Optimization-Based Separations for Neural Networks

Itay Safran   ISAFRAN@PRINCETON.EDU
Jason D. Lee   JASONLEE@PRINCETON.EDU
Princeton University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Depth separation results propose a possible theoretical explanation for the benefits of deep neural networks over shallower architectures, establishing that the former possess superior approximation capabilities. However, there are no known results in which the deeper architecture leverages this advantage into a provable optimization guarantee. We prove that when the data are generated by a distribution with radial symmetry which satisfies some mild assumptions, gradient descent can efficiently learn ball indicator functions using a depth 2 neural network with two layers of sigmoidal activations, and where the hidden layer is held fixed throughout training. By building on and refining existing techniques for approximation lower bounds of neural networks with a single layer of non-linearities, we show that there are d-dimensional radial distributions on the data such that ball indicators cannot be learned efficiently by any algorithm to accuracy better than $\Omega(d^{-4})$, nor by a standard gradient descent implementation to accuracy better than a constant. These results establish what is to the best of our knowledge, the first optimization-based separations where the approximation benefits of the stronger architecture provably manifest in practice. Our proof technique introduces new tools and ideas that may be of independent interest in the theoretical study of both the approximation and optimization of neural networks.

1. Introduction

In recent years, several theoretical papers have provided a possible explanation for the benefits of using deep neural networks over shallower architectures, by proving *depth separation* results (Eldan and Shamir, 2016; Telgarsky, 2016; Poole et al., 2016; Daniely, 2017; Yarotsky, 2017; Liang and Srikant, 2017; Safran and Shamir, 2017; Poggio et al., 2017; Safran et al., 2019; Venturi et al., 2021). Simply put, these results show the existence of some function that can be approximated efficiently by a neural network with a certain depth, whereas a shallower architecture might require exponentially many more neurons to achieve the same accuracy. Such results may suggest that the reason that deeper networks perform better in practice stems from their superior *approximation* capabilities. In contrast, the network configuration that is achieved in practice when a neural network is trained is strongly dictated by the learning algorithm used in its *optimization*. While it is clear that the ability of a certain architecture to approximate a function efficiently is a necessary condition for being able to learn it successfully, it is not obvious to what extent - if at all - this is also a sufficient condition. Indeed, several recent works have shown that in some cases, the same 'extreme' properties that give rise to the inapproximability of these hard functions when using shallower architectures may also prove detrimental for optimization using the deeper architecture (Malach and Shalev-Shwartz, 2019; Malach et al., 2021). This suggests that certain results used to demonstrate the benefits of depth such as the one shown in Telgarsky (2016) are in a sense 'too strong', and that arguably, one of the main incentives for studying the approximation power of neural networks (namely, its necessity

for proving a positive optimization result) is absent in such cases. Moreover, if inapproximability using a shallow architecture may in fact turn out to imply the inability to optimize efficiently using the deep architecture in a more general setting, then this would suggest that the approximation capabilities of a neural network have little impact on its optimization, perhaps further demotivating the study of neural network approximation from a practical perspective.

In light of the above, to theoretically explain the practical success of depth via the lens of function approximation, we would need to show that the benefits of depth will also manifest in practice. While a learning algorithm can simply deterministically return the efficient approximation of the hard function by utilizing depth for a specific problem setting, this is unsatisfactory; such a setting would not explain why deeper neural networks achieve better results in practice, since these are typically trained using first-order optimization methods such as gradient descent (GD) or its more sophisticated variants. We are now lead to the following question:

> *Can we show an optimization-based separation result, in which a simple architecture cannot approximate efficiently (implying the hardness of learning with such architecture), yet a stronger architecture can provably learn with a polynomial sample size, network size and running time, using a first-order optimization method?*

In this paper, we answer this question in the affirmative by proving that a neural network with a single hidden layer with sigmoidal activations and an additional non-linearity in the output neuron can learn ball indicator functions efficiently (functions of the form $\mathbf{x} \mapsto 1$ if and only if $\|\mathbf{x}\|_2 \leq \lambda$ for some $\lambda > 0$). In contrast, by using a reduction technique of Safran and Shamir (2017), we show that ball indicators cannot be approximated efficiently to accuracy better than $\Omega(d^{-4})$ using depth 2 neural networks when we are not allowed to use an additional non-linearity in the output neuron. Moreover, using a technique introduced in Daniely (2017), we provide new lower bounds for approximating ball indicators, showing that getting better than constant accuracy requires that either the width or the Euclidean norm of the weights of the network grow exponentially in either d or $\frac{1}{\sqrt{\epsilon}}$ (ϵ being the desired accuracy). By bounding the progress of GD in a standard setting, this leads to a similar exponential lower bound on its running time. Overall, these establish what is to the best of our knowledge, the first separation results where the benefits of the stronger class of networks for function approximation manifests in practice and facilitates efficient optimization.

The remainder of this paper is structured as follows: After presenting our contributions in more detail below, we turn to discuss related work in Subsection 1.2. In Sec. 2, we first introduce relevant definitions and notation before we present our setting and problem formulation in Subsection 2.1. Thereafter, our main results are presented in Subsection 2.2, followed by a formal definition of the distributions they use in Subsection 2.3. Sec. 3 contains our approximation lower bounds for ball indicator functions that our main results build upon. In Sec. 4, we turn to show a positive optimization result for learning ball indicators.

1.1. Our Contributions

- We show that there exist a sequence $\{\mathcal{D}_d\}_{d=2}^{\infty}$ of d-dimensional heavy-tailed distributions and a sequence of constants $\{\lambda_d\}_{d=2}^{\infty}$ where $\lambda_d \in [1, 2]$ for all d, such that no neural network of depth 2 with a single layer of non-linearities and width less than $\Omega\left(\exp(\Omega(d))\right)$ can approximate ball indicator functions with radii $\{\lambda_d\}_{d=2}^{\infty}$ to accuracy better than $\Omega(d^{-4})$ (Thm. 4).

- In a slightly different approximation setting, we show that there exist a sequence $\{\mathcal{D}'_d\}_{d=2}^{\infty}$ of d-dimensional compactly-supported distributions and a sequence of constants $\{\lambda_d\}_{d=2}^{\infty}$ where $\lambda_d \in [1, 2]$ for all d, such that any neural neural network of depth 2 with a single layer of non-linearities which achieves accuracy $\mathcal{O}(m^{-2})$ for natural $m \geq 1$ on ball indicators with radii $\{\lambda_d\}_{d=2}^{\infty}$ must have width or weight norm exponential in either d or m (Thm. 5).

- We prove that under certain assumptions on the distribution of the data (Assumption 6), a neural network with two layers of non-linear activations (see Eq. (2) for the formal architecture) can efficiently learn data labeled by a ball indicator function with radius in $[1, 2]$, and attain arbitrarily small population loss by using GD with a standard initialization (Assumption 7), where the hidden layer is held fixed throughout training (Thm. 8).

- Welding our lower bounds together with the positive optimization result, we derive our main theorems (Thm. 2 and Thm. 3) which are optimization-based separation results. We show that no algorithm returning a depth 2 network with a single layer of non-linearities can efficiently learn ball indicators with respect to $\{\mathcal{D}_d\}_{d=2}^{\infty}$ to accuracy better than $\Omega(d^{-4})$, whereas optimizing using a standard GD setting on a depth 2 network cannot efficiently learn ball indicators with respect to $\{\mathcal{D}'_d\}_{d=2}^{\infty}$ to more than constant accuracy with overwhelming probability. In contrast, optimizing using a standard GD setting on our stronger architecture which employs more than a single layer of non-linear activations will succeed in learning with respect to both distributions with high probability.

Next, we turn to discuss and compare related work in the literature that is most relevant to ours.

1.2. Related Work

Depth Separations in Neural Networks. For concreteness, let us focus here on works that separate depth 2 from 3 where the target function and the domain of approximation are continuous. In their seminal work, Eldan and Shamir (2016) provided the first such depth separation. The authors show that a certain radial function consisting of a superposition of polynomially many thin shells cannot be approximated to accuracy better than an absolute constant using a depth 2 network, unless the width is exponential in the input dimension. Building on their work, Safran and Shamir (2017) provide a more natural separating target function by reducing the more complicated hard function to a ball indicator. Additionally, the authors also demonstrate empirically that depth 3 networks outperform depth 2 networks in learning ball indicators, even if the shallow network has many more trainable weights. While we build on a similar proof strategy as in the aforementioned paper, our work differs from theirs in that they show that for a ball indicator function of any radius, there exists a distribution under which it is hard to approximate using depth 2 networks, whereas our work shows that for a specific distribution there exists a ball indicator function with radius in the interval $[1, 2]$ which is hard to approximate using a depth 2 network. Moreover, our main result rigorously proves the learnability of the target function used in their experiment. An additional similar separation is given by Daniely (2017), where the approximation is in a compact domain and the target function oscillates polynomially many times in the input dimension, and where an exponential bound on the magnitude of the weights of the depth 2 network is imposed. Venturi et al. (2021) provide separation results with respect to product distributions on the data and target functions that do not necessarily possess radial symmetry.

Unlearnability of Depth Separations. As discussed earlier in the introduction, there are known cases where the same property that gives rise to a negative approximation result for shallow networks is also detrimental for optimization using the deeper architecture. Malach and Shalev-Shwartz (2019) have shown a depth separation result on fractal-like distributions with a natural coarse-to-fine structure that prompts the use of depth to achieve efficient approximation. On the other hand, they show that such distributions are intrinsically difficult to learn using gradient methods since the gradient will be exponentially close to zero with overwhelming probability, even if the architecture being optimized is able to achieve zero loss on the distribution. Building on a similar technique, Malach et al. (2021) show that if a depth 3 neural network cannot approximate a target function efficiently, then a deeper architecture will also be prone to failure if optimized using gradient methods. As a specific corollary, this implies that the depth separation result shown in Telgarsky (2016), which exploits the ability of deep ReLU networks to compute piece-wise linear functions with a large number of segments, also results in a function that is hard to optimize over using gradient methods. In Vardi et al. (2021a), the authors show that there is a natural proof barrier for proving depth separations between networks of depth 4 and deeper architectures, when the target functions have 'benign' properties, such as polynomially-bounded Lipschitz constant and being able to be computed efficiently. This, in a certain sense, suggests the opposite implication of Malach and Shalev-Shwartz (2019); Malach et al. (2021); that it might be difficult to show depth separations with functions that have some properties that are desired in practice to ensure learnability.

Learning a Single Neuron. In our positive optimization result, by holding the weights of the hidden layer fixed throughout training, we essentially reduce our problem to that of learning with a single neuron using gradient methods. This setting has been studied in several recent works. Yehudai and Shamir (2020); Vardi et al. (2021b) show the convergence of GD and some of its variants on the population loss objective, with constant or high probability and with or without a bias term, under various distributional assumptions on the data and in the realizable setting. In contrast, our work deals with the non-realizable case, assumes a ball indicator target function, and provides a convergence guarantee with high rather than constant probability for radial distributions. Frei et al. (2020) provide an agnostic (in the non-realizable setting) analysis for learning with finite data while making few assumptions on the problem. However, their result does not imply linear convergence of GD as opposed to ours, and does not guarantee attaining loss which is arbitrarily close to the global minimum.

Neural Network Approximation Using Random Features. Many works used the technique of random features to approximate various target functions, which is potentially used to derive positive optimization results for neural networks (Bach, 2017; Ji et al., 2019; Bai and Lee, 2019; Allen-Zhu et al., 2019; Yehudai and Shamir, 2019; Chen et al., 2020; Ghorbani et al., 2021). A major difference that sets our work apart from these is that other works perform their approximation on a bounded domain, whereas we show the convergence of random features for heavy-tailed distributions that generate data with a very large norm. On the downside, we merely show the approximation of ball indicator target functions, whereas the works cited above show approximation of a family of polynomials or even more general target function classes.

Learning More than a Single Layer of Non-linearities. The problem of learning a neural network which has more than a single layer of non-linearities has been addressed by several papers in recent years. In Goel and Klivans (2019), the authors study a similar architecture to the one we use to show a positive optimization result, where an additional non-linear activation is employed on the output neuron. In Allen-Zhu et al. (2019), the authors prove that a certain class of depth

3 target functions can be improperly learned using depth 3 networks with a different architecture. Allen-Zhu and Li (2020) show that deep networks with quadratic activations can heirarchically learn certain concept classes. However, the learning algorithm used in Goel and Klivans (2019) is different than GD, and all three papers crucially rely on the input having unit norm or some other form of boundedness, whereas our work accommodates for certain inputs whose norm distribution does not have finite moments. In contrast, our paper merely establishes convergence for the target class of ball indicator functions, and not for a broader family of target functions analyzed in the aforementioned papers. Chen et al. (2020) show an optimization-based separation between a depth 3 neural network in a suitably-defined quadratic NTK regime and a network of arbitrary depth trained in the NTK regime, where the former learns polynomials of degree p using width $d^{p/2}$ compared to d^p of the latter. However, this separation does not exclude the existence of a different shallow learner that operates in the non-NTK regime, whereas our work excludes efficient learnability in a standard GD setting, or even more generally with any shallow model due to inapproximability, thus providing a stronger separation.

2. Preliminaries and Main Results

In this section, we formally define our problem setting before stating our main theorems. We will begin however with introducing some of the notation that will be used throughout the paper in the following subsection.

2.1. Notation and Terminology

We use $[n]$ as shorthand for the set $\{1, \ldots, n\}$. We denote vectors using bold-faced letters (e.g. $\mathbf{x}$) and matrices or random variables using upper-case letters (e.g. X). Given a vector $\mathbf{x}$, we let $\|\mathbf{x}\|_2$ denote its Euclidean norm, where the subscript is occasionally omitted when clear from context. Given a square matrix A, we denote its spectral norm by $\|A\|_{\mathrm{sp}}$. Given two non-zero vectors $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, we denote the angle between them using $\theta_{\mathbf{w}, \mathbf{v}} := \arccos\left(\frac{\langle \mathbf{w}, \mathbf{v}\rangle}{\|\mathbf{w}\| \cdot \|\mathbf{v}\|}\right)$. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *radial* if for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, $\|\mathbf{x}\| = \|\mathbf{x}'\|$ implies $f(\mathbf{x}) = f(\mathbf{x}')$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *log-concave* if $f = \exp(-\varphi)$ where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is convex. A distribution is radial (resp., log-concave) if its density function is radial (resp., log-concave). For some $L > 0$, a function $f \in C^2(\mathbb{R})$ is *L-smooth* in a domain $A \subseteq \mathbb{R}^d$ if $f(\mathbf{x}) - f(\mathbf{x}') \leq \nabla f(\mathbf{x}')^\top (\mathbf{x} - \mathbf{x}') + \frac{L}{2} \|\mathbf{x} - \mathbf{x}'\|^2$ for all $\mathbf{x}, \mathbf{x}' \in A$. When optimizing an L-smooth objective, a learning rate (step size) η is *stable* if $\eta < 2/L$. We let $\mathbb{S}_R^{d-1} := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = R\}$ denote the d-dimensional hypersphere of radius R centered at the origin. We use $\mathcal{N}(\mu, \sigma^2)$ to denote a normal random variable with mean μ and variance σ^2, and $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ to denote a multivariate normal random variable with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

Our shallow class of neural networks with inferior approximation capabilities is the following class of depth 2 networks with architecture defined by

$$\mathbf{x} \mapsto \sum_{j=1}^r w_j \sigma\left(\langle \mathbf{u}_j, \mathbf{x}\rangle + b_j\right) + b_0, \tag{1}$$

where for terseness we occasionally denote the trainable parameters of this architecture in vectorized form using $\boldsymbol{\theta} := (\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{w}, \mathbf{b}, b_0)$. In the above architecture, σ is some non-linear activation function which satisfies the following assumption which we adopt from Eldan and Shamir (2016):

Assumption 1 (Polynomially-Bounded Activation) *The activation function σ is Lebesgue measurable and satisfies*

$$|\sigma(x)| \leq C_\sigma \left(1 + |x|^{\alpha_\sigma}\right)$$

for all $x \in \mathbb{R}$ and for some constants $C_\sigma, \alpha_\sigma > 0$.

We remark that the above assumption is very mild and is satisfied by all standard activation functions used in the literature, which includes common examples such as ReLU and sigmoidal activations, and in particular includes the error function and clipped ReLU activations which we use in the architecture to be defined next: For our stronger class with superior approximation capabilities, we consider the following 'depth 3' neural networks with a simplified structure,

$$\mathbf{x} \mapsto \left[\sum_{j=1}^{r} w_j \operatorname{erf}\left(\langle \mathbf{u}_j, \mathbf{x}\rangle + b_j\right)\right]_+^1. \tag{2}$$

In this architecture, we use error function activations $z \mapsto \operatorname{erf}(z)$ in the hidden layer, and a clipped ReLU activation $[z]_+^1 := \min\left\{[z]_+, 1\right\}$ in the output neuron, where $[z]_+ := \max\left\{0, z\right\}$ denotes the ReLU activation. We point out that strictly speaking, this is not a depth 3 neural network since there are no weights applied after the clipped ReLU activation. However, the additional non-linearity as we will see later in this section will prove to be necessary and sufficient for approximating the target functions considered in this paper. Our particular choice of the error function activation is motivated mainly by technical requirements, since it allows us to greatly simplify certain necessary calculations (e.g. in the proofs of Thm. 10 and Proposition 21). Likewise, the choice of the clipped ReLU is also motivated by technical considerations, and we conjecture our result to hold for a wider class of activation functions, leaving such generalizations to future work.

Since our goal is to separate different architectures according to their asymptotic performance with respect to the input dimension $d \geq 2$, our assumptions formally refer to sequences of distributions and target functions rather than individual objects. We assume that we are given access to a finite i.i.d. data set $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ sampled from the corresponding distribution $\mathcal{D}_d$ in a sequence $\{\mathcal{D}_d\}_{d=2}^{\infty}$, where the target values are determined according to a sequence of ball indicator functions, $y_i = \mathbb{1}\left\{\|\mathbf{x}_i\| \leq \lambda_d\right\}$ for $\{\lambda_d\}_{d=2}^{\infty}$ where $\lambda_d \in [1, 2]$. Since we will use a random features technique where the hidden layer remains fixed during training to derive our positive optimization result, we will use normal bold-faced letters ($\mathbf{x} \in \mathbb{R}^d$) and tilde bold-faced letters ($\hat{\mathbf{x}} \in \mathbb{R}^r$) to make the distinction between the d-dimensional data we are given and the r-dimensional random features produced by the output of the hidden layer in Eq. (2).

2.2. Main Results

We now turn to state our main theorems. The first shows that there exists a sequence of radial distributions with unbounded support such that efficiently learning a ball indicator to accuracy $\mathcal{O}(d^{-4})$ is not possible using any learning algorithm which must return a network with a weak architecture as defined in Eq. (1), whereas the strong architecture in Eq. (2) can efficiently learn such ball indicators using GD.

Theorem 2 *There exists a sequence of radial distributions $\{\mathcal{D}_d\}_{d=2}^{\infty}$ such that the following holds:*

- *There exist universal constants $c_1, c_2, c_3, c_4 > 0$ and a sequence of target functions $\{f_d(\mathbf{x}) := \mathbb{1}\{\|\mathbf{x}\| \leq \lambda_d\}\}_{d=c_1}^{\infty}$ where $\lambda_d \in [1, 2]$ for all d, such that for all $d \geq c_1$, any algorithm that returns a depth 2 neural network N with architecture as defined in Eq. (1) which employs an activation function σ satisfying Assumption 1 must satisfy*

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_d} \left[(N(\mathbf{x}) - f_d(\mathbf{x}))^2 \right] \geq \frac{c_2}{d^4},$$

  unless the algorithm runs in time at least $c_3 \exp(c_4 d)$.

- *For all $\epsilon, \delta \in (0, 1)$ and dimension $d \geq 2$, there exist r, T and sample size n that are $poly(1/\epsilon, \ln(1/\delta), d)$ and learning rate which is $1/poly(1/\epsilon, \ln(1/\delta), d)$, such that if we run at most T iterations of GD using a neural network with architecture as defined in Eq. (2), where the network's weights are initialized using a standard initialization scheme (see Assumption 7) and the hidden layer is held fixed during training. Then for all $\lambda \in [1, 2]$, with probability at least $1 - \delta$ over the randomness in the initialization of the network and the sampling of the data, the algorithm returns a neural network N such that*

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_d} \left[(N(\mathbf{x}) - \mathbb{1}\{\|\mathbf{x}\| \leq \lambda\})^2 \right] \leq \epsilon.$$

The proof of the above theorem appears in Appendix B.1. The main property of the distributions $\{\mathcal{D}_d\}_{d=2}^{\infty}$ used in our theorem is that they are heavy-tailed in the sense that they produce data instances with very large norm. Essentially, we will be penalized severely for not approximating data well even when it's far away from the origin. This however also results in a more challenging positive optimization result to achieve, since data with large norm is harder to learn; for example, this results in an objective with gradients having unbounded norm which significantly complicates our analysis. To partially circumvent this difficulty we use error function activations with image bounded in $[-1, +1]$ which mitigates the input's large norm to some extent.

We stress that in our result, the hardness of learning with a depth 2 neural network applies regardless of whether we are given access to a sufficiently large sample or not. In fact, since the hardness stems from the inability of depth 2 neural networks to approximate the target functions $\{f_d\}_{d=2}^{\infty}$, the hardness result persists even with perfect knowledge of the distributions and of the constants $\{\lambda_d\}_{d=2}^{\infty}$. On the other hand, our result only holds in a high-accuracy regime, and does not exclude the existence of an efficient algorithm which returns an approximation of a ball indicator to accuracy slightly larger than $c_2 d^{-4}$, which might be considered sufficient for certain practical applications in machine learning. The dependence on d^{-4} is an artifact of our proof technique (which relies on a reduction from the main result of Eldan and Shamir (2016)). To complement the above theorem, we present the following result which establishes the hardness of learning ball indicators to any accuracy smaller than a constant using a standard GD setting, for data generated by a certain compactly-supported distribution.

Theorem 3 *There exists a sequence of compactly-supported radial distributions $\{\mathcal{D}_d'\}_{d=2}^{\infty}$ such that the following holds:*

- *There exist constants $c_1, c_2 > 0$ and a sequence of target functions $\{f_d(\mathbf{x}) := \mathbb{1}\{\|\mathbf{x}\| \leq \lambda_d\}\}_{d=4}^{\infty}$ where $\lambda_d \in [1, 2]$ for all d, such that for all $d \geq 4$, suppose we run GD on the architecture defined in Eq. (1) which employs an activation function $\sigma \in C^2(\mathbb{R})$ satisfying Assumption 1 with the following assumptions:*

- – *The algorithm runs for $T \geq 4$ iterations.*
- – *The learning rate satisfies $\eta \leq \frac{2}{L} - \frac{1}{T}$ where L is the smoothness of the objective when training the weak architecture on n samples labeled according to a ball indicator.*
- – *We initialize at some $\boldsymbol{\theta}_0$ such that $\mathbb{P}\left[\|\boldsymbol{\theta}_0\|_2 \leq p(d, r)\right] \geq 1 - \exp(-d)$, for some polynomial p.*

Then with probability at least $1 - \exp(-d)$, for all $t \in \{0, \dots, T\}$, the network $N_{\boldsymbol{\theta}_t}(\cdot)$ at iteration t satisfies

$$\epsilon := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}'_d}\left[\left(N_{\boldsymbol{\theta}_t}(\mathbf{x}) - f_d(\mathbf{x})\right)^2\right] > \frac{1}{400},$$

unless the algorithm runs in time at least

$$c_1 \exp\left(c_2 \min\left\{\frac{1}{\sqrt{\epsilon}} \ln\left(d\sqrt{\epsilon} + 2\right), d \ln\left(\frac{1}{d\sqrt{\epsilon}} + 2\right)\right\}\right),$$

where $c_1, c_2 > 0$ depend solely on σ.

- • *For all $\epsilon, \delta \in (0, 1)$ and dimension $d \geq 3$, there exist r, T and sample size n that are $poly(1/\epsilon, \ln(1/\delta), d)$ and learning rate which is $1/poly(1/\epsilon, \ln(1/\delta), d)$, such that if we run at most T iterations of GD using a neural network with architecture as defined in Eq. (2), where the network's weights are initialized using a standard initialization scheme (see Assumption 7). Then for all $\lambda \in [1, 2]$, with probability at least $1 - \delta$ over the randomness in the initialization of the network and the sampling of the data, the algorithm returns a neural network N such that*

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}'_d}\left[\left(N(\mathbf{x}) - \mathbb{1}\left\{\|\mathbf{x}\|_2 \leq \lambda\right\}\right)^2\right] \leq \epsilon.$$

The proof of the theorem, which appears in Appendix B.2, builds on a reduction to the technique introduced in Daniely (2017), showing that an approximation to accuracy ϵ is only possible if either of the width of the approximating network or the Euclidean norm of its weights is exponential in either d or $\frac{1}{\sqrt{\epsilon}}$. Since the progress of GD and the norm of the weights upon initialization are bounded due to the assumptions in the theorem statement, we have that GD must perform a large number of iterations or that the network must be very wide to achieve a good approximation, resulting in an exponential running time of the algorithm in either of the parameters. This result essentially combines the approximation limitations of our weak architecture and the inability of GD with reasonably-sized steps to move much in weight space to produce a negative result.

Turning to discuss the assumptions made in the theorem in more detail, we first note that our assumptions are mostly mild. The assumption that $\sigma \in C^2(\mathbb{R})$ is to simplify the analysis, however since the non-smooth setting typically tends to result in worse optimization guarantees, we do not expect the running time to drastically change for the better if σ is assumed to be a ReLU for example. The polynomial boundedness assumption of the norm upon initialization is also very mild, and holds for essentially any initialization scheme used in the literature, which includes in particular the common Xavier initialization (see Glorot and Bengio (2010)), as well as the initialization scheme we use in our positive optimization result (see Assumption 7). Lastly, the bound on the learning rate is a necessary condition for the local convergence of GD in a general setting, since there are cases where no non-trivial guarantee can be given otherwise, as we demonstrate in Appendix H.

It is interesting to note that a similar trade-off between the dimension and accuracy parameters was given in Hsu et al. (2021), where it is shown that a certain high-frequency function cannot be approximated efficiently using depth 2 networks unless the width incurs exponential dependence on either d or $\frac{1}{\epsilon^2}$. Moreover, it can be shown that approximating a ball indicator (in the L_∞ norm) on a compact domain can be done efficiently in d if we only wish to obtain constant accuracy $\epsilon > 0$.[1] This provides a partial indication that the lower bound we derive here is somewhat tight, since a lower bound with exponential dependence on both parameters (i.e. exponential in $\max\{d, 1/\sqrt{\epsilon}\}$) is not possible.

We now turn to provide an explicit construction of the sequences of distributions used in our main theorems in the following subsection.

2.3. The Distributions Used

In this subsection, we formally define the sequence of distributions used in Thm. 2 and Thm. 3. We define a density of a d-dimensional distribution by

$$\bar{\mu}_d(\mathbf{x}) := \left(\frac{R_d}{\|\mathbf{x}\|}\right)^d J_{d/2}^2(2\pi\alpha\sqrt{d}R_d\|\mathbf{x}\|), \tag{3}$$

where $R_d = \frac{1}{\sqrt{\pi}}\left(\Gamma\left(\frac{d}{2}+1\right)\right)^{1/d}$ is the radius of the unit-volume d-dimensional Euclidean ball, $\alpha \geq 1$ is the absolute constant used in the derivation of the main result in Eldan and Shamir (2016), and J_ν is the Bessel function of the first kind, of order ν (see the aforementioned reference for further discussion about these objects). Such similar density functions (identical up to a linear change of variables) were used to show various inapproximability results for radial functions (Eldan and Shamir, 2016; Safran and Shamir, 2017; Safran et al., 2019). We now define the densities of our sequence of distributions $\{\mathcal{D}_d\}_{d=2}^\infty$ used in Thm. 2, which are a mixture of the distribution with density $\bar{\mu}_d$ defined in Eq. (3) and a multivariate normal with covariance matrix $\frac{1}{\sqrt{d}}I_d$:

$$\mu_d(\mathbf{x}) := 0.5\bar{\mu}_d(\mathbf{x}) + 0.5\left(\frac{d}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}d^d\|\mathbf{x}\|^2\right). \tag{4}$$

Turning to define the sequence of distributions $\{\mathcal{D}_d'\}_{d=2}^\infty$ used in Thm. 3, we define the d-dimensional distribution $\mathcal{D}_d'$ to be the distribution of $\mathbf{x} \in \mathbb{R}^d$ which is given by the sum

$$\mathbf{x} := \mathbf{x}_1 + \mathbf{x}_2, \quad \text{where} \quad \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}_1^{d-1} \text{ are i.i.d. and uniformly distributed.} \tag{5}$$

We denote the density of $\mathcal{D}_d'$ by φ_d.

Having defined the sequence of distributions used in our main results, we now turn to formally present our approximation lower bounds in the next section.

3. Lower Bounds for Approximating Ball Indicators

In this section, we present our lower bounds for the approximation of ball indicator functions used to derive the optimization lower bounds in our main results. As discussed in the related work section,

---

1. This can be achieved by first approximating a ball indicator to accuracy $\epsilon/2$ using a $2/\epsilon$-Lipschitz approximation, and then using the main result in Safran et al. (2019) to approximate the Lipschitz approximation to accuracy $\epsilon/2$ using width polynomial in d (albeit exponential in the constant $1/\epsilon$).

the inapproximability of ball indicator functions was previously shown in Safran and Shamir (2017), and our goal here is merely to adapt and extend similar bounds to our setting and assumptions.

The following theorem establishes the inapproximability of ball indicators with respect to the heavy-tailed distribution $\mathcal{D}_d$ defined in the previous section, which is used to prove Thm. 2.

Theorem 4 *The following holds for some universal constants $c_1, c_2, c_3, c_4 > 0$, and any network employing an activation function satisfying Assumption 1: For all $d \geq c_1$, there exist a sequence $\{\lambda_d\}_{d=c_1}^{\infty}$ of constants $\lambda_d \in [1, 2]$ and a sequence of radial distributions $\{\mathcal{D}_d\}_{d=2}^{\infty}$ with density functions $\{\mu_d\}_{d=2}^{\infty}$ defined in Eq. (4) such that for any depth 2 neural network N with architecture as defined in Eq. (1), we have*

$$\int_{\mathbb{R}^d} \left(N(\mathbf{x}) - \mathbb{1}\left\{\|\mathbf{x}\| \leq \lambda_d\right\}\right)^2 \mu_d(\mathbf{x})d\mathbf{x} \geq \frac{c_2}{d^4},$$

unless N has width at least $c_3 \exp(c_4 d)$.

The proof of the above theorem, which appears in Appendix C.1, relies on the main result of Eldan and Shamir (2016). Essentially, it is shown in the aforementioned work that a superposition of $\mathcal{O}(d^2)$ thin shells with varying radii cannot be approximated efficiently using depth 2 neural networks, with respect to a heavy-tailed distribution with density similar to the one defined in Eq. (3). This implies that there must *exist* a ball indicator function that is inapproximable with respect to the distribution with such a density, since otherwise we can efficiently construct an approximation of $\mathcal{O}(d^2)$ thin shells and concatenate them using a depth 2 neural network, contradicting the result of Eldan and Shamir (2016).

The following theorem establishes our second inapproximability result of ball indicators with respect to the compactly-supported distribution $\mathcal{D}'_d$ defined in the previous section, which is used to prove Thm. 3.

Theorem 5 *The following holds for any neural network employing an activation function σ which satisfies Assumption 1, and constants $c_1, c_2 > 0$ that may depend solely on σ: For all $d \geq 4$, there exist a sequence $\{\lambda_d\}_{d=4}^{\infty}$ of constants $\lambda_d \in [1, 2]$ and a sequence of radial distributions $\{\mathcal{D}'_d\}_{d=4}^{\infty}$ defined in Eq. (5) such that for any natural $m \geq 1$ and any depth 2 neural network $N_{\boldsymbol{\theta}}$ with architecture as defined in Eq. (1), we have*

$$\int_{\mathbb{R}^d} \left(N_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbb{1}\left\{\|\mathbf{x}\| \leq \lambda_d\right\}\right)^2 \varphi_d(\mathbf{x})d\mathbf{x} > \frac{1}{196(m+1)^2},$$

unless N has width r or weight norm $\|\boldsymbol{\theta}\|_2$ at least $c_1 \exp\left(c_2 \min\left\{m \ln\left(\frac{d}{m}+2\right), d \ln\left(\frac{m}{d}+2\right)\right\}\right)$.

The proof of the above theorem, which appears in Appendix C.2, builds on and refines the technique developed in Daniely (2017). We first show that approximating a radial function $f : \mathbb{R}^d \to \mathbb{R}$ consisting of $2m$ thin shells using a depth 2 neural network must require a large width or a large norm of the weights, since its radial component function (namely, the univariate function $g : \mathbb{R} \to \mathbb{R}$ such that $f(\mathbf{x}) = g(\|\mathbf{x}\|_2)$) cannot be approximated well using polynomials of degree at most m. Then, a similar reduction argument as the one used in the proof of Thm. 4 is used to derive a lower bound for a ball indicator function.

The benefit in the introduction of the parameter m which is independent of d is that it allows us to separate the dependence on the dimension and the accuracy parameters, leading to a hardness of

approximation result that persists beyond the high-accuracy regime used in the previous theorem, and which holds even for a (sufficiently small) constant accuracy requirement. On the flip side, since we cannot rule out the existence of a good approximation using a moderately-wide network with exponential weights, this comes at the cost of a somewhat weaker algorithmic implication where the hardness of learning is shown for just the GD algorithm and not for any algorithm as is the case in Thm. 2. Nevertheless, exponential weights have limitations beyond those that we exploit to show an optimization-based lower bound in Thm. 3. For example, the class of depth 2 neural networks with exponentially-bounded weights has an exponential FAT-shattering dimension and is therefore not statistically learnable in polynomial time (if we consider a broader class of target distributions than ball indicators). We leave the derivation of such statistically-based lower bounds and the generalization to a broader class of optimization algorithms as important future work directions.

4. Learning Ball Indicators Using Random Features

In this section, we prove that under certain assumptions on the distribution of the data, the family of ball indicator functions defined as $\{\mathbb{1}\{\|\mathbf{x}\| \leq \lambda\}\}_{\lambda \in [1,2]}$ can be learned efficiently by GD using a neural network with architecture defined in Eq. (2), with a polynomial network width, sample size and number of GD iterations. Recalling that we use a tilde to distinguish a d-dimensional data point $\mathbf{x} \in \mathbb{R}^d$ and r-dimensional random features $\tilde{\mathbf{x}} \in \mathbb{R}^r$ produced by the hidden layer, we now define our empirical objective function which takes the following form

$$\hat{F}(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^n \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - y_i \right)^2. \tag{6}$$

Additionally, we define the risk of a predictor $\mathbf{w}$ over a distribution $\mathcal{D}_d$ to be

$$F(\mathbf{w}) := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_d} \left[\left(\left[\mathbf{w}^\top \tilde{\mathbf{x}} \right]_+^1 - \mathbb{1}\{\|\mathbf{x}\| \leq \lambda\} \right)^2 \right] = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_d} \left[\hat{F}(\mathbf{w}) \right].$$

To facilitate our analysis, we make the following assumptions on the distributions of the data and the initialization of our neural network. Specifically, for the distribution of our data, we assume the following.

Assumption 6 (Data Distribution Assumption) *The sequence of distributions $\{\mathcal{D}_d\}_{d=2}^\infty$ possesses radial symmetry and the density γ_d of the distribution of its norm given by the random variable X_d satisfy the following properties:*

1. *There exists a constant $C > 0$ such that for all d*

$$\int_0^2 \gamma_d(x)dx \geq C.$$

2. *$\gamma_d(x) > 0$ for all $x \geq 0$, and there exists a polynomial q such that for all d and all $x \in [0.9, 2.05]$*

$$\gamma_d(x) \leq q(d).$$

11

3. *There exist $c_1, c_2 > 0$ such that for all $x \geq c_1$ and all d,*

$$\mathbb{P}\left[X_d \geq x\right] \leq c_2 \cdot \frac{1}{x}.$$

We remark that the above assumption is mild. The main non-trivial requirement is that $\{\mathcal{D}_d\}_{d=2}^{\infty}$ possess radial symmetry, but the remaining properties merely require that γ_d maintains a non-negligible portion of its mass within a constant distance from the origin as d grows, is polynomially-bounded for any d, and has a tail that decays at most proportionally to its distance from the origin. Note that this includes heavy-tailed distributions such as Cauchy for example.

Moving to our initialization of the network with architecture defined in Eq. (2), we assume the following:

Assumption 7 (Network Initialization)

1. *The weights in the hidden layer $U_j \in \mathbb{R}^d$ and the bias terms B_j satisfy*

$$U_j \sim \mathcal{N}\left(0, \frac{1}{4} \cdot I_d\right), \quad B_j \sim \mathcal{N}\left(0, \frac{1}{4}\right).$$

2. *The weights of the output neuron $W \in \mathbb{R}^r$ satisfy*

$$W \sim \mathcal{N}\left(0, \frac{1}{r^2} \cdot I_r\right).$$

We now turn to present our main result of this section. The following theorem shows that under Assumptions 6 and 7, for any $\epsilon, \delta \in (0,1)$, if r is polynomial in $d, 1/\epsilon$ and $\ln(1/\delta)$, and if the sample size n is cubic in r up to poly-logarithmic factors, then $\mathcal{O}(r \ln(r/\epsilon))$ iterations of GD will suffice to achieve a risk of at most ϵ with probability at least $1 - \delta$.

Theorem 8 *Under Assumptions 6 and 7, for all $\epsilon, \delta \in (0,1)$, suppose we run GD on the objective in Eq. (6) using an architecture as defined in Eq. (2) where the hidden layer is held fixed, with the following parameters:*

- *Network width $r \geq \max\left\{12000^4, c\epsilon^{-5}\left(q^4(d) + \ln^2(1/\delta)\right)\right\}$,*

- *Sample size $n = \lceil cr^3 \log_2^2(r) \rceil$,*

- *Fixed learning rate $\eta < \frac{\nu}{8r}$, where $\nu = \frac{C}{840 \cdot 9^4 \cdot 10^6}$.*

until the iterate $\mathbf{w}_t$ satisfies $\hat{F}(\mathbf{w}_t) \leq \frac{\epsilon}{2}$. Then with probability at least $1 - \delta$, we have that after at most $T = 2\eta^{-1}\nu^{-1} \ln(r/8\epsilon)$ iterations GD returns $\mathbf{w}_t$ such that $F(\mathbf{w}_t) \leq \epsilon$, where C, q, c_1, c_2 are defined in Assumption 6, and $c \geq 1$ is a constant that depends solely on C, c_1, c_2.

In a nutshell, the proof of the above theorem which appears in Appendix G.2, relies on a random features approach where the weights of the neurons in the hidden layer remain fixed throughout the optimization process, and then showing that the weights learned by GD in the output neuron converge with high probability to a point with risk at most ϵ. We remark that we did not attempt

to decrease the constants in the theorem statement, and that even when we fix the hidden layer, the resulting optimization problem is still non-convex in general and may even contain exponentially many sub-optimal local minima (Auer et al., 1996; Safran and Shamir, 2016). In a bit more detail, the proof can be broken down into four parts. In the first part, we show that with high probability, sufficiently many random features will result in the existence of a weight of the output neuron that attains small empirical loss. In the second part, we establish that the random features produced by the initialization scheme defined in Assumption 7 have well-behaved two-dimensional marginal distributions. This property is important for showing that the gradient of our objective points in the direction of the point with small empirical risk whose existence is shown in the first part. In the third part, we prove that GD converges with high probability to a point with small empirical loss. In the fourth and last part, we show generalization bounds which establish that our small empirical loss also results in a small risk on the distribution of the data. For a more detailed breakdown of the proof of the theorem and its intermediate steps, we refer the reader to Appendix A.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 2019.

Peter Auer, Mark Herbster, Manfred K Warmuth, et al. Exponentially many local minima for single neurons. *Advances in neural information processing systems*, pages 316–322, 1996.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Minshuo Chen, Yu Bai, Jason D Lee, Tuo Zhao, Huan Wang, Caiming Xiong, and Richard Socher. Towards understanding hierarchical learning: Benefits of neural representations. *arXiv preprint arXiv:2006.13436*, 2020.

Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, pages 690–696. PMLR, 2017.

DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.3 of 2021-09-15. URL http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. *arXiv preprint arXiv:2005.14426*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Surbhi Goel and Adam R Klivans. Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory*, pages 1470–1499. PMLR, 2019.

Oded Goldreich. *Foundations of Cryptography, Volume 2*. Cambridge university press, 2003.

Daniel Hsu, Clayton Sanford, Rocco A Servedio, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. *arXiv preprint arXiv:2102.02336*, 2021.

Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Neural tangent kernels, transportation mappings, and universal approximation. *arXiv preprint arXiv:1910.06956*, 2019.

Shiyu Liang and R Srikant. Why deep neural networks for function approximation? In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Eran Malach and Shai Shalev-Shwartz. Is deeper better only when shallow is good? *Advances in Neural Information Processing Systems*, 32:6429–6438, 2019.

Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. The connection between approximation, depth separation and learnability in neural networks. *arXiv preprint arXiv:2102.00434*, 2021.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Donald B Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090, 1956.

Donald Bruce Owen. A table of normal integrals: A table. *Communications in Statistics-Simulation and Computation*, 9(4):389–419, 1980.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29:3360–3368, 2016.

Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pages 774–782, 2016.

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International Conference on Machine Learning*, pages 2979–2987. PMLR, 2017.

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: what is actually being separated? In *Conference on Learning Theory*, pages 2664–2666. PMLR, 2019.

Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

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

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

Gal Vardi and Ohad Shamir. Neural networks with small weights and depth-separation barriers. *Advances in Neural Information Processing Systems*, 33, 2020.

Gal Vardi, Daniel Reichman, Toniann Pitassi, and Ohad Shamir. Size and depth separation in approximating benign functions with neural networks. In *Conference on Learning Theory*, pages 4195–4223. PMLR, 2021a.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Learning a single neuron with bias using gradient descent. *arXiv preprint arXiv:2106.01101*, 2021b.

Luca Venturi, Samy Jelassi, Tristan Ozuch, and Joan Bruna. Depth separation beyond radial functions. *arXiv preprint arXiv:2102.01621*, 2021.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2019.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. *arXiv preprint arXiv:2001.05205*, 2020.

Appendix A. Proof Sketch of Thm. 8

This appendix details the proof stages of the convergence of GD when learning ball indicator functions with respect to radial distributions.

A.1. Random Feature Approximation via Truncation

In this subsection, we show that random features can approximate ball indicator functions having radius in $[1, 2]$. The main idea is that we can truncate a particular random feature (neuron in the hidden layer) by setting its corresponding weight in the output neuron to zero. This way we can manipulate the expectation of the network's output to approximate a desired function, and then use concentration of measure to guarantee a good approximation with high probability. We now present the main theorem of this subsection:

Theorem 9 *Given any $\delta \in (0, 1)$, suppose that r and n satisfy*

- $r \geq \max\left\{12000^4, 500^4 \max\{c_1, 10\}^4 \ln^2(8n/\delta)\right\},$

- $\sqrt{\frac{\ln(8/\delta)}{2n}} \leq p := \frac{1800q(d) + 500c_2\sqrt{\ln(8n/\delta)}}{r^{0.25}}.$

Then for any $\lambda \in [1, 2]$, with probability at least $1 - \delta$ over drawing a sample $\{\mathbf{x}_i, \mathbb{1}\{\|\mathbf{x}_i\| \leq \lambda\}\}_{i=1}^n$ from $\mathcal{D}_d$ under Assumption 6 with parameters c_1, c_2, q, and over the randomness in the initialization of $\|\mathbf{w}_0\|$ and the remaining network's weights under Assumption 7, there exists $\mathbf{v} \in \mathbb{R}^r$ which satisfies the following properties:

1. $\|\mathbf{w}_0 - \mathbf{v}\|^2 < \|\mathbf{v}\|^2 \leq \frac{1}{\sqrt{r}}.$

2. *There exists a partition $\bigcup_{k=1}^4 I_k$ of $[0, \infty)$ into disjoint intervals I_k such that $\left[\mathbf{v}^\top \tilde{\mathbf{x}}_j\right]_+^1 = \mathbb{1}\{\|\mathbf{x}_j\| \leq \lambda\} = y_j$ for all j that satisfy $\|\mathbf{x}_j\| \in I_1 \cup I_3$, and $\left|\mathbf{v}^\top \tilde{\mathbf{x}}_j\right| \leq 3$ for all j that satisfy $\|\mathbf{x}_j\| \in I_2 \cup I_4$. Moreover, we have that*

$$|\{j : \|\mathbf{x}_j\| \in I_2 \cup I_4\}| \leq 2pn.$$

Putting it in simple words, the above theorem establishes that under some technical assumptions on the magnitude of r and n, there exists a point $\mathbf{v} \in \mathbb{R}^r$ in our optimization space that classifies 'most' of the data instances correctly, and thus attains an empirical loss which decays with r and n to zero. Moreover, the theorem also implies that this point is close enough to our initialization point $\mathbf{w}_0$ to attract GD (see Proposition 11 for a formal statement).

The proof of Thm. 9 which appears in Appendix D relies on concentration of measure and on the observation that due to the symmetry in Assumption 7, a neural network that is initialized in this manner will approximate the zero function in expectation. To bias the initialization of the network to approximate non-trivial functions, we can truncate undesired random features by setting their corresponding weight in the output neuron to zero. For example, by truncating random features with a negative bias term realization, we can bias the function computed by the network to return a positive prediction on data instances with small norm. Note that by doing so we do not change the initialization scheme defined in Assumption 7, but rather show that such a manipulation performed in the optimization space of the output neuron is equivalent to initializing using a different initialization scheme, which can effectively approximate various target functions using our architecture.

16

A.2. Radial Distributions Produce Well-Behaved Random Features

Using random features and keeping the hidden layer fixed throughout the optimization essentially reduces our learning problem to that of learning with a single neuron. To facilitate the convergence of GD in such a setting, a useful property to show is that the two-dimensional marginal distribution of the data has a strictly positive density in a neighborhood around the origin (Yehudai and Shamir, 2020; Frei et al., 2020). Such a property is desired since it allows one to show that if the target vector $\mathbf{v}$ that we wish to approximate and our current GD iterate $\mathbf{w}_t$ satisfy some certain properties, then the density assumption in the subspace spanned by $\mathbf{v}, \mathbf{w}_t$ guarantees that the gradient points in the direction of $\mathbf{v}$. To establish that this property is satisfied in our setting, we would need to analyze the r-dimensional distribution of the random features produced by Assumptions 6 and 7. To this end, we have the following theorem.

Theorem 10 *Suppose that the sequence of distributions $\{\mathcal{D}_d\}_{d=2}^\infty$ satisfies Assumption 6(1) with a constant $C > 0$, and that a neural network with an architecture defined in Eq. (2) is initialized according to Assumption 7. Then the sequence of distributions $\{\tilde{\mathcal{D}}_d\}_{d=2}^\infty$ of the random features produced by the first hidden layer satisfies the following: For any vectors $\mathbf{w}_1 \neq \mathbf{w}_2$, let $\tilde{\mathcal{D}}_{\mathbf{w}_1,\mathbf{w}_2}$ denote the marginal distribution of $\tilde{\mathbf{x}}$ on the subspace spanned by $\mathbf{w}_1, \mathbf{w}_2$ (as a distribution over $\mathbb{R}^2$). Then any such distribution has a density function $p_{\mathbf{w}_1,\mathbf{w}_2}(\mathbf{x})$ which satisfies $\inf_{\mathbf{x}:\|\mathbf{x}\|\leq\alpha} p_{\mathbf{w}_1,\mathbf{w}_2}(\mathbf{x}) \geq \beta$, where*

$$\alpha = \frac{1}{9}, \quad \beta = C \cdot 10^{-6}.$$

The proof of the above theorem, which appears in Appendix E, relies for the most part on the analysis of log-concave distributions performed in Lovász and Vempala (2007). The main challenge in the proof is that we cannot guarantee that the distributions $\{\tilde{\mathcal{D}}_d\}_{d=2}^\infty$ are log-concave. Instead, we show that the density of these distributions can be lower bounded by a log-concave density, from which the theorem follows. With Thms. 9 and 10 at hand, we can turn to show a convergence result for GD in the next subsection.

A.3. Convergence of Gradient Descent

In this subsection, we establish the technical tools that facilitate the convergence of GD. We begin with the following proposition, which establishes that as long as some technical inequality is satisfied, then the gradient of the objective in Eq. (6) points us in the direction of the vector $\mathbf{v}$ which achieves small empirical loss and whose existence is established by Thm. 9. More formally, we have the following:

Proposition 11 *Under Assumptions 6 and 7, there exist constants $\alpha, \beta > 0$ such that for all $\delta \in (0, 1)$, with probability at least $1 - \delta$ over sampling n data instances from $\mathcal{D}_d$ and over the randomness in the initialization of the network, for all $\mathbf{w}, \mathbf{v} \in \mathbb{R}^r$ that satisfy $\|\mathbf{w} - \mathbf{v}\| \leq \|\mathbf{v}\| \leq 0.5$ and the following technical inequality*

$$\frac{1}{n}\sum_{i=1}^n \left(\left[\mathbf{w}^\top\tilde{\mathbf{x}}_i\right]_+^1 - \left[\mathbf{v}^\top\tilde{\mathbf{x}}_i\right]_+^1 \right) \cdot \mathbb{1}\left\{\mathbf{w}^\top\tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \left(\mathbf{w}^\top\tilde{\mathbf{x}}_i - \mathbf{v}^\top\tilde{\mathbf{x}}_i\right)$$

$$\geq \frac{2}{n}\sum_{i=1}^n \left(\left[\mathbf{v}^\top\tilde{\mathbf{x}}_i\right]_+^1 - y_i \right) \cdot \mathbb{1}\left\{\mathbf{w}^\top\tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \left(\mathbf{w}^\top\tilde{\mathbf{x}}_i - \mathbf{v}^\top\tilde{\mathbf{x}}_i\right), \tag{7}$$

we have that

$$\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{v} \rangle \geq \frac{1}{2} \left\| \mathbf{w} - \mathbf{v} \right\|^2 \left(\frac{\alpha^4 \beta}{210} - 4r \sqrt{\frac{(8r+8)\log_2(n)}{n}} - \sqrt{\frac{2\ln(2/\delta)}{n}} \right).$$

The proof of the above proposition, which appears in Appendix F.1, relies on lower bounding the dot product with the gradient using the assumed inequality, and then using the marginal density property from the previous subsection to guarantee that the gradient of the objective is correlated with the direction leading to $\mathbf{v}$. In a sense, this result can be viewed as a finite data analog in the non-realizable setting of Thm. 4.2 in Yehudai and Shamir (2020). The main difficulty that arises in the finite data case is that in the application of the proposition, $\mathbf{w}$ is an iterate returned by GD and thus we need to derive the above result uniformly for all the $\mathbf{w}$'s that satisfy the above condition.

The following proposition establishes that when the technical inequality assumption in Eq. (7) is violated then we have already achieved a small empirical loss.

Proposition 12 *Under Assumptions 6 and 7, for all $\delta \in (0,1)$, with probability at least $1 - \delta$ over sampling n data instances from $\mathcal{D}_d$ and over the randomness in the initialization of the network, for all $\mathbf{w}, \mathbf{v} \in \mathbb{R}^r$ that satisfy $\|\mathbf{w} - \mathbf{v}\| \leq \|\mathbf{v}\| \leq r^{-0.25}$ and the following technical inequality*

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right) \cdot \mathbb{1} \left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1) \right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i \right)$$

$$< \frac{2}{n} \sum_{i=1}^{n} \left(\left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 - y_i \right) \cdot \mathbb{1} \left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1) \right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i \right),$$

if r and n satisfy the following inequalities

- $r \geq \max \left\{ 12000^4, 500^4 \max \{c_1, 10\}^4 \ln^2(16n/\delta) \right\}$,

- $\sqrt{\frac{\ln(16/\delta)}{2n}} \leq p' := \frac{1800q(d) + 500c_2 \sqrt{\ln(16n/\delta)}}{r^{0.25}}$.

Then we have that

$$\hat{F}(\mathbf{w}) \leq 20p' + \frac{2}{r} + 8\sqrt{\frac{(8r+8)\log_2(n)}{n}} + 2\sqrt{\frac{2\ln(4/\delta)}{n}}.$$

The proof of the above proposition which appears in Appendix F.2 is technical, and is based on lower bounding the smaller term in the proposition assumption by the empirical loss. This is achieved by analyzing several different cases depending on the values attained by the dot products of $\mathbf{v}$ and $\mathbf{w}$ with the data instances. We then upper bound the larger term in the proposition assumption by using Thm. 9. Combining the two propositions stated in this subsection, the convergence of GD follows from the fact that we converge to $\mathbf{v}$ at a linear rate by Proposition 11 as long as the technical inquality holds, and if it is violated then Proposition 12 guarantees that the empirical loss we achieved is sufficiently small.

The final component in our proof of Thm. 8 is to show that the small empirical loss achieved using the above propositions translates to a generalization bound. To this end, we have the following theorem:

Theorem 13 *For all $\delta \in (0,1)$, with probability at least $1 - \delta$ over drawing a sample of size n from any distribution $\hat{\mathcal{D}}_d$, we have for all $\mathbf{w}$ satisfying $\|\mathbf{w}\| \leq 1$ that*

$$F(\mathbf{w}) \leq \hat{F}(\mathbf{w}) + 4\sqrt{\frac{r}{n}} + 4\sqrt{\frac{2\ln(4/\delta)}{n}}.$$

The proof of the above theorem, which appears in Appendix G.1, is simple and relies on standard Rademacher complexity arguments.

Appendix B. Proofs from Sec. 2 – Main Results

B.1. Proof of Thm. 2

Proof Starting with the lower bound in the first item in the theorem statement, we have that it follows from Thm. 4, since if an algorithm has returned a depth 2 neural network N satisfying

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_d} \left[(N(\mathbf{x}) - f(\mathbf{x}))^2 \right] < \frac{c_2}{d^4},$$

then it must hold that the width of the network is at least $c_3 \exp(c_4 d)$. Thus for the algorithm to return such an output it must run in time at least $c_3 \exp(c_4 d)$.[2]

Turning to show the second item, we first argue that if learning is computationally tractable for a certain sequence of distributions, then it is also tractable for any other sequence of distributions which is indistinguishable by polynomial-time sampling (see Goldreich (2003, Def. 3.2.4)). This is true since otherwise we would be able to distinguish the two sequences using the learning algorithm which would succeed on the first sequence but would fail on the second.

Next, we will show that a sequence which is indistinguishable by polynomial-time sampling from the sequence defined by the densities in Eq. (4), satisfies Assumption 6. Consider a random variable X_d which follows a distribution with density given in Eq. (4). Since this is a mixture of radial distributions, it is clearly radial. We will now show that the density of $\|X_d\|$ is given by

$$\gamma_d(x) := 0.5\bar{\gamma}_d(x) + 0.5p_d(x),$$

where

$$\bar{\gamma}_d(x) := \frac{d}{x} J_{d/2}^2 (2\pi\alpha\sqrt{d}R_d x), \quad p_d(x) := \frac{2(d/2)^{d/2}}{\Gamma(d/2)} x^{d-1} \exp\left(-0.5dx^2\right). \tag{8}$$

For the multivariate normal component, we have that it follows a scaled chi-distribution with d degrees of freedom, given by $\frac{1}{\sqrt{d}}\chi_d$. Since a χ_d distribution is a special case of the generalized gamma distribution $GG(\sqrt{2}, d, 2)$, and since scaling a generalized gamma random variable by a constant $\frac{1}{\sqrt{d}}$ results in a generalized gamma random variable with parameters $GG(\sqrt{2/d}, d, 2)$, we have that its cumulative distribution function is given by

$$P\left(\frac{d}{2}, \frac{dx^2}{2}\right),$$

---

2. Formally, we have that a Turing machine returning a network of width at least $c_3 \exp(c_4 d)$ will require exponential time to output the network's weights on its tape.

where $P(\cdot, \cdot)$ denotes the regularized (lower) incomplete gamma function. Its density is therefore given by $p_d(x)$.

Moving to the density of the norm of the $\bar{\mu}_d$ component, we let $A_d := \frac{d\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ denote the volume of the unit hypersphere in $\mathbb{R}^d$ and we let Y denote a random variable with density $\bar{\mu}_d$. Changing to polar coordinates, we have

$$
\mathbb{P}\left[\|Y\| \leq x\right] = \int_{\{\mathbf{x}:\|\mathbf{x}\| \leq x\}} \bar{\mu}_d(\mathbf{x}) d\mathbf{x} = \int_{\{\mathbf{x}:\|\mathbf{x}\| \leq x\}} \left(\frac{R_d}{\|\mathbf{x}\|}\right)^d J_{d/2}^2(2\pi\alpha\sqrt{d}R_d \|\mathbf{x}\|) d\mathbf{x}
$$

$$
= \int_0^x \int_{\mathbf{y} \in \mathbb{S}_t^{d-1}} \left(\frac{R_d}{t}\right)^d J_{d/2}^2 \left(2\pi\alpha\sqrt{d}R_d t\right) d\mathbf{y} dt
$$

$$
= \int_0^x A_d t^{d-1} \left(\frac{R_d}{t}\right)^d J_{d/2}^2 \left(2\pi\alpha\sqrt{d}R_d t\right) dt = \int_0^x \frac{d}{t} J_{d/2}^2 (2\pi\alpha\sqrt{d}R_d t) dt.
$$

Thus, the density of the norm of a random variable with density given in Eq. (3) is given by $\bar{\gamma}_d(\cdot)$.

We now prove each of the required properties.

1. It will suffice to lower bound the norm of the multivariate normal component. Using a standard bound on $P(\cdot, \cdot)$ (see Eq. (8.10.13) in DLMF) we have

$$
P\left(\frac{d}{2}, \frac{d}{2}\right) > \frac{1}{2},
$$

and thus

$$
\mathbb{P}\left[X_d \leq 2\right] \geq \mathbb{P}\left[X_d \leq 1\right] \geq \frac{1}{4}.
$$

2. $p(x) > 0$ for all $x > 0$ follows immediately from the definition of $p_d(\cdot)$ in Eq. (8). Note that even though $p(0) = 0$, we can mix the multivariate normal distribution with another radial d-dimensional distribution that has positive density at the origin (e.g. the density on the closed ball $B_r(\mathbf{0}) = \{\mathbf{x} : \|\mathbf{x}\| \leq r\}$ whose pushforward via the function $\mathbf{x} \mapsto x_1$ onto $[-r, r]$ is the uniform measure). By taking r exponentially small in d and using an appropriately sufficiently small weight, we can make the two distributions statistically indistinguishable by polynomial-time sampling, and therefore we may assume without loss of generality that $p(0) > 0$.

To upper bound the density γ_d, we begin with upper bounding $0.5 p_d(x)$. The mode of a generalized gamma random variable $GG(\sqrt{2/d}, d, 2)$ where $d > 1$ is attained at $x = \sqrt{1 - 1/d}$, and we thus have

$$
\sup_x 0.5 p_d(x) = \frac{(d/2)^{d/2}}{\Gamma(d/2)} \left(1 - \frac{1}{d}\right)^{(d-1)/2} \exp\left(-\frac{d}{2} + 0.5\right).
$$

Plugging $d = 2$ in the above yields an upper bound of 1. Using a standard lower bound on the gamma function we have for all $d \geq 3$

$$
\Gamma(d/2) > \sqrt{2\pi} \left(\frac{d}{2} - 1\right)^{\frac{d}{2}-\frac{1}{2}} \exp\left(-\frac{d}{2} + 1\right),
$$

20

which when plugged in the above and simplified a bit yields

$$\sup_x 0.5 p_d(x) < \frac{1}{\sqrt{2\pi}} \exp(-0.5) \sqrt{\frac{d}{2}} \left(1 + \frac{1}{d-2}\right)^{(d-1)/2} \leq \sqrt{\frac{d}{2}},$$

where the second inequality is due to $(1 + 1/(d-2))^{(d-1)/2} \leq 2$ for all $d \geq 3$ and $\frac{1}{\sqrt{2\pi}} \exp(-0.5) \leq 0.5$. Upper bounding $0.5\bar{\gamma}_d(x)$ we have from Eq. (10.14.4) in DLMF that

$$\gamma_d(x) = \frac{d}{x} J_{d/2}^2(\pi\alpha\sqrt{d}R_d x) \leq \frac{d}{x} \cdot \frac{(\pi\alpha\sqrt{d}R_d x)^d}{\Gamma^2(d/2+1)}.$$

Assuming $x \leq (\pi\alpha\sqrt{d}R_d)^{-2} < 1$ and recalling that $d \geq 2$, the above is at most

$$\frac{d}{x} \cdot \frac{x^{d/2}}{\Gamma^2(d/2+1)} \leq d.$$

Otherwise, if $x > (\pi\alpha\sqrt{d}R_d)^{-2}$, we use Eq. (10.14.1) in DLMF and Eldan and Shamir (2016, Lemma 1) to deduce that

$$\gamma_d(x) = \frac{d}{x} J_{d/2}^2(\pi\alpha\sqrt{d}R_d x) \leq \frac{d}{x} < d(\pi\alpha\sqrt{d}R_d)^2 \leq \frac{\pi^2\alpha^2}{4} d^3.$$

We conclude that there exists a polynomial q of degree 3 which satisfies Item 2 in the assumption.

3. By Eldan and Shamir (2016, Lemma 14) we have for all $x \geq 1$ that

$$\int_x^\infty \gamma_d(t)dt = \int_x^\infty \frac{d}{t} J_{d/2}^2(\pi\alpha\sqrt{d}R_d t)dt \leq \int_x^\infty \frac{d}{t} \cdot \frac{1.3}{t\alpha d}dt = \frac{1.3}{\alpha}\left[-t^{-1}\right]_x^\infty = \frac{1.3}{\alpha x}.$$

For the multivariate component, we compute for any $x \geq 2$

$$\mathbb{P}\left[\frac{1}{\sqrt{d}}\chi_d \geq x\right] = \mathbb{P}\left[\chi_d^2 \geq x^2 d\right] \leq (x^2\exp(1-x^2))^{d/2} \leq (x^2\exp(1-x^2)) \leq \frac{1}{2x},$$

where the first inequality is a standard tail bound on chi-squared random variables, the second inequality is due to $\sup_x x^2\exp(1-x^2) \leq 1$ and $d \geq 2$, and the last inequality holds for all $x \geq 2$. Combining the last two displayed inequalities we have for all $x \geq 2$ that

$$\mathbb{P}[X_d \geq x] \leq \frac{1}{2}\left(\frac{1.3}{\alpha x} + \frac{1}{2x}\right) \leq \frac{1}{x},$$

where we used the fact that $\alpha \geq 1$.

$\blacksquare$

B.2. Proof of Thm. 3

Before we prove the theorem, we will first state and prove the following lemmas. Throughout this subsection of the appendix, we use $\mathcal{L}(\boldsymbol{\theta})$ to denote the objective function obtained by making predictions with an architecture as defined in Eq. (1), over n data instances sampled from the distribution $\mathcal{D}'_d$ defined in Eq. (5), and using the squared loss. We also define the set

$$\mathcal{F}(d,m) := \Big\{ f : \mathbb{R}^2_+ \to \mathbb{R} \mid \exists c_1, c_2 > 0 \text{ s.t. } \forall d, m > 0,$$
$$f(d,m) \geq c_1 \exp \left(c_2 \min \left\{ m \ln \left(\frac{d}{m} + 2 \right), d \ln \left(\frac{m}{d} + 2 \right) \right\} \right) \Big\}, \quad (9)$$

which represents the set of functions that grow exponentially fast in at least one of the parameters d or m. The lemma below establishes some useful properties of the function class $\mathcal{F}(d,m)$.

Lemma 14 *Suppose that $f \in \mathcal{F}(d,m)$ where $\mathcal{F}(d,m)$ is defined in Eq. (9). Then*

1. *For any $c, k > 0$,*
$$c \sqrt[k]{f(d,m)} \in \mathcal{F}(d,m).$$

2. *For any $c > 0$,*
$$f(c \cdot d, m) \in \mathcal{F}(d,m) \quad \text{and} \quad f(d, c \cdot m) \in \mathcal{F}(d,m).$$

3. *For any polynomial p in d, m, we have*
$$\frac{1}{p(d,m)} f(d,m) \in \mathcal{F}(d,m).$$

Proof

1. This follows immediate from the definition of $\mathcal{F}(d,m)$ by taking the constants $c \sqrt[k]{c_1} > 0$ and $c \cdot c_2/k > 0$ instead of c_1 and c_2.

2. From symmetry, it suffices to show $f(c \cdot d, m) \in \mathcal{F}(d,m)$. First suppose that $c \geq 1$, then we have
$$\frac{cd}{m} + 2 \geq \frac{d}{m} + 2$$
and
$$cd \ln \left(\frac{m}{cd} + 2 \right) \geq d \ln \left(\frac{m}{d} + 2 \right),$$
since $x \mapsto x \ln(a/x + 2)$ is increasing in x for all $a, x > 0$. We therefore have
$$f(c \cdot d, m) \geq c_1 \exp \left(c_2 \min \left\{ m \ln \left(\frac{d}{m} + 2 \right), d \ln \left(\frac{m}{d} + 2 \right) \right\} \right) \in \mathcal{F}(d,m).$$

   Suppose that $c < 1$. Then
$$m \ln \left(\frac{cd}{m} + 2 \right) \geq cm \ln \left(\frac{cd}{cm} + 2 \right) = cm \ln \left(\frac{d}{m} + 2 \right),$$

where again we used the fact that $x \mapsto x \ln (a/x + 2)$ is increasing in x for all $a, x > 0$. Furthermore, we have

$$cd \ln \left(\frac{m}{cd} + 2 \right) \geq cd \ln \left(\frac{cm}{cd} + 2 \right) = cd \ln \left(\frac{m}{d} + 2 \right).$$

Combining these two inequalities we obtain

$$f(c \cdot d, m) \geq c_1 \exp \left(c \cdot c_2 \min \left\{ m \ln \left(\frac{d}{m} + 2 \right), d \ln \left(\frac{m}{d} + 2 \right) \right\} \right) \in \mathcal{F}(d, m).$$

3. Since any polynomial in d, n can be upper bound by $cd^k m^k$ for sufficiently large $c > 0$ and natural k, it suffices to show this item for $p(d, m) = d^k m^k$. By symmetry, showing this for d^k and m^k is equivalent, and by applying this rule iteratively for $\frac{1}{m}$ (or equivalently for $\frac{1}{d}$), it suffices to prove this item for just the case $p(d, m) = \frac{1}{m}$. We have by assumption that there exist $c_1, c_2 > 0$ such that for all $d, m > 0$ we have

$$\frac{1}{m} f(d, m) \geq \frac{1}{m} c_1 \exp \left(c_2 \min \left\{ m \ln \left(\frac{d}{m} + 2 \right), d \ln \left(\frac{m}{d} + 2 \right) \right\} \right).$$

We first handle the case where $m > d$, in which the above is at least

$$\begin{aligned}
\frac{1}{m} c_1 \exp \left(c_2 d \ln \left(\frac{m}{d} + 2 \right) \right) &= \frac{1}{m} c_1 \left(\frac{m}{d} + 2 \right)^{c_2 d} \geq \frac{1}{m} c_1 \left(\frac{m}{d} + 1 \right)^{c_2 d} \\
&= \frac{1}{m} c_1 \left(\frac{m}{d} + 1 \right)^{c_2 d/2} \left(\frac{m}{d} + 1 \right)^{c_2 d/2} \\
&\geq \frac{1}{m} c_1 \left(\frac{c_2 m}{2} + 1 \right) \left(\frac{m}{d} + 1 \right)^{c_2 d/2} \geq \frac{1}{2} c_1 c_2 \left(\frac{m}{d} + 1 \right)^{c_2 d/2} \\
&= \frac{1}{2} c_1 c_2 \exp \left(c_2 \frac{d}{2} \ln \left(\frac{m}{d} + 1 \right) \right) \\
&\geq \frac{1}{2} c_1 c_2 \exp \left(c_2 \frac{d}{4} \ln \left(\frac{m}{d} + 2 \right) \right) \\
&= \frac{1}{2} c_1 c_2 \exp \left(\frac{c_2}{4} \min \left\{ d \ln \left(\frac{m}{d} + 2 \right), m \ln \left(\frac{d}{m} + 2 \right) \right\} \right) \\
&\in \mathcal{F}(d, m).
\end{aligned}$$

In the above, the second inequality follows from Bernoulli's inequality and the fourth inequality follows from the inequality $0.5 \ln(2 + x) / \ln(1 + x) \leq 1$ which holds for all $x \geq 1$ and since we assume $m > d$. Turning to the case $m \leq d$, we have

$$\frac{1}{m} f(d, m) \geq \frac{1}{d} f(d, m),$$

where by symmetry we have that the proof follows from the same reasoning as in the previous case, where the roles of d and m are switched.

■

The following lemma bounds the loss of the objective function in terms of the Euclidean norm of the weights and the width of the network.

Lemma 15 *Suppose that σ satisfies Assumption 1. Then there exists a polynomial p with coefficients and degree that depend solely on σ such that*

$$\mathcal{L}(\boldsymbol{\theta}) \leq r^2 p\left(\|\boldsymbol{\theta}\|_2\right).$$

Proof We begin with bounding the output of the hidden layer. For any data point $\mathbf{x}_i$ we have from Cauchy-Schwartz that $\langle \mathbf{u}_j, \mathbf{x}_i \rangle \leq \|\mathbf{u}_j\|_2 \|\mathbf{x}_i\|_2 \leq 2\|\mathbf{u}_j\|$, where we used the fact that $\mathcal{D}'_d$ is supported on a ball of radius 2. By Assumption 1 we have constants $C_\sigma, \alpha_\sigma > 0$ such that

$$\sigma\left(\langle \mathbf{u}_j, \mathbf{x}_i \rangle + b_j\right) \leq C_\sigma \left(1 + \left|2\|\mathbf{u}_j\|_2 + |b_j|\right|^{\alpha_\sigma}\right) \leq C_\sigma \left(1 + |3\|\boldsymbol{\theta}\|_2|^{\alpha_\sigma}\right).$$

Using the above, we get

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{r} w_j \sigma\left(\langle \mathbf{u}_j, \mathbf{x}_i \rangle + b_j\right) + b_0 - y_i\right)^2 \\
&\leq \frac{2}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{r} w_j \sigma\left(\langle \mathbf{u}_j, \mathbf{x}_i \rangle + b_j\right) + b_0\right)^2 + 2 \\
&\leq \frac{2}{n} \sum_{i=1}^{n} \left(rC_\sigma \|\boldsymbol{\theta}\|_2 \left(1 + |3\|\boldsymbol{\theta}\|_2|^{\alpha_\sigma}\right) + \|\boldsymbol{\theta}\|_2\right)^2 + 2 \\
&\leq 2r^2 \left(\left(C_\sigma \|\boldsymbol{\theta}\|_2 \left(1 + |3\|\boldsymbol{\theta}\|_2|^{\alpha_\sigma}\right) + \|\boldsymbol{\theta}\|_2\right)^2 + 1\right),
\end{aligned}$$

where the first inequality uses the inequality $(a + b)^2 \leq 2(a^2 + b^2)$ which holds for all a, b. The lemma then follows from the above bound. ∎

Proof [of Thm. 3] We start the proof with showing the lower bound. Suppose we train a neural network with architecture defined in Eq. (1) using GD for $T \geq 4$ iterations. Then we have that the running time of the algorithm is lower bounded by $\max\{r, T\}$. Further assume that the algorithm returned a set of weights $\boldsymbol{\theta}_T$ such that

$$\mathcal{L}(\boldsymbol{\theta}_T) \leq \epsilon \leq \frac{1}{400}.$$

The bulk of the remainder of the lower bound proof focuses on lower bounding T in terms of the norm of $\boldsymbol{\theta}_T$. Since our bound on $\|\boldsymbol{\theta}_T\|_2$ will be shown to be increasing with T, it suffices to derive the bound for just the iterate at the final iteration T in order to obtain a lower bound on the loss of all the other iterates $t < T$ as well. We compute

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_0\|_2 \leq \sum_{t=1}^{T} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_2 = \eta \sum_{t=1}^{T} \|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2 \leq \eta \sqrt{T \sum_{t=1}^{T} \|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2^2}, \quad (10)$$

where the first inequality is due to the triangle inequality, the equality is due to the update rule $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_t)$, and the second inequality is due to Cauchy-Schwartz. We now argue that $\mathcal{L}(\cdot)$ is L-smooth for some positive $L > 0$. This follows from our assumption $\sigma \in C^2(\mathbb{R})$ which

implies that $\nabla^2 \mathcal{L}(\cdot)$ is a continuous mapping, and thus $\left\| \nabla^2 \mathcal{L}(\boldsymbol{\theta}) \right\|_{\mathrm{sp}}$ attains its maximum on any compact domain. In particular, we may define

$$L := \max_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\|_2 \leq cTd^\ell r^\ell} \left\| \nabla^2 \mathcal{L}(\boldsymbol{\theta}) \right\|_{\mathrm{sp}},$$

for sufficiently large $c, \ell > 0$ (this norm bound will be justified later in the proof). Since $\mathcal{L}(\cdot)$ is L-smooth, we have that

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) - \mathcal{L}(\boldsymbol{\theta}_t) \leq \nabla \mathcal{L}(\boldsymbol{\theta}_t)^\top (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + \frac{L}{2} \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_2^2.$$

Plugging the update rule in the above, simplifying and rearranging yields

$$\left(\eta - \eta^2 \frac{L}{2} \right) \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 \leq \mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}).$$

Summing over $t \in [T]$ we have that the right-hand side telescopes, and by the fact that $\mathcal{L}(\boldsymbol{\theta}_T) \geq 0$ we get

$$\sum_{t=1}^{T} \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 \leq \left(\eta - \eta^2 \frac{L}{2} \right)^{-1} \mathcal{L}(\boldsymbol{\theta}_0).$$

Plugging the above back in Eq. (10) we have

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_0\|_2 \leq \sqrt{T\mathcal{L}(\boldsymbol{\theta}_0)} \cdot \eta \left(\eta - \eta^2 \frac{L}{2} \right)^{-0.5} = \sqrt{T\mathcal{L}(\boldsymbol{\theta}_0)} \cdot \sqrt{\eta} \left(1 - \eta \frac{L}{2} \right)^{-0.5}.$$

Since the function $x \mapsto \sqrt{x} \left(1 - x \frac{a}{2} \right)^{-0.5}$ is increasing in x for all $x \in (0, 2/a)$ where $a > 0$, we have by our assumption $\eta \leq \frac{2}{L} - \frac{1}{T}$ that

$$\sqrt{\eta} \left(1 - \eta \frac{L}{2} \right)^{-0.5} \leq \sqrt{\frac{2}{L} - \frac{1}{T}} \left(1 - \left(\frac{2}{L} - \frac{1}{T} \right) \frac{L}{2} \right)^{-0.5} \leq \sqrt{\frac{2}{L}} \left(\frac{L}{2T} \right)^{-0.5} = \frac{2}{L} \sqrt{T},$$

which when plugged in the previous inequality yields

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_0\|_2 \leq \frac{2T}{L} \sqrt{\mathcal{L}(\boldsymbol{\theta}_0)}. \tag{11}$$

To lower bound L, observe that $\mathcal{L} \in C^2(\mathbb{R})$ since $\sigma \in C^2(\mathbb{R})$ by assumption. Moreover, it is readily seen that in this case we have that $\frac{\partial^2}{\partial^2 b_0} \mathcal{L}(\boldsymbol{\theta}) = 2$. Consider the unit vector $\mathbf{e}_1 = (1, 0, \ldots, 0) \in \mathbb{R}^k$, where $k := r(d+2) + 1$ is the dimension of our optimization space. If we assume that $\frac{\partial^2}{\partial^2 b_0} \mathcal{L}(\boldsymbol{\theta})$ is the first diagonal entry in the Hessian $H(\boldsymbol{\theta}) := \nabla^2 \mathcal{L}(\boldsymbol{\theta})$, then we have that it equals 2 for all $\boldsymbol{\theta} \in \mathbb{R}^k$. Along with the fact that $H(\boldsymbol{\theta})$ is symmetric, this implies that for all $\boldsymbol{\theta}$

$$2 = \mathbf{e}_1^\top H(\boldsymbol{\theta}) \mathbf{e}_1 = \mathbf{v}^\top D(\boldsymbol{\theta}) \mathbf{v}$$

where $H(\boldsymbol{\theta}) = O(\boldsymbol{\theta})^\top D(\boldsymbol{\theta}) O(\boldsymbol{\theta})$ with $D(\boldsymbol{\theta})$ diagonal where its entries are the eigenvalues of $H(\boldsymbol{\theta})$, $O(\boldsymbol{\theta})$ is orthogonal and $\mathbf{v} = O(\boldsymbol{\theta}) \mathbf{e}_1$. Since $O(\boldsymbol{\theta})$ is an isometry we have $\|\mathbf{v}\|_2 = 1$, implying that

$$2 = \sum_{i=1}^{k} d_i v_i^2.$$

Now, if $d_i < \frac{2}{k}$ for all i we have

$$\sum_{i=1}^{k} d_i v_i^2 < \frac{2}{k} \sum_{i=1}^{k} v_i^2 = 2,$$

which leads to a contradiction. We therefore have that the largest eigenvalue of $H(\boldsymbol{\theta})$ is at least $2/k$, implying $L \geq 2/k$ since $\mathcal{L} \in C^2(\mathbb{R})$. Plugging this back in Eq. (11) we have arrived at

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_0\|_2 \leq Tk\sqrt{\mathcal{L}(\boldsymbol{\theta}_0)} \leq 3drT\sqrt{\mathcal{L}(\boldsymbol{\theta}_0)},$$

where we assume $d \geq 2$. By the triangle inequality this implies

$$\|\boldsymbol{\theta}_T\|_2 \leq \|\boldsymbol{\theta}_T - \boldsymbol{\theta}_0\|_2 + \|\boldsymbol{\theta}_0\|_2 \leq 3drT\sqrt{\mathcal{L}(\boldsymbol{\theta}_0)} + \|\boldsymbol{\theta}_0\|_2 \leq 3drT\left(\sqrt{\mathcal{L}(\boldsymbol{\theta}_0)} + \|\boldsymbol{\theta}_0\|_2\right).$$

From Lemma 15 and our initialization assumption, we have that the above can be upper bounded by $cTd^\ell r^\ell$ with probability at least $1 - \exp(-d)$ for sufficiently large $c, \ell > 0$. Note that this entails a lower bound on T as follows

$$T \geq \frac{\|\boldsymbol{\theta}_T\|_2}{cd^\ell r^\ell},$$

which implies a lower bound on the running time given by

$$\max\{r, T\} \geq \max\left\{r, \frac{\|\boldsymbol{\theta}_T\|_2}{cd^\ell r^\ell}\right\} \geq \frac{1}{d^\ell} \min\left\{1, \frac{1}{c}\right\} \cdot \max\left\{r, \frac{\|\boldsymbol{\theta}_T\|_2}{r^\ell}\right\}.$$

By Lemma 14, it will suffice to show that

$$\max\left\{r, \frac{\|\boldsymbol{\theta}_T\|_2}{r^\ell}\right\} \in \mathcal{F}\left(d, \frac{1}{\sqrt{\epsilon}}\right).$$

To this end, we first invoke Thm. 5 which by our assumption guarantees that

$$\max\left\{r, \|\boldsymbol{\theta}_T\|_2\right\} \in \mathcal{F}(d, m),$$

where m is the largest integer such that $\epsilon \leq \frac{1}{196(m+1)^2}$, implying that $m \geq 2$ by our assumption $\epsilon \leq \frac{1}{400}$. We now split our analysis into several cases.

- First assume that $r > \|\boldsymbol{\theta}_T\|_2$. Then in this case we immediately have that

$$\max\left\{r, \frac{\|\boldsymbol{\theta}_T\|_2}{r^\ell}\right\} = r = \max\left\{r, \|\boldsymbol{\theta}_T\|_2\right\} \in \mathcal{F}(d, m).$$

- Assuming $r \leq \|\boldsymbol{\theta}_T\|_2$ and $\sqrt{\|\boldsymbol{\theta}_T\|_2} r^{-\ell} \geq 1$, we have

$$\max\left\{r, \frac{\|\boldsymbol{\theta}_T\|_2}{r^\ell}\right\} \geq \max\left\{r, \sqrt{\|\boldsymbol{\theta}_T\|_2}\right\} \geq \sqrt{\|\boldsymbol{\theta}_T\|_2} \in \mathcal{F}(d, m),$$

where the inclusion follows from Lemma 14.

- Assuming $r \leq \|\boldsymbol{\theta}_T\|_2$ and $\sqrt{\|\boldsymbol{\theta}_T\|_2} r^{-\ell} < 1$, we have

$$\max\left\{r, \frac{\|\boldsymbol{\theta}_T\|_2}{r^\ell}\right\} \geq r > \sqrt[2\ell]{\|\boldsymbol{\theta}_T\|_2} \in \mathcal{F}(d,m),$$

where once again the inclusion follows from Lemma 14.

We conclude the proof by recalling that m is the largest integer such that $\frac{1}{196(m+1)^2} \geq \epsilon$ and observing that our bound on ϵ of $\frac{1}{400}$ implies that

$$3 \leq m + 1 \leq \frac{1}{14\sqrt{\epsilon}},$$

from which we have $1 \leq \frac{1}{42\sqrt{\epsilon}}$ and thus

$$2 \leq \frac{1}{21\sqrt{\epsilon}} = \frac{1}{14\sqrt{\epsilon}} - \frac{1}{42\sqrt{\epsilon}} \leq m + 1 \leq \frac{1}{14\sqrt{\epsilon}}.$$

That is, for any $\epsilon \leq \frac{1}{400}$ we have an integer m such that $m = \Theta(1/\sqrt{\epsilon})$ and $\max\{r, T\} \in \mathcal{F}(d, m)$, and therefore from Lemma 14 we get

$$\max\{r, T\} \in \mathcal{F}\left(d, \frac{1}{\sqrt{\epsilon}}\right).$$

Turning to the positive optimization result, we first argue that $\mathcal{D}'_d$ has a radial density which follows from the fact that the sum of two radial distributions is also radial, since each individual distribution is invariant to radial transformations and therefore so is their sum. Next, we prove each item in Assumption 6, where we denote the random variable representing the norm of the distribution $\mathcal{D}'_d$ by X'_d, and its density by γ'_d.

1. Since $\mathcal{D}'_d$ is supported on $\{\mathbf{x} : \|\mathbf{x}\| \leq 2\}$, we have that

$$\int_0^2 \gamma'_d(x) dx = 1.$$

2. By Vardi and Shamir (2020, Lemma A.2) and a straightforward change of variables, we have for all $x \in [0, 2]$ that

$$\gamma'_d(x) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} x^{d-2}(1 - 0.25x^2)^{\frac{d-3}{2}}$$

(where for $x > 2$ the density is 0). Upper bounding the expression above, we compute

$$x^{d-2}(1 - 0.25x^2)^{\frac{d-3}{2}} = (x^2)^{\frac{d-2}{2}}(1 - 0.25x^2)^{\frac{d-3}{2}} \leq (x^2 - 0.25x^4)^{\frac{d-3}{2}} \leq 1,$$

where the last inequality holds by the assumption $d \geq 3$ and the fact that $\max_{x \in [0,2]} x^2 - 0.25x^4 = 1$. Using Eq. 5.6.4 in DLMF, we have $\Gamma\left(\frac{d}{2}\right)/\Gamma\left(\frac{d-1}{2}\right) < \sqrt{\frac{d}{2}}$, allowing us to deduce an upper bound of

$$\gamma'_d(x) \leq \sqrt{d}$$

for all $x \in [0, 2]$ and $d \geq 3$. We remark that $\gamma'_d(x) > 0$ holds for all $x \in [0, 2]$ except for the boundary, but this poses no problem since we may without loss of generality mix the distribution with another with positive density at the boundary while keeping the analysis unchanged (see the explanation in the beginning of Appendix B.1 for justification).

3. Once again, using the fact that $\mathcal{D}'_d$ is supported on $\{\mathbf{x} : \|\mathbf{x}\| \leq 2\}$, we have for all $x \geq 2$ that

$$\mathbb{P}\left[X'_d \geq x\right] = 0 \leq \frac{1}{x}.$$

$\blacksquare$

Appendix C. Proofs from Sec. 3 – Approximation Lower Bounds

C.1. Proof of Thm. 4

Proof Our proof builds on the following main result in Eldan and Shamir (2016), which is restated here for completeness in a slightly different manner to fit our setting.

Theorem 16 (Eldan and Shamir (2016)) *The following holds for some universal constants $c_1, c_2, c_3, c_4 > 0$, integer $\gamma \geq 1$, constant $\alpha \geq 1$ and any network employing an activation function satisfying Assumptions 1 and 2 in Eldan and Shamir (2016): For all $d \geq c_1$, there exist constants $\lambda_d^{(1)}, \ldots, \lambda_d^{(k)} \in [1, 2]$ and $\epsilon_d^{(1)}, \ldots, \epsilon_d^{(k)} \in \{-1, 1\}$ where $k = \gamma d^2$, such that for any neural network N of depth 2 and width at most $c_3 \exp(c_4 d)$ we have*

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_d}\left[\left(\sum_{i=1}^{k} \epsilon_d^{(i)} \mathbb{1}\left\{\|\mathbf{x}\| \leq \alpha\sqrt{d}\lambda_d^{(i)}\right\} - N(\mathbf{x})\right)^2\right] \geq c_2,$$

where the density of $\mathcal{D}_d$ is given by

$$\hat{\mu}(\mathbf{x}) := \left(\frac{R_d}{\|\mathbf{x}\|}\right)^d J_{d/2}^2(2\pi R_d \|\mathbf{x}\|).$$

Writing the implication of the above theorem in integral form, we have that

$$\int_{\mathbb{R}^d} \left(\sum_{i=1}^{k} \epsilon_d^{(i)} \mathbb{1}\left\{\|\mathbf{x}\| \leq \alpha\sqrt{d}\lambda_d^{(i)}\right\} - N(\mathbf{x})\right)^2 \hat{\mu}(\mathbf{x})d\mathbf{x} \geq c_2.$$

We now perform a change of variables $\mathbf{x} = \alpha\sqrt{d}\mathbf{z}$, $d\mathbf{x} = \frac{1}{(\alpha\sqrt{d})^d}d\mathbf{z}$ to obtain

$$\int_{\mathbb{R}^d} \left(\sum_{i=1}^{k} \epsilon_d^{(i)} \mathbb{1}\left\{\|\mathbf{z}\| \leq \lambda_d^{(i)}\right\} - N(\alpha\sqrt{d}\mathbf{z})\right)^2 \frac{1}{(\alpha\sqrt{d})^d}\hat{\mu}(\alpha\sqrt{d}\mathbf{z})d\mathbf{z} \geq c_2.$$

Since $N(\alpha\sqrt{d}\mathbf{z})$ computes the same function as the neural network $N'(\mathbf{z})$ where the weights of the neurons in the hidden layer are multiplied by $\alpha\sqrt{d}$, and by defining the measure $\bar{\mu}(\mathbf{x}) := \frac{1}{(\alpha\sqrt{d})^d}\hat{\mu}(\alpha\sqrt{d}\mathbf{x})$ (note that this is indeed a measure as readily seen by the change of variables $\mathbf{y} = \alpha\sqrt{d}\mathbf{x}$, $d\mathbf{y} = \frac{1}{(\alpha\sqrt{d})^d}d\mathbf{x}$, which yields $\int \bar{\mu}(\mathbf{x})d\mathbf{x} = \int \frac{1}{(\alpha\sqrt{d})^d}\hat{\mu}(\alpha\sqrt{d}\mathbf{x})d\mathbf{x} = \int \hat{\mu}(\mathbf{y})d\mathbf{y} = 1$), we can rewrite the above as

$$\int_{\mathbb{R}^d} \left(\sum_{i=1}^{k} \epsilon_d^{(i)} \mathbb{1}\left\{\|\mathbf{x}\| \leq \lambda_d^{(i)}\right\} - N(\mathbf{x})\right)^2 \bar{\mu}(\mathbf{x})d\mathbf{x} \geq c_2, \tag{12}$$

for any depth 2 neural network N of width at most $c_3 \exp(c_4 d)$. Fix any $d \geq c_1$. We assume by contradiction that for all $i \in [k]$, we have a depth 2 neural network N_i of width less than $c_3 \exp(c_4 d)/(\gamma d^2)$ which satisfies

$$\int_{\mathbb{R}^d} \left(\mathbb{1}\left\{ \|\mathbf{x}\| \leq \lambda_d^{(i)} \right\} - N_i(\mathbf{x}) \right)^2 \bar{\mu}(\mathbf{x})d\mathbf{x} < \frac{c_2}{\gamma^2 d^4}.$$

Letting $\|f\|_{L(\bar{\mu})} := \sqrt{\int_{\mathbb{R}^d} f^2(\mathbf{x})\bar{\mu}(\mathbf{x})d\mathbf{x}}$, we now compute

$$\left\| \sum_{i=1}^{k} \epsilon_d^{(i)} \mathbb{1}\left\{ \|\mathbf{x}\| \leq \lambda_d^{(i)} \right\} - \sum_{i=1}^{k} \epsilon_d^{(i)} N_i(\mathbf{x}) \right\|_{L(\bar{\mu})} \leq \sum_{i=1}^{k} \left| \epsilon_d^{(i)} \right| \cdot \left\| \mathbb{1}\left\{ \|\mathbf{x}\| \leq \lambda_d^{(i)} \right\} - N_i(\mathbf{x}) \right\|_{L(\bar{\mu})}$$

$$< \sum_{i=1}^{k} \left| \epsilon_d^{(i)} \right| \cdot \sqrt{\frac{c_2}{\gamma^2 d^4}} \leq \sqrt{c_2},$$

where the first inequality is the triangle inequality and the second is by our assumption. But since $\sum_{i=1}^{k} \epsilon_d^{(i)} N_i(\mathbf{x})$ can be computed by a depth 2 neural network of width less than $c_3 \exp(c_4 d)$ by concatenating the k networks N_i, $i = 1, \ldots, k$, the square of the above contradicts Eq. (12). We thus have that for all $d \geq c_1$, there exists $i \in [k]$ such that for any depth 2 neural network N of width less than $c_3 \exp(c_4 d)/(\gamma d^2)$, we have

$$\int_{\mathbb{R}^d} \left(\mathbb{1}\left\{ \|\mathbf{x}\| \leq \lambda_d^{(i)} \right\} - N(\mathbf{x}) \right)^2 \bar{\mu}(\mathbf{x})d\mathbf{x} \geq \frac{c_2}{\gamma^2 d^4},$$

which for the measure μ defined in Eq. (4) implies

$$\int_{\mathbb{R}^d} \left(\mathbb{1}\left\{ \|\mathbf{x}\| \leq \lambda_d^{(i)} \right\} - N(\mathbf{x}) \right)^2 \mu(\mathbf{x})d\mathbf{x} \geq \frac{c_2}{2\gamma^2 d^4}.$$

Scaling c_2 by $0.5\gamma^{-2}$ and since we can take $c_3', c_4' > 0$ small enough so that

$$\frac{c_3 \exp(c_4 d)}{\gamma d^2} \geq c_3' \exp(c_4' d)$$

for all $d \geq c_1$, the statement of the theorem follows. $\blacksquare$

C.2. Proof of Thm. 5

To prove the theorem, we will first need the following lemma and proposition. The lemma below shows that functions that oscillate $2m + 2$ many times cannot be approximated well using polynomials of degree at most m with respect to a certain weight function.

Lemma 17 *For any polynomial p of degree at most m, we have*

$$\int_{-1}^{1} \left(p(x) - \bar{f}_m(x) \right)^2 \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} (1 - x^2)^{\frac{d-3}{2}} dx \geq \frac{1}{36}.$$

29

Proof Since the integrand is non-negative, we can lower bound the integral in the lemma by

$$\int_{-\frac{1}{2\sqrt{d}}}^{0} \left(p(x) - \bar{f}_m(x)\right)^2 \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}(1-x^2)^{\frac{d-3}{2}}dx.$$

Assuming $d \geq 4$, then $\Gamma\left(\frac{d}{2}\right)/\Gamma\left(\frac{d-1}{2}\right) \geq \frac{\sqrt{d}}{2}$ (see Eq. 5.6.4 in DLMF) and since

$$\frac{1}{\sqrt{\pi}} \inf_{x \in [-0.5d^{-0.5}, 0]} (1-x^2)^{(d-3)/2} \geq \frac{4}{9}$$

for all $d \geq 4$ we have that the above integral is lower bounded by

$$\frac{2\sqrt{d}}{9} \int_{-\frac{1}{2\sqrt{d}}}^{0} \left(p(x) - \bar{f}_m(x)\right)^2 dx.$$

By its definition, $\bar{f}_m$ alternates signs at least $2m$ times on the domain of integration, and since a degree m polynomial can only change signs at most m times, there are at least m intervals each of which of length $\frac{1}{4m\sqrt{d}}$ (we view the intervals at the boundary of the domain of integration as a single interval of this length whenever $\sqrt{d}$ is not an integer) where p does not change its sign. Moreover, on at least $m/2$ of these intervals the signs of p and $\bar{f}_m$ are opposite, and the integral on these intervals is lower bounded by their length $\frac{1}{4m\sqrt{d}}$. Combining this with our previously derived lower bound concludes the lemma. ∎

The following proposition shows that a depth 2 neural network cannot approximate a certain dot-product function that oscillates $2m + 2$ many times, unless its width or the Euclidean norm of its weights is large.

Proposition 18 *Suppose that a neural network $N_{\boldsymbol{\theta}}(\cdot)$ with architecture defined in Eq. (1) satisfies*

$$r, \|\boldsymbol{\theta}\|_2 \leq \frac{1}{10(1 + \sqrt{C_\sigma})} N_{d,m}^{\alpha'_\sigma/6},$$

where

$$N_{d,m} := \frac{(2m + d - 2)(m + d - 3)!}{m!(d-2)!},$$

$\alpha'_\sigma := \begin{cases} 1 & \alpha_\sigma \leq 1 \\ \alpha_\sigma^{-1} & \alpha_\sigma > 1 \end{cases}$, *and C_σ, α_σ are defined in Assumption 1 and $\boldsymbol{\theta}$ are all the trainable weights of $N_{\boldsymbol{\theta}}(\cdot)$ in vectorized form. Then*

$$\left\|N_{\boldsymbol{\theta}}(\mathbf{x}) - \bar{f}_m(\langle\mathbf{x}_1, \mathbf{x}_2\rangle)\right\|_{L_2(\varphi_d)}^2 > \frac{1}{49}$$

Proof By Daniely (2017, Thm. 4), we have that

$$\left\|N_{\boldsymbol{\theta}}(\mathbf{x}) - \bar{f}_m(\langle\mathbf{x}_1, \mathbf{x}_2\rangle)\right\|_{L_2(\varphi_d)}^2$$
$$\geq \left\|\mathcal{P}_m\bar{f}_m\right\|_{L_2(\varphi_d)} \left(\left\|\mathcal{P}_m\bar{f}_m\right\|_{L_2(\varphi_d)} - \frac{2\sum_{j=1}^{r}|w_j| \cdot \|\sigma(\langle\mathbf{u}_j, \mathbf{x}\rangle + b_j)\|_{L_2(\varphi_d)} + 2|b_0|}{\sqrt{N_{d,m}}}\right),$$

where $\mathcal{P}_m \bar{f}_m$ is the best degree m polynomial approximation of $x \mapsto \bar{f}_m(x)$ with respect to the density $w(x) := \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}(1-x^2)^{\frac{d-3}{2}}$. By Lemma 17 we have

$$\left\|N(\mathbf{x}) - \bar{f}_m(\langle\mathbf{x}_1, \mathbf{x}_2\rangle)\right\|_{L_2(\varphi_d)}^2 \geq \frac{1}{6}\left(\frac{1}{6} - \frac{2\sum_{j=1}^{r}|w_j|\cdot\|\sigma\left(\langle\mathbf{u}_j,\mathbf{x}\rangle + b_j\right)\|_{L_2(\varphi_d)} + 2|b_0|}{\sqrt{N_{d,m}}}\right).$$
(13)

We now bound the norm of the function implemented by a neuron

$$\|\sigma\left(\langle\mathbf{u}_j,\mathbf{x}\rangle + b_j\right)\|_{L_2(\varphi_d)} = \sqrt{\int_{\mathbb{R}^d}\left(\sigma\left(\langle\mathbf{u}_j,\mathbf{x}\rangle + b_j\right)\right)^2\varphi_d(\mathbf{x})d\mathbf{x}}$$

$$\leq \sqrt{\int_{\mathbb{R}^d}C_\sigma\left(1 + |2\|\mathbf{u}_j\| + |b_j||^{\alpha_\sigma}\right)^2\varphi_d(\mathbf{x})d\mathbf{x}}$$

$$= \sqrt{C_\sigma}\left(1 + |2\|\mathbf{u}_j\| + |b_j||^{\alpha_\sigma}\right)$$

$$\leq \sqrt{C_\sigma}\left(1 + (3\|\boldsymbol{\theta}\|)^{\alpha_\sigma}\right),$$

where the first inequality holds due to Cauchy-Schwartz and Assumption 1. From the above and an additional application of Cauchy-Schwartz we get

$$2\sum_{j=1}^{r}|w_j|\cdot\|\sigma\left(\langle\mathbf{u}_j,\mathbf{x}\rangle + b_j\right)\|_{L_2(\varphi_d)} + 2|b_0| \leq 2r\sqrt{C_\sigma}\|\boldsymbol{\theta}\|\left(1 + (3\|\boldsymbol{\theta}\|)^{\alpha_\sigma}\right) + 2\|\boldsymbol{\theta}\|$$

$$\leq 2r\|\boldsymbol{\theta}\|\left(1 + \sqrt{C_\sigma} + 3^{\alpha_\sigma}\|\boldsymbol{\theta}\|^{\alpha_\sigma}\right).$$

By our assumption in the proposition statement and the definition of α_σ' which entails $\alpha_\sigma', \alpha_\sigma\alpha_\sigma' \leq 1$, we have that $2r\|\boldsymbol{\theta}\| \leq \frac{1}{50}N_{d,m}^{1/3}$ and $1 + \sqrt{C_\sigma} + 3^{\alpha_\sigma}\|\boldsymbol{\theta}\|^{\alpha_\sigma} \leq 2N_{d,m}^{1/6}$. Plugging these two inequalities in Eq. (13) and simplifying, the proposition follows. ∎

With the above lemma and proposition, we are ready to prove the theorem.

Proof [of Thm. 5] Let $\mathcal{D}_d'$ denote the distribution on $\mathbb{R}^d$ defined in Eq. (5). For natural $m \geq 1$, consider the function

$$f_m(z) := \begin{cases} \text{sign}\left(\sin\left(2\pi m\sqrt{d}z^2\right)\right) & z^2 \in \left[2 - \frac{1}{\sqrt{d}}, 2\right] \\ 0 & \text{otherwise} \end{cases},$$

where we define

$$\text{sign}(z) := \begin{cases} 1 & z \geq 0 \\ -1 & z < 0 \end{cases}.$$

By the definition of $f_m(\cdot)$, we have that we can write

$$f_m(\|\mathbf{x}\|_2) = \sum_{i=1}^{2m+2}\epsilon_d^{(i)}\mathbb{1}\left\{\|\mathbf{x}\|_2 \leq \lambda_d^{(i)}\right\}$$

for appropriately chosen $\epsilon_d^{(1)}, \dots, \epsilon_d^{(2m+2)} \in \{-1, 1\}$ and $\lambda_d^{(1)}, \dots, \lambda_d^{(2m+2)} \in [1, 2]$. Suppose that for all $i \in [2m + 2]$ there exists a neural network $N_{\boldsymbol{\theta}_i}(\cdot)$ of width r_i, having weights $\boldsymbol{\theta}_i$, and with architecture defined by Eq. (1) such that

$$\left\| N_{\boldsymbol{\theta}_i}^{(i)}(\mathbf{x}) - \mathbb{1}\left\{ \|\mathbf{x}\|_2 \leq \lambda_d^{(i)} \right\} \right\|_{L_2(\varphi_d)} \leq \frac{1}{14(m+1)}.$$

Then we can construct a depth 2 neural network $N_{\boldsymbol{\theta}'}(\cdot)$ of width r' which is at most $(2m + 2)r$, having weights satisfying $\left\| \boldsymbol{\theta}' \right\| \leq \sqrt{2m + 2} \cdot \max_i \|\boldsymbol{\theta}_i\|$ such that

$$N_{\boldsymbol{\theta}'}(\mathbf{x}) = \sum_{i=1}^{2m+2} \epsilon_d^{(i)} N_{\boldsymbol{\theta}_i}^{(i)}(\mathbf{x}),$$

by taking all the neural networks $N_{\boldsymbol{\theta}_i}$ and concatenating them appropriately. From the above and the triangle inequality we have

$$\|N_{\boldsymbol{\theta}'}(\mathbf{x}) - f_m(\|\mathbf{x}\|_2)\|_{L_2(\varphi_d)} \leq \sum_{i=1}^{2m+2} \left| \epsilon_d^{(i)} \right| \cdot \left\| N_{\boldsymbol{\theta}_i}^{(i)}(\mathbf{x}) - \mathbb{1}\left\{ \|\mathbf{x}\|_2 \leq \lambda_d^{(i)} \right\} \right\|_{L_2(\varphi_d)} \leq \frac{1}{7}. \quad (14)$$

Next, we compute for any $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ such that $\mathbf{x}_1, \mathbf{x}_2$ have unit norm and $\|\mathbf{x}\|_2^2 \in \left[2 - \frac{1}{\sqrt{d}}, 2\right]$ to obtain that

$$f_m(\|\mathbf{x}\|_2) = \operatorname{sign}\left(\sin\left(2\pi m \sqrt{d} \, \|\mathbf{x}\|_2^2\right)\right) = \operatorname{sign}\left(\sin\left(4\pi m \sqrt{d}(1 + \langle \mathbf{x}_1, \mathbf{x}_2 \rangle)\right)\right),$$

for any unit $\mathbf{x}_1, \mathbf{x}_2$ such that $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle \in \left[-\frac{1}{2\sqrt{d}}, 0\right]$. Define

$$\bar{f}_m(z) := \begin{cases} \operatorname{sign}\left(\sin\left(4\pi m \sqrt{d}(1 + z)\right)\right) & z \in \left[-\frac{1}{2\sqrt{d}}, 0\right] \\ 0 & \text{otherwise} \end{cases}.$$

By the above we have

$$f_m(\|\mathbf{x}\|_2) = \bar{f}_m(\langle \mathbf{x}_1, \mathbf{x}_2 \rangle),$$

which when plugged in Eq. (14) yields

$$\left\| N_{\boldsymbol{\theta}'}(\mathbf{x}) - \bar{f}_m(\langle \mathbf{x}_1, \mathbf{x}_2 \rangle) \right\|_{L_2(\varphi_d)} \leq \frac{1}{7}.$$

By Proposition 18 we have that

$$\max\left\{ r', \left\| \boldsymbol{\theta}' \right\|_2 \right\} > \frac{1}{10(1 + \sqrt{C_\sigma})} N_{d,m}^{\alpha'_\sigma/6},$$

implying

$$\max_i \left\{ \max\left\{ r_i, \|\boldsymbol{\theta}_i\|_2 \right\} \right\} > \frac{1}{20(m+1)(1 + \sqrt{C_\sigma})} N_{d,m}^{\alpha'_\sigma/6} \geq \frac{1}{40(1 + \sqrt{C_\sigma})m} N_{d,m}^{\alpha'_\sigma/6}. \quad (15)$$

Lower bounding $\frac{1}{m}N_{d,m}$, we first assume that $m > d$ to obtain

$$
\frac{1}{m}N_{d,m} \geq \frac{1}{m}\binom{m+d-2}{d-2} \geq \frac{1}{m}\left(1+\frac{m}{d-2}\right)^{d-2} \geq \frac{1}{m}\left(1+\frac{m}{d}\right)^{d/2}
$$
$$
= \frac{1}{m}\left(1+\frac{m}{d}\right)^{d/4}\cdot\left(1+\frac{m}{d}\right)^{d/4} \geq \frac{1}{m}\left(1+\frac{m}{4}\right)\cdot\left(1+\frac{m}{d}\right)^{d/4}
$$
$$
\geq \frac{1}{4}\left(1+\frac{m}{d}\right)^{d/4} = \frac{1}{4}\exp\left(\frac{d}{4}\ln\left(1+\frac{m}{d}\right)\right) \geq \frac{1}{4}\exp\left(\frac{d}{8}\ln\left(\frac{m}{d}+2\right)\right).
$$

In the above; in the first line, the second inequality follows from the inequality $\binom{n}{k} \geq (n/k)^k$ which holds for all natural $n \geq k$ and the second inequality follows from the fact that $x \mapsto (1+a/x)^x$ is increasing in $x \geq 0$ for all $a > 0$ and by assuming $d \geq 4$ which entails $d - 2 \geq d/2$; in the second line the inequality follows from Bernoulli's inequality; in the third line, the last inequality follows from the fact that $0.5\ln(2+x)/\ln(1+x) \leq 1$ for all $x \geq 1$ and since we assume $m > d$. Now assuming that $m \leq d$, we have

$$
\frac{1}{m}N_{d,m} \geq \frac{1}{d}N_{d,m} \geq \frac{1}{d}\binom{m+d-2}{m},
$$

from which by symmetry and using the same reasoning as in the previous case we get

$$
\frac{1}{m}N_{d,m} \geq \frac{1}{4}\exp\left(\frac{m}{8}\ln\left(\frac{d}{m}+2\right)\right).
$$

Combining these two cases together we obtain a lower bound of

$$
\frac{1}{m}N_{d,m} \geq \frac{1}{4}\exp\left(\frac{1}{8}\min\left\{m\ln\left(\frac{d}{m}+2\right), d\ln\left(\frac{m}{d}+2\right)\right\}\right),
$$

which when plugged back in Eq. (15), along with Lemma 14, implies the existence of some $i \in [2m+2]$ such that

$$
\max\{r_i, \|\boldsymbol{\theta}_i\|_2\} \geq c_1\exp\left(c_2\min\left\{m\ln\left(\frac{d}{m}+2\right), d\ln\left(\frac{m}{d}+2\right)\right\}\right),
$$

for some $c_1, c_2 > 0$, concluding the proof of the theorem. ∎

Appendix D. Proofs from Subsection A.1

To prove Thm. 9, we will need the following auxiliary lemmas and propositions. The following technical lemma provides useful bounds on the special function Owen's T which we encounter when computing the expectation of truncated random features (see Owen (1956) for more information about this function).

Lemma 19 *Let*

$$
T(h,a) := \frac{1}{2\pi}\int_0^a \frac{\exp\left(-\frac{1}{2}h^2(1+t^2)\right)}{1+t^2}dt \tag{16}
$$

denote Owen's T function. Then we have for any $(h,a) \in [0,\infty)^2$ that

$$
\frac{1}{2\pi}\exp\left(-\frac{1}{2}h^2\left(1+a^2\right)\right)\arctan(a) \leq T(h,a) \leq \frac{1}{2\pi}\exp\left(-\frac{1}{2}h^2\right)\arctan(a).
$$

Proof Starting with the lower bound, since $\exp\left(-\frac{1}{2}h^2\left(1+t^2\right)\right)$ is minimized at $t = a$ in the domain of integration $[0, a]$ and since the integrand is positive, we can lower bound $T(h, a)$ by

$$\frac{1}{2\pi}\int_0^a \frac{\exp\left(-\frac{1}{2}h^2(1+a^2)\right)}{1+t^2}dt = \frac{\exp\left(-\frac{1}{2}h^2(1+a^2)\right)}{2\pi}\int_0^a \frac{1}{1+t^2}dt$$

$$= \frac{\exp\left(-\frac{1}{2}h^2(1+a^2)\right)}{2\pi}\arctan(a),$$

where the upper bound follows analogously from the exponent being maximized at 0 over the integration domain. ∎

The following lemma establishes some useful properties of the distribution of the random features that are produced by i.i.d. normal random variables.

Lemma 20 *Suppose that $U_j \sim \mathcal{N}(0, \sigma^2 \cdot I_d)$ and $B_j \sim \mathcal{N}(0, \sigma^2)$ for all $j \in [r]$. Then the pre-activation output of a neuron in the hidden layer conditioned on the input's norm $\|X\|$ attaining the value x, given by the random variable $(\langle U_j, X\rangle + B_j \mid \|X\| = x)$,[3] follows a normal distribution with variance $\sigma^2(1 + x^2)$. Moreover, under the conditioning on the input's norm, all such outputs for all the neurons are mutually independent random variables.*

Proof Fix some $\mathbf{x} = (x_1, \ldots, x_d)$ such that $\|\mathbf{x}\| = x$, then we have that

$$\langle U_j, \mathbf{x}\rangle + B_j = \sum_{k=1}^{d} U_{j,k}x_k + B_j$$

is normally distributed with variance $\sigma^2(1 + x^2)$ since the sum of independent normal random variables is normally distributed with the sum of the expectations and variances as its parameters. Moreover, Since the coordinate in each weight $U_{j,k}$ and B_j are mutually independent by our initialization assumption it is readily seen that $\mathbb{E}\left[\prod_{j\in A}\left(\langle U_j, \mathbf{x}\rangle + B_j\right)\right] = 0$ for any subset of indices $A \subseteq \{1, \ldots, r\}$. Since uncorrelatedness implies independence for multivariate normal variables, the lemma follows. ∎

The following proposition computes the expectation of the random features produced under Assumption 7 when we truncate the bias terms.

Proposition 21 *Suppose that the random variables $U \in \mathbb{R}^d$ and $B \in \mathbb{R}$ are distributed according to Assumption 7. Then if B is truncated from below at 0 we have for all $\mathbf{x} \in \mathbb{R}^d$ that*

$$\mathbb{E}_{U,B}\left[\mathrm{erf}\left(\langle U, \mathbf{x}\rangle + B\right)\right] = \frac{2}{\pi}\arctan\left(\frac{1}{\sqrt{2 + \|\mathbf{x}\|^2}}\right).$$

Further, if B is instead truncated from below at $1/\sqrt{2}$ we have for all $\mathbf{x} \in \mathbb{R}^d$ that

$$\mathbb{E}_{U,B}\left[\mathrm{erf}\left(\langle U, \mathbf{x}\rangle + B\right)\right] = \frac{4}{1 - \mathrm{erf}(1)}\left(T\left(\sqrt{2}, \frac{1}{\sqrt{2 + \|\mathbf{x}\|^2}}\right)\right).$$

---

3. To avoid ambiguity stemming from the Borel-Kolmogorov paradox, we formally define this random variable using the limit $\lim_{\epsilon\to 0}\left(\langle U_j, X\rangle + B_j \mid \|X\| \in (x - \epsilon, x + \epsilon)\right)$. Note that this is a valid definition since by Assumption 6(2) $\|X\|$ has a strictly positive density.

Proof We will use of the following identities which appear in Owen (1980):

$$\int_{-\infty}^{\infty} \Phi(c+dz)\phi(z)dz = \Phi\left(\frac{c}{\sqrt{1+d^2}}\right), \tag{17}$$

$$\int \Phi(cz)\phi(z)dz = -T(z,c) + \frac{1}{2}\Phi(z), \tag{18}$$

$$T(0,a) = \frac{1}{2\pi}\arctan(a), \tag{19}$$

where $T(\cdot,\cdot)$ is Owen's T function defined in Eq. (16), $\phi(z) = (2\pi)^{-0.5}\exp(-0.5z^2)$ is the density function of a standard normal random variable and $\Phi(z)$ is its cumulative distribution function.

Let U, B as in the proposition statement. We have by Lemma 20 that $\langle U, \mathbf{x}\rangle = \|\mathbf{x}\|^2 \cdot Y$ where $Y \sim \mathcal{N}(0, \sigma^2)$. We compute the expectation iteratively, starting with the expectation over Y by using the law of total expectation as follows

$$\mathbb{E}_{Y,B}\left[\text{erf}\left(\|\mathbf{x}\|^2 Y + B\right)\right] = \mathbb{E}_B\left[\mathbb{E}_{Y|B}\left[\text{erf}\left(\|\mathbf{x}\|^2 Y + b|B = b\right)\right]\right]. \tag{20}$$

Applying the law of unconscious statistician to Y, the inner expectation above equals

$$\int_{-\infty}^{\infty} \text{erf}\left(\|\mathbf{x}\|^2 y + b\right)\frac{1}{\sigma}\phi\left(\frac{y}{\sigma}\right)dy,$$

which by a simple change of variables, the identity $\text{erf}(z) = 2\Phi(\sqrt{2}z) - 1$ and Eq. (17), can be simplified to

$$\int_{-\infty}^{\infty} \text{erf}\left(\|\mathbf{x}\|\sigma y + b\right)\phi(y)\,dy = 2\int_{-\infty}^{\infty} \Phi\left(\sqrt{2}\|\mathbf{x}\|\sigma y + \sqrt{2}b\right)\phi(y)\,dy - \int_{-\infty}^{\infty}\phi(y)dy$$

$$= 2\int_{-\infty}^{\infty} \Phi\left(\sqrt{2}\|\mathbf{x}\|\sigma y + \sqrt{2}b\right)\phi(y)\,dy - 1$$

$$= 2\Phi\left(\frac{\sqrt{2}b}{\sqrt{1 + 2\|\mathbf{x}\|^2\sigma^2}}\right) - 1.$$

To compute the expectation over the bias term, assume we truncate its values (by setting the corresponding weight in the output neuron to zero) to the interval $[\alpha, \beta]$ for some $\beta > \alpha$. Then plugging the above in Eq. (20), we have again from the law of unconscious statistician, the density of a truncated normal and a simple change of variables that this equals

$$\mathbb{E}_B\left[2\Phi\left(\frac{\sqrt{2}B}{\sqrt{1 + 2\|\mathbf{x}\|^2\sigma^2}}\right) - 1\right]$$

$$= 2\int_{\alpha}^{\beta} \Phi\left(\frac{\sqrt{2}b}{\sqrt{1 + 2\|\mathbf{x}\|^2\sigma^2}}\right)\frac{1}{\sigma\cdot(\Phi(\beta/\sigma) - \Phi(\alpha/\sigma))}\phi\left(\frac{b}{\sigma}\right)db - 1$$

$$= \frac{2}{\Phi(\beta/\sigma) - \Phi(\alpha/\sigma)}\int_{\alpha/\sigma}^{\beta/\sigma} \Phi\left(\frac{\sqrt{2}\sigma b}{\sqrt{1 + 2\|\mathbf{x}\|^2\sigma^2}}\right)\phi(b)\,db - 1.$$

From Eq. (18), the above equals

$$\frac{2}{\Phi(\beta/\sigma) - \Phi(\alpha/\sigma)} \left[-T\left(z, \frac{\sqrt{2}\sigma}{\sqrt{1 + 2\|\mathbf{x}\|^2 \sigma^2}} \right) + \frac{1}{2}\Phi(z) \right]_{\alpha/\sigma}^{\beta/\sigma} - 1$$

$$= \frac{2}{\Phi(\beta/\sigma) - \Phi(\alpha/\sigma)} \left(T\left(\frac{\alpha}{\sigma}, \frac{\sqrt{2}\sigma}{\sqrt{1 + 2\|\mathbf{x}\|^2 \sigma^2}} \right) - T\left(\frac{\beta}{\sigma}, \frac{\sqrt{2}\sigma}{\sqrt{1 + 2\|\mathbf{x}\|^2 \sigma^2}} \right) \right)$$

$$= \frac{4}{\mathrm{erf}(\beta/\sqrt{2}\sigma) - \mathrm{erf}(\alpha/\sqrt{2}\sigma)} \left(T\left(\frac{\alpha}{\sigma}, \frac{\sqrt{2}\sigma}{\sqrt{1 + 2\|\mathbf{x}\|^2 \sigma^2}} \right) - T\left(\frac{\beta}{\sigma}, \frac{\sqrt{2}\sigma}{\sqrt{1 + 2\|\mathbf{x}\|^2 \sigma^2}} \right) \right).$$

Plugging $\sigma = \frac{1}{2}$ and taking the limit $\beta \to \infty$, the above reduces to

$$\frac{4}{1 - \mathrm{erf}(\sqrt{2}\alpha)} \left(T\left(2\alpha, \frac{1}{\sqrt{2 + \|\mathbf{x}\|^2}} \right) \right),$$

which for the special cases of $\alpha = 0$ and $\alpha = 1/\sqrt{2}$, by virtue of Eq. (19), equals

$$\frac{2}{\pi} \arctan\left(\frac{1}{\sqrt{2 + \|\mathbf{x}\|^2}} \right),$$

and

$$\frac{4}{1 - \mathrm{erf}(1)} \left(T\left(\sqrt{2}, \frac{1}{\sqrt{2 + \|\mathbf{x}\|^2}} \right) \right).$$

∎

The following proposition establishes some of the crucial properties of the functions we use to approximate ball indicators with random features.

Proposition 22 *Let $\xi \in I$ where $I = [0.45, 0.472]$ and define for $z \geq 0$*

$$f_\xi(z) := \frac{2}{\pi} \arctan\left(\frac{1}{\sqrt{2 + z^2}} \right) - \xi \cdot \frac{4}{1 - \mathrm{erf}(1)} \left(T\left(\sqrt{2}, \frac{1}{\sqrt{2 + z^2}} \right) \right).$$

Then $\{f_\xi\}_{\xi \in I}$ satisfy the following properties:

1. *For all $\xi \in I$, $\lim_{z \to \infty} f_\xi(z) = 0$.*

2. *For all $\lambda \in [1, 2]$ there exists $\xi \in I$ such that $f_\xi(\lambda) = 0$.*

3. *For all $\xi \in I$, $f_\xi(z)$ has a global minimum at*

$$z^* = \sqrt{\frac{1}{\ln\left(\frac{\xi}{1 - \mathrm{erf}(1)} \right) - 1} - 2},$$

where $z^* \in [2.8, 4.2]$ for any $\xi \in I$. Moreover, $f_\xi(z)$ is decreasing for all $z \in (0, z^*)$ and increasing for all $z \in (z^*, \infty)$.

4. *For all $\xi \in I$ and all $z \geq 0$*

$$\frac{2}{\pi} \arctan\left(\frac{1}{\sqrt{2+z^2}}\right) \in [0, 0.5]$$

and

$$\xi \cdot \frac{4}{1 - \mathrm{erf}(1)} \left(T\left(\sqrt{2}, \frac{1}{\sqrt{2+z^2}}\right)\right) \in [0, 0.5].$$

5. *For all $\xi \in I$ and all $z \geq 10$,*

$$f_\xi(z) \leq -\frac{1}{50z}.$$

6. *For all $\xi \in I$ and all $z \in [0.9, 2.1]$,*

$$\left|\frac{\partial}{\partial z} f_\xi(z)\right| \geq \frac{1}{600}.$$

Proof

1. Since $\frac{1}{\sqrt{2+z^2}} \to 0$ as $z \to \infty$, the limit follows from continuity and the definition of $T(\cdot, \cdot)$.

2. Given $\lambda \in [1, 2]$, since $T\left(\sqrt{2}, \frac{1}{\sqrt{2+\lambda^2}}\right) > 0$, we define

$$\xi = g(\lambda) := \frac{1 - \mathrm{erf}(1)}{2\pi} \cdot \frac{\arctan\left(\frac{1}{\sqrt{2+\lambda^2}}\right)}{T\left(\sqrt{2}, \frac{1}{\sqrt{2+\lambda^2}}\right)}.$$

Clearly, $f_\xi(\lambda) = 0$ by its definition. To bound the set of values ξ takes for all $\lambda \in [1, 2]$, we bound the quotient

$$\frac{\arctan\left(\frac{1}{\sqrt{2+\lambda^2}}\right)}{T\left(\sqrt{2}, \frac{1}{\sqrt{2+\lambda^2}}\right)}$$

from below and above. It will suffice to show the monotonicity of this quotient in the domain $[1, 2]$ and compute its values at the boundary. Since $(2 + \lambda^2)^{-0.5}$ is monotone, it will suffice to show monotonicity after the change of variables $z \mapsto (2 + \lambda^2)^{-0.5}$, which yields

$$\frac{\arctan(z)}{T\left(\sqrt{2}, z\right)}.$$

We now need only show that the derivative of the above does not change its sign. Using the fundamental theorem of calculus, consider the numerator of the derivative of the quotient which is given by

$$\frac{1}{1+z^2} T\left(\sqrt{2}, z\right) - \frac{1}{1+z^2} \cdot \frac{1}{2\pi} \arctan(z) \exp(-(1+z^2)),$$

which is non-negative by Lemma 19. We now conclude by verifying that $g(1), g(2) \in I$ using a symbolic computation package.

37

3. Using the chain rule and the fundamental theorem of calculus, we have

$$
\frac{\partial}{\partial z} f_\xi(z) = -\frac{2}{\pi} \cdot \frac{z}{(3+z^2)\sqrt{2+z^2}} + \xi \cdot \frac{4}{2\pi(1-\mathrm{erf}(1))} \cdot \frac{z \exp\left(-\left(1 + \frac{1}{2+z^2}\right)\right)}{(2+z^2)^{1.5}\left(1 + \frac{1}{2+z^2}\right)}
$$
$$
= \frac{2z}{\pi(1-\mathrm{erf}(1))\left(3+z^2\right)\sqrt{2+z^2}}\left(\xi \exp\left(-\left(1 + \frac{1}{2+z^2}\right)\right) - (1 - \mathrm{erf}(1))\right).
$$
(21)

To find the critical point z^*, equate the above to zero and solve for z to get

$$
z^* = \sqrt{\frac{1}{\ln\left(\frac{\xi}{1-\mathrm{erf}(1)}\right) - 1} - 2},
$$

which for any $\xi \in I$ is contained inside the interval $[2.8, 4.2]$. Lastly, since

$$
\exp\left(-\left(1 + \frac{1}{2+z^2}\right)\right)
$$

is increasing in z and since the fraction on the left-hand side of Eq. (21) is positive for positive z, we have from that the derivative is negative in $(0, z^*)$ and positive in (z^*, ∞).

4. We have
$$
0 \leq \frac{2}{\pi}\arctan\left(\frac{1}{\sqrt{2+z^2}}\right) \leq \frac{2}{\pi}\arctan\left(\frac{1}{\sqrt{2}}\right) \leq 0.5,
$$
since clearly the above is positive for all $z \geq 0$. Moreover, by Lemma 19
$$
0 \leq \xi \cdot \frac{4}{1-\mathrm{erf}(1)}\left(T\left(\sqrt{2}, \frac{1}{\sqrt{2+z^2}}\right)\right)
$$
$$
\leq \xi \cdot \frac{4}{1-\mathrm{erf}(1)} \cdot \frac{\exp(-1)}{2\pi}\arctan\left(\frac{1}{\sqrt{2+z^2}}\right)
$$
$$
\leq \xi \cdot \frac{4}{1-\mathrm{erf}(1)} \cdot \frac{\exp(-1)}{2\pi}\arctan\left(\frac{1}{\sqrt{2}}\right) \leq 0.5,
$$

where the first inequality is immediate from the definition of $T(\cdot, \cdot)$ and in the last inequality we used $\xi \leq 0.472$.

5. Using Lemma 19 to lower bound $T\left(\sqrt{2}, \frac{1}{\sqrt{2+z^2}}\right)$, we have

$$
f_\xi(z) \leq \frac{2}{\pi}\arctan\left(\frac{1}{\sqrt{2+z^2}}\right)
$$
$$
-\xi \cdot \frac{4}{1-\mathrm{erf}(1)}\left(\frac{1}{2\pi}\exp\left(-\left(1 + \frac{1}{2+z^2}\right)\right)\arctan\left(\frac{1}{\sqrt{2+z^2}}\right)\right)
$$
$$
= \frac{2}{\pi}\arctan\left(\frac{1}{\sqrt{2+z^2}}\right)\left(1 - \frac{\xi}{1-\mathrm{erf}(1)}\exp\left(-\left(1 + \frac{1}{2+z^2}\right)\right)\right)
$$
$$
\leq \frac{2}{\pi}\arctan\left(\frac{1}{\sqrt{2+z^2}}\right)\left(1 - \frac{0.45}{1-\mathrm{erf}(1)}\exp(-1.01)\right)
$$
$$
\leq -\frac{0.08}{\pi}\arctan\left(\frac{1}{\sqrt{2+z^2}}\right) \leq -\frac{0.064}{\pi\sqrt{2+z^2}} \leq -\frac{0.02}{z},
$$

where the second inequality follows from $\xi \geq 0.45$ and the assumption $z \geq 10$ which implies $-\left(1 + 1/(2 + z^2)\right) \geq -1.01$, the third inequality follows from $1 - \frac{0.45}{1-\mathrm{erf}(1)} \exp\left(-1.01\right) \leq -0.04$, the fourth from $\arctan(z) \geq 0.8z$ for all $z \in [0, 2^{-0.5}]$, and the last inequality follows from $2 \leq 0.02z^2$ for any $z \geq 10$ and from $0.064/(\sqrt{1.02}\pi) \geq 0.02$.

6. Plugging 0.9 or 2.1 into Eq. (21), depending on whether a certain expression in the derivative is monotonically increasing or decreasing, we can lower bound the derivative in absolute value by

$$\left| \frac{2 \cdot 0.9}{\pi(1 - \mathrm{erf}(1))(3 + 2.1^2)\sqrt{2 + 2.1^2}} \left(0.472 \exp\left(-\left(1 + \frac{1}{2 + 2.1^2}\right)\right) - (1 - \mathrm{erf}(1))\right) \right|$$

where we also used the fact that $\xi \leq 0.472$. The lower bound then follows by verifying that the above is at least 0.00169.

■

With the above propositions, we can now prove Thm. 9.

Proof [of Thm. 9] We first sample $\mathbf{w}_0 = (w_{1,0}, \ldots, w_{r,0})$ according to Assumption 7, and then we proceed to define $\mathbf{v}$ in the following manner: Split the r neurons in the hidden layer into two equally sized groups, and set $v_j = r^{-0.75}$ for all the neurons in the first group and $v_j = -\xi_\lambda \cdot r^{-0.75}$ in the second group where $\xi_\lambda \in [0.45, 0.472]$ is a constant that depends on λ and is to be specified later. Next, we describe which undesired random features get truncated (by setting their corresponding v_j to 0). For each hidden neuron in the first group, we truncate bias terms with value below 0, and for the second group we truncate bias terms below $1/\sqrt{2}$. Furthermore, if the sign of any weight w_j in the output neuron's initialization disagrees with the sign of the corresponding value we had set for v_i, then it is also truncated.

Before we proceed to prove the theorem, we will introduce some notation to be used throughout the remainder of the proof. We let A_1, A_2 denote the indices of the neurons that do not get truncated in the first and second groups, respectively. For some $j \in [r]$, we let W_j denote the random variable which is the j-th coordinate of $\mathbf{w}_0$. Fix some $i \in [n]$ and let $Y_{j,k} = \mathrm{erf}\left(\langle U_j, \mathbf{x}_i \rangle + B_{j,a}\right)$ denote the random variable which is the output of the j-th neuron in the k-th group on the i-th data instance, where $B_{j,a}$ is a normal random variable with zero mean and variance 0.25 which is truncated from below at a (i.e. $a = 0$ for $k = 1$ and $a = 2^{-0.5}$ for $k = 2$). Let $\mathbf{v} = (v_1, \ldots, v_r)$ denote the coordinates of $\mathbf{v}$.

Next, we restate the inequalities assumed in the statement of the theorem to be used throughout the proof. Observe that after some calculations, the assumed bound on r implies the following

inequalities (if we also assume $n \geq 4$)

$$r^{-0.25} \leq \frac{1}{12000}, \tag{22}$$

$$10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{r}} \leq 10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}} \leq \frac{1}{500} < 1, \tag{23}$$

$$-\frac{1}{50c_1} \leq -10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}}, \tag{24}$$

$$\sqrt{\frac{\ln\left(\frac{32}{\delta}\right)}{r}} \leq 0.02, \tag{25}$$

$$1.2^{-r} \leq \frac{\delta}{16}. \tag{26}$$

We begin the proof with showing the first item in the theorem statement, where the upper bound follows simply from

$$\|\mathbf{v}\|^2 = \sum_{i=1}^{r} v_i^2 \leq \sum_{i=1}^{r} r^{-1.5} = \frac{1}{\sqrt{r}},$$

where the inequality is by the definition of ξ which satisfies $\xi \in [0.45, 0.472]$ (and appears at a later stage of the proof) and the fact that truncating any coordinate v_j only decreases the norm of $\mathbf{v}$. Lower bounding $\|\mathbf{v}\|$ requires that we first show that the number of neurons that do not get truncated is large enough. To this end, we first observe that the probability of neurons in the first group to not get truncated is exactly $\frac{1}{4}$. This is because we obtain a bias term realization that is negative with probability 0.5, the sign of the corresponding coordinate in $\mathbf{w}_0$ is positive with probability 0.5 by Assumption 7, and since the two events are independent. Similarly, we have that the probability of a neuron in the second group to not get truncated is $0.25 - 0.25 \operatorname{erf}(1) \geq 0.03$. This guarantees that with high probability, the number of untruncated neurons in each group is at least a constant fraction of r: We have from Hoeffding's inequality that for each group that

$$\mathbb{P}\left[\left|\frac{2}{r}\sum_{j=1}^{r/2} \mathbb{1}\left\{v_j \neq 0\right\} - \mathbb{E}\left[\mathbb{1}\left\{v_j \neq 0\right\}\right]\right| > \sqrt{\frac{\ln(32/\delta)}{r}}\right] \leq \frac{\delta}{16},$$

where the expectations for the first and second groups are at least $0.25, 0.03$, respectively, since the expectation of an indicator is the probability of the event. By Eq. (25) this implies

$$\mathbb{P}\left[\frac{2}{r}\sum_{i=1}^{r/2} \mathbb{1}\left\{v_i \neq 0\right\} \geq 0.01\right] \geq 1 - \frac{\delta}{16}.$$

Taking a union bound over the two truncation groups, we have that

$$\mathbb{P}\left[\min\left\{|A_1|, |A_2|\right\} \geq \frac{r}{200}\right] \geq 1 - \frac{\delta}{8}. \tag{27}$$

Assuming the above holds, we can derive the lower bound on $\|\mathbf{v}\|$. Focusing on the neurons in the first truncation group, we have by Assumption 7 that for any $j \in A_1$,

$$(W_j \mid v_j \neq 0) \sim (W_j \mid W_j > 0) \sim |W_j|.$$

That is, given the knowledge that a weight $\mathbf{v}_j$ did not get truncated, its corresponding weight in the hidden layer W_j follows a half-normal distribution. This is true since the bias term B_j and W_j are independent by Assumption 7, and since $v_j \neq 0$ implies that the realization of W_j was positive which by symmetry implies it is half-normally distributed. Since the PDF of a half-normal random variable is $\mathrm{erf}\left(\frac{x}{\sigma\sqrt{2}}\right)$, we have that $\mathbb{P}\left[|W_j| \geq \frac{1}{r}\right] > 0.3$. We can now use Hoeffding's inequality on all the neurons in A_1 (note that they are i.i.d. even after we are given the knowledge that they were not truncated) and obtain

$$\mathbb{P}\left[\left|\frac{1}{|A_1|}\sum_{j \in A_1} \mathbb{1}\left\{|W_j| \geq \frac{1}{r}\right\} - \mathbb{E}\left[\mathbb{1}\left\{|W_j| \geq \frac{1}{r}\right\}\right]\right| > 10\sqrt{\frac{\ln(32/\delta)}{r}}\right] \leq \frac{\delta}{16},$$

which by Eqs. (25,27) implies

$$\mathbb{P}\left[\left|\frac{1}{|A_1|}\sum_{j \in A_1} \mathbb{1}\left\{|W_j| \geq \frac{1}{r}\right\}\right| \leq 0.1\right] \leq \frac{\delta}{16},$$

or equivalently using Eq. (27) again

$$\mathbb{P}\left[\left|\sum_{j \in A_1} \mathbb{1}\left\{|W_j| \geq \frac{1}{r}\right\}\right| > \frac{|A_1|}{10} \geq \frac{r}{2000}\right] \geq 1 - \frac{\delta}{16}. \tag{28}$$

With the above at hand we can lower bound $\langle \mathbf{w}_0, \mathbf{v}\rangle$ in a straightforward manner; by Assumption 7 and the construction of $\mathbf{v}$, we have that $v_j = r^{-0.75}$ and $w_j \geq \frac{1}{r}$ for at least $r/2000$ neurons, therefore

$$\langle \mathbf{w}_0, \mathbf{v}\rangle \geq \frac{r}{2000} \cdot \frac{1}{r^{0.75}} \cdot \frac{1}{r} = \frac{1}{2000r^{0.75}}. \tag{29}$$

Upper bounding $\|\mathbf{w}_0\|^2$, we use a standard bound on Chi-squared random variables to obtain

$$\mathbb{P}\left[\|\mathbf{w}_0\|^2 \geq \frac{2}{r}\right] \leq (2\exp(-1))^{r/2} \leq 1.2^{-r} \leq \frac{\delta}{16},$$

where the last inequality follows from Eq. (26). Equivalently, the above can be stated as

$$\mathbb{P}\left[\|\mathbf{w}_0\|^2 < \frac{2}{r}\right] \geq 1 - \frac{\delta}{16}, \tag{30}$$

which along with Eq. (29) yields

$$\|\mathbf{w}_0\|^2 < \frac{2}{r} \leq \frac{1}{10^3 r^{0.75}} \leq 2\langle \mathbf{w}_0, \mathbf{v}\rangle,$$

where the second inequality follows from Eq. (22). Adding $\|\mathbf{v}\|^2$ to both sides of the above inequality and rearranging, the lower bound on $\|\mathbf{v}\|^2$ follows.

Moving to the second item in the theorem statement, we now wish to define the intervals $I_1, \ldots, I_4$. To this end, we first define the functions

$$f_{\xi_\lambda}(z) := \frac{2}{\pi}\arctan\left(\frac{1}{\sqrt{2+z^2}}\right) - \xi_\lambda \cdot \frac{4}{1-\mathrm{erf}(1)}\left(T\left(\sqrt{2}, \frac{1}{\sqrt{2+z^2}}\right)\right),$$

for $\xi_\lambda \in [0.45, 0.472]$ that depend on λ so that $f_{\xi_\lambda}(\lambda) = 0$ for all $\lambda \in [1, 2]$ (see Item 2 in Proposition 22). By Item 3 in Proposition 22, we can define two inverse functions which we denote by $f_{\xi_\lambda,1}^{-1}(z), f_{\xi_\lambda,2}^{-1}(z)$ on the intervals $(0, z^*)$ and (z^*, ∞), respectively, where it is guaranteed that $z^* \in [2.8, 4.2]$. We can now define the intervals $I_1, \ldots, I_4$. Starting with I_1, I_2, we let

$$I_1 := \left[0, f_{\xi_\lambda,1}^{-1}\left(2r^{-0.25}\right) \right], \quad I_2 := \left(f_{\xi_\lambda,1}^{-1}\left(2r^{-0.25}\right), f_{\xi_\lambda,1}^{-1}\left(-r^{-0.25}\right) \right).$$

We remark that the above intervals are well-defined since by Items 2 and 6 in Proposition 22, we have $f_{\xi_\lambda}(0.9) \geq \frac{1}{6000}$ and $f_{\xi_\lambda}(2.05) \leq -\frac{1}{12000}$, which along with Eq. (22) imply that

$$0.9 \leq f_{\xi_\lambda,1}^{-1}\left(2r^{-0.25}\right) < f_{\xi_\lambda,1}^{-1}\left(-r^{-0.25}\right) \leq 2.05.$$

To define I_3 and I_4, we observe that the above also guarantees that $z^* \notin I_1 \cup I_2$, and from Item 1 in Proposition 22 we have that the image of $f_{\xi_\lambda,2}^{-1}(\cdot)$ is $(0, f_{\xi_\lambda,2}^{-1}(z^*))$. We can thus define I_3 and I_4 as the intervals given by

$$I_3 := \left[f_{\xi_\lambda,1}^{-1}\left(-r^{-0.25}\right), f_{\xi_\lambda,2}^{-1}\left(-10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{r}}\right) \right], \quad I_4 := \left(f_{\xi_\lambda,2}^{-1}\left(-10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{r}}\right), \infty \right).$$

To see why the above intervals are well-defined, we use Item 5 in Proposition 22 and Eq. (23) to deduce that

$$f_{\xi_\lambda}(10) \leq -\frac{1}{500} \leq -10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}} \leq -10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{r}}. \tag{31}$$

Since $f_{\xi_\lambda,2}^{-1}(\cdot)$ is increasing by its definition (and defined for $z = 10$ since $10 \geq z^*$), we can apply it to the above inequality to obtain

$$10 \leq f_{\xi_\lambda,2}^{-1}\left(-10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{r}}\right), \tag{32}$$

which verifies that I_3, I_4 are well-defined.

We now turn to show that with high probability, the fraction of data instances with norm in $I_2 \cup I_4$ decays to zero as r grows. Recall that $I_2 \subseteq [0.9, 2.05]$, we can upper bound the length of I_2 using Item 6 in Proposition 22 by $1800r^{-0.25}$. Using Assumption 6, let q denote the polynomial bounding the density γ_d on I_2. Then

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{D}_d}\left[\|\mathbf{x}\| \in I_2\right] \leq 1800q(d)r^{-0.25}. \tag{33}$$

To bound $\mathbb{P}_{\mathbf{x} \sim \mathcal{D}_d}\left[\|\mathbf{x}\| \in I_4\right]$, apply $f_{\xi_\lambda,2}^{-1}(\cdot)$ to the inequality in Eq. (31) (ignoring the last term) to get

$$f_{\xi_\lambda,2}^{-1}\left(-10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}}\right) \geq 10. \tag{34}$$

We now argue that the constant c_1 from Assumption 6 satisfies

$$c_1 \leq f_{\xi_\lambda,2}^{-1}\left(-10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}}\right). \tag{35}$$

If $c_1 \leq 10$, the above is immediate from Eq. (34). Otherwise, we can use Eq. (24) and Item 5 in Proposition 22 to deduce that

$$f(c_1) \leq -\frac{1}{50c_1} \leq -10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}},$$

where Eq. (35) follows by applying the increasing function $f_{\xi_\lambda,2}^{-1}$ to both sides of the inequality. Using Eq. (35) and Assumption 6, we have the following bound

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_d}\left[\|\mathbf{x}\| \in I_4\right] \leq c_2 \left(f_{\xi_\lambda,2}^{-1}\left(-10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}}\right)\right)^{-1}. \tag{36}$$

To bound the expression above, note that Eq. (23) entails

$$\frac{r^{0.25}}{500\sqrt{\ln\left(\frac{8n}{\delta}\right)}} \geq 10,$$

which allows us to use Item 5 in Proposition 22 to deduce that

$$f_{\xi_\lambda}\left(\frac{r^{0.25}}{500\sqrt{\ln\left(\frac{8n}{\delta}\right)}}\right) \leq -10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}}.$$

Applying the increasing function $f_{\xi_\lambda,2}^{-1}$ to both sides of the inequality above and rearranging while noting that $f_{\xi_\lambda,2}^{-1}(z) > 0$ for all z in its domain which does not change the sign upon division, we have

$$\frac{1}{f_{\xi_\lambda,2}^{-1}\left(-10\sqrt{\ln\left(\frac{8n}{\delta}\right)/\sqrt{r}}\right)} \leq \frac{500\sqrt{\ln\left(\frac{8n}{\delta}\right)}}{r^{0.25}},$$

which when plugged in Eq. (36) and combined with Eq. (33) using a union bound implies

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_d}\left[\|\mathbf{x}\| \in I_2 \cup I_4\right] \leq p := \frac{1800q(d) + 500c_2\sqrt{\ln\left(\frac{8n}{\delta}\right)}}{r^{0.25}}.$$

Since the data instances are i.i.d., we can use Hoeffding's inequality to bound the probability of getting a significantly larger portion of the data in the intervals I_2, I_4 as follows

$$\mathbb{P}_{\mathbf{x}_i\sim\mathcal{D}_d^n}\left[\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left\{\|\mathbf{x}_i\| \in I_2 \cup I_4\right\} - \mathbb{E}\left[\mathbb{1}\left\{\|\mathbf{x}_i\| \in I_2 \cup I_4\right\}\right]\right| > \sqrt{\frac{\ln(8/\delta)}{2n}}\right] \leq \frac{\delta}{4},$$

which by the assumption $\sqrt{\frac{\ln(8/\delta)}{2n}} \leq p$ in the theorem statement implies

$$\mathbb{P}_{\mathbf{x}_i\sim\mathcal{D}_d^n}\left[\sum_{i=1}^n \mathbb{1}\left\{\|\mathbf{x}_i\| \in I_2 \cup I_4\right\} > 2pn\right] \leq \frac{\delta}{4}. \tag{37}$$

To show that with high probability, data instances with norm in $I_1 \cup I_3$ are classified correctly and data instances in $I_2 \cup I_4$ have a bounded misclassification margin, we would need to obtain sufficient concentration of the hidden neurons around their means. We thus now turn to bound the uniform deviation over the data for each group of neurons from its expected value. Using Hoeffding's inequality on the random variables $Y_{i,k}$ for each k separately (which is justified since given $\|\mathbf{x}_j\|$, the outputs of the neurons in the hidden layer are mutually independent by Lemma 20), we have from Item 4 in Proposition 22, that $\mathbb{P}\left[Y_{i,1} \in [0, 0.5]\right] = 1$, and therefore

$$\mathbb{P}\left[\left|\sum_{i \in A_1} r^{-0.75} Y_{i,1} - r^{-0.75} \mathbb{E}\left[Y_{i,1}\right]\right| > \sqrt{\frac{25 \ln(8n/\delta)}{\sqrt{r}}}\right] \leq \frac{\delta}{4n}.$$

Likewise, from a similar bound on the average of $Y_{i,2}$ where $\mathbb{P}\left[\xi_\lambda Y_{i,2} \in [0, 0.5]\right] = 1$ by Item 4 in Proposition 22, we have

$$\mathbb{P}\left[\left|\sum_{i \in A_2} \xi_\lambda r^{-0.75} Y_{i,2} - \xi_\lambda r^{-0.75} \mathbb{E}\left[Y_{i,2}\right]\right| > \sqrt{\frac{25 \ln(8n/\delta)}{\sqrt{r}}}\right] \leq \frac{\delta}{4n}.$$

Combining the above two bounds using a union bound and the triangle inequality and taking another union bound over all the n instances in the data, it then follows from Proposition 21 that

$$\mathbb{P}\left[\sup_{j \in [n]} \left|\mathbf{v}^\top \tilde{\mathbf{x}}_j - r^{0.25} f_{\xi_\lambda}(\|\mathbf{x}_j\|)\right| \leq 10\sqrt{\frac{\ln(8n/\delta)}{\sqrt{r}}}\right] \geq 1 - \frac{\delta}{2}. \tag{38}$$

It only remains to show that $\left[\mathbf{v}^\top \tilde{\mathbf{x}}_j\right]_+^1 = \mathbb{1}\left\{\|\mathbf{x}_j\| \leq \lambda\right\} = y_j$ for any $\mathbf{x}_j$ such that $\|\mathbf{x}_j\| \in I_1 \cup I_3$, and that $\left|\mathbf{v}^\top \tilde{\mathbf{x}}_j\right| \leq 3$ for any $\mathbf{x}_j$ such that $\|\mathbf{x}_j\| \in I_2 \cup I_4$. To this end, we will consider each interval separately:

- Suppose that $\|\mathbf{x}_j\| \in I_1$, then since f_{ξ_λ} is decreasing on I_1 we have that $r^{0.25} f_{\xi_\lambda}(\|\mathbf{x}_j\|) \geq r^{0.25} f_{\xi_\lambda}\left(f_{\xi_{\lambda,1}}^{-1}\left(2r^{-0.25}\right)\right) = 2$. Combining this with Eqs. (23,38) we have

$$\mathbf{v}^\top \tilde{\mathbf{x}}_j \geq r^{0.25} f_{\xi_\lambda}(\|\mathbf{x}_j\|) - 10\sqrt{\frac{\ln(8n/\delta)}{\sqrt{r}}} \geq 2 - 1 = 1.$$

  We thus have

$$\left[\mathbf{v}^\top \tilde{\mathbf{x}}_j\right]_+^1 = 1 = \mathbb{1}\left\{\|\mathbf{x}_j\| \leq \lambda\right\} = y_j.$$

- Suppose that $\mathbf{x}_j \in I_2$, then since f_{ξ_λ} is decreasing on I_2 we have that

$$r^{0.25} f_{\xi_\lambda}(\|\mathbf{x}_j\|) \in (-1, 2),$$

  which together with Eqs. (23,38) implies

$$\mathbf{v}^\top \tilde{\mathbf{x}}_j \in (-2, 3).$$

44

- Suppose that $\mathbf{x}_j \in I_3$. Since $z^* \in I_3$ by Eq. (32), we have that f_{ξ_λ} is decreasing and then increasing on I_3 by Item 3 in Proposition 22. Therefore, f_{ξ_λ} attains its maximum over I_3 at the boundary, and we have

$$f_{\xi_\lambda}(\|\mathbf{x}_j\|) \leq \max \left\{ f_{\xi_\lambda}\left(f_{\xi_\lambda,1}^{-1}\left(-r^{-0.25}\right)\right), f_{\xi_\lambda}\left(f_{\xi_\lambda,2}^{-1}\left(-10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{r}}\right)\right) \right\}$$

$$= \max \left\{ -r^{-0.25}, -10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{r}} \right\}.$$

From the above and Eqs. (23,38), we have

$$\mathbf{v}^\top \tilde{\mathbf{x}}_j \leq r^{0.25} f_{\xi_\lambda}(\|\mathbf{x}_j\|) + 10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}} \leq \max \left\{ -1, -10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}} \right\} + 10\sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{\sqrt{r}}} \leq 0,$$

implying that

$$\left[\mathbf{v}^\top \tilde{\mathbf{x}}_j\right]_+^1 = 0 = \mathbb{1}\left\{\|\mathbf{x}_j\| \leq \lambda\right\} = y_j.$$

- Suppose that $\mathbf{x}_j \in I_4$, then since f_{ξ_λ} is increasing on I_4 we have that

$$r^{0.25} f_{\xi_\lambda}(\|\mathbf{x}_j\|) \in (-1, 0),$$

which together with Eqs. (23,38) implies

$$\mathbf{v}^\top \tilde{\mathbf{x}}_j \in (-2, 1).$$

We conclude the proof of Thm. 9 by applying a union bound to the events in Eqs. (27,28,30,37,38). ∎

Appendix E. Proofs from Subsection A.2

The proof of Thm. 10 relies on the following lemmas. The first lemma below provides a technical inequality which allows us to lower bound the density of the distribution of our random features by a log-concave density.

Lemma 23 *The function*

$$g(x, z) = \frac{\exp\left(\mathrm{erf}^{\text{-}1}(z)^2\left(1 - \frac{2}{1+x^2}\right)\right)}{\sqrt{2 + 2x^2}}$$

restricted to the domain $(x, z) \in \mathbb{R} \times [-1/9, 1/9]$ *is non-increasing in* x.

45

Proof Differentiating with respect to x we obtain (by hand or by using a symbolic computation package)

$$\frac{\partial}{\partial x}g(x,z) = \frac{x\exp\left(\operatorname{erf}^{-1}(z)^2\left(1-\frac{2}{1+x^2}\right)\right)\left(2\operatorname{erf}^{-1}(z)^2 - 0.5(1+x^2)\right)}{0.5^3(2+2x^2)^{2.5}}.$$

The above is non-positive if

$$2\operatorname{erf}^{-1}(z)^2 - 0.5(1+x^2) \leq 0.$$

It can be verified that $\sup_{z\in[-1/9,1/9]}\operatorname{erf}^{-1}(z)^2 \leq 0.01$, and therefore

$$2\operatorname{erf}^{-1}(z)^2 - 0.5(1+x^2) \leq 0.02 - 0.5 < 0,$$

which concludes the proof of the lemma. ∎

The following lemma leverages the analysis done in Lovász and Vempala (2007) to assert that log-concave distributions have well-behaved two-dimensional marginal distributions in any directions.

Lemma 24 *Suppose that $\tilde{X}_1,\ldots,\tilde{X}_r$ are i.i.d. log-concave random variables with $\mathbb{E}[\tilde{X}_1] = