)$. For the other half, take $(b_i, c_i, d_i) = (-b^*, d^*, c^*)$. Then the objective value is exactly half of the optimum of the optimum of Opt 5 with $P_5 = P_4$.

For $i \in H$, let $s(i)$ be the optimality of the respective instance of Opt 5, that is, $\frac{\gamma_i}{C_4\left(P_5^{(i)}\right)^{\frac{h}{2}}}$. If the solution to Opt 4 is $(1-\epsilon)$-suboptimal, then we have

$$\mathbb{E}_i s(i) C_4\left(P_5^{(i)}\right)^{\frac{h}{2}} \geq (1-\epsilon)\, C_4 (P_4)^{\frac{h}{2}}; \tag{D.171}$$

$$\mathbb{E} P_5^{(i)} \leq P_4. \tag{D.172}$$

Plugging the second equation into the first, and applying Jensen's inequality to the concavity of the function $x \to x^{h/2}$, we obtain

$$\mathbb{E}_i s(i) \left(P_5^{(i)}\right)^{\frac{h}{2}} \geq (1-\epsilon)\, (\mathbb{E} P_5^{(i)})^{\frac{h}{2}} \geq (1-\epsilon)\, \mathbb{E} (P_5^{(i)})^{\frac{h}{2}}. \tag{D.173}$$

Let $\alpha(i) := \left(P_5^{(i)}\right)^{\frac{h}{2}}$. In the remainder of the lemma, we use the α and s to denote random variables over the randomness of i, and all expectation are over i uniformly from H.

For any δ, we have,

$$\begin{aligned}
(1-\epsilon)\mathbb{E}[\alpha] &\leq \mathbb{E}[\alpha s] && \text{(D.174)}\\
&= \mathbb{E}[\alpha s \mathbb{1}(s \geq 1-\delta)] + \mathbb{E}[\alpha s \mathbb{1}(s < 1-\delta)] && \text{(D.175)}\\
&\leq \mathbb{E}[\alpha \mathbb{1}(s \geq 1-\delta)] + (1-\delta)\mathbb{E}[\alpha \mathbb{1}(s < 1-\delta)] && \text{(D.176)}\\
&= \mathbb{E}[\alpha \mathbb{1}(s \geq 1-\delta)] + (1-\delta)\left(\mathbb{E}[\alpha] - \mathbb{E}[\alpha \mathbb{1}(s \geq 1-\delta)]\right) && \text{(D.177)}\\
&= \delta \mathbb{E}[\alpha \mathbb{1}(s \geq 1-\delta)] + (1-\delta)\mathbb{E}[\alpha], && \text{(D.178)}
\end{aligned}$$

so

$$\mathbb{E}[\alpha \mathbb{1}(s \geq 1-\delta)] \geq \left(1 - \frac{\epsilon}{\delta}\right) \mathbb{E}[\alpha], \tag{D.179}$$

and hence,

$$\mathbb{E}[s\alpha \mathbb{1}(s \geq 1-\delta)] \geq (1-\delta)\left(1 - \frac{\epsilon}{\delta}\right) \mathbb{E}[\alpha]. \tag{D.180}$$

Let $\epsilon_{D.15}$ and η be the constants ϵ and η from Lemma D.15. By Lemma D.15, for any i with $s(i) \geq 1 - \epsilon_{D.15}$, we have $\phi(b_i) \geq \eta\phi(b_i + c_i)$. Thus

$$\begin{aligned}
\mathbb{E}_{i\in H}[\phi(b_i)] &\geq \mathbb{E}_{i\in H}[\phi(b_i)\mathbb{1}(s(i) \geq \epsilon_{D.15})] && \text{(D.181)}\\
&\geq \eta \mathbb{E}_{i\in H}[\phi(b_i + c_i)\mathbb{1}(s(i) \geq 1 - \epsilon_{D.15})] && \text{(D.182)}\\
&= \eta\left(\mathbb{E}_{i\in H}[\phi(b_i + c_i)] - \mathbb{E}_{i\in H}[\phi(b_i + c_i)\mathbb{1}(s(i) < 1 - \epsilon_{D.15})]\right) && \text{(D.183)}\\
&\geq \eta\left(\mathbb{E}_{i\in H}[\phi(b_i + c_i)] - \mathbb{E}_{i\in H}[(\phi(b_i + c_i) + \phi(-b_i + d_i))\, \mathbb{1}(s(i) < 1 - \epsilon_{D.15})]\right) && \text{(D.184)}\\
&= \eta\left(\mathbb{E}_{i\in H}[\phi(b_i + c_i)] - \mathbb{E}[\alpha s \mathbb{1}(s < 1 - \epsilon_{D.15})]\right) && \text{(D.185)}\\
&\geq \eta\left(\mathbb{E}_{i\in H}[\phi(b_i + c_i)] - \mathbb{E}[\alpha \mathbb{1}(s < 1 - \epsilon_{D.15})]\right) && \text{(D.186)}\\
&\geq \eta\left(\mathbb{E}_{i\in H}[\phi(b_i + c_i)] - \frac{\epsilon}{\epsilon_{D.15}}\mathbb{E}[\alpha]\right) && \text{(D.187)}
\end{aligned}$$

We need one more claim:

Claim D.30. *If the solution to Opt 4 is $(1-\epsilon)$-optimal, then*

$$\mathbb{E}_{i\in H}[\phi(b_i + c_i)] \geq \frac{1}{2}(1-\epsilon)\, \mathbb{E}_{i\in H}[\phi(b_i + c_i) + \phi(-b_i + d_i)]. \tag{D.188}$$

$$\mathbb{E}_{i\in H}[\phi(-b_i + d_i)] \geq \frac{1}{2}(1-\epsilon)\, \mathbb{E}_{i\in H}[\phi(b_i + c_i) + \phi(-b_i + d_i)]. \tag{D.189}$$

Proof. Suppose without loss of generality that $\mathbb{E}_{i\in H}[\phi(b_i + c_i)] = q\left(\mathbb{E}_{i\in H}[\phi(b_i + c_i) + \phi(-b_i + d_i)]\right)$ for some $q \leq \frac{1}{2}$.

Then the optimum of the program is at most q, so we have

$$qC_4(2P_4)^{\frac{h}{2}} \geq q\mathbb{E}_{i\in H}[\phi(b_i + c_i) + \phi(-b_i + d_i)] \geq (1-\epsilon)\frac{1}{2}C_4(2P_4)^{\frac{h}{2}}. \tag{D.190}$$

The conclusion follows.

□

Using the claim and Equation D.173,

$$\mathbb{E}_{i\in H}[\phi(b_i+c_i)] \geq \frac{1}{2}(1-\epsilon)\,\mathbb{E}_{i\in H}[\phi(b_i+c_i)+\phi(-b_i+d_i)] = \frac{1}{2}(1-\epsilon)\,\mathbb{E}[\alpha s] \geq \frac{1}{2}(1-\epsilon)^2\,\mathbb{E}[\alpha]. \tag{D.191}$$

Thus plugging this into the Equation D.181, we have

$$\mathbb{E}_{i\in H}[\phi(b_i)] \geq \eta\left(\mathbb{E}_{i\in H}[\phi(b_i+c_i)] - \frac{\epsilon}{\epsilon_{D.15}}\mathbb{E}[\alpha]\right) \tag{D.192}$$

$$\geq \eta\left(\mathbb{E}_{i\in H}[\phi(b_i+c_i)]\right)\left(1-\frac{2\epsilon}{\epsilon_{D.15}(1-\epsilon)^2}\right). \tag{D.193}$$

The second statement of the lemma can be proved identically, but using the second result of Lemma D.15.

Now we consider the case when additionally we have $\kappa < \kappa_{\text{uc}}$.

We can bound

$$\mathbb{E}_{i\in H}[\phi(b_i)] \leq \mathbb{E}_{i\in H}[\phi(b_i)\mathbb{1}(s(i)\geq 1-\epsilon_{D.15})] + \mathbb{E}_{i\in H}[\phi(b_i)\mathbb{1}(s(i)<1-\epsilon_{D.15})] \tag{D.194}$$

$$\leq \left(\frac{1+q(\kappa)}{2}\right)\mathbb{E}_{i\in H}[\phi(b_i+c_i)\mathbb{1}(s(i)\geq 1-\epsilon_{D.15})] + \mathbb{E}_{i\in H}[\alpha(i)\mathbb{1}(s(i)<1-\epsilon_{D.15})] \tag{D.195}$$

$$\leq \left(\frac{1+q(\kappa)}{2}\right)\mathbb{E}_{i\in H}[\phi(b_i+c_i)] + \frac{\epsilon}{\epsilon_{D.15}}\mathbb{E}[\alpha] \tag{D.196}$$

$$\leq \left(\frac{1+q(\kappa)/2}{2}\right)\mathbb{E}_{i\in H}[\phi(b_i+c_i)] \tag{D.197}$$

for ϵ a small enough constant. Here in the second inequality we used Lemma D.15 and additionally the fact that for any i, we have $\phi(b_i)\leq\alpha(i)$ in any feasible solution. In the third inequality, we used Equation D.179. In the final inequality, we used Equation D.191 and chose ϵ small enough in terms of $q(\kappa)$ and $\epsilon_{D.15}$.

The same argument holds for d_i and $-b_i$. □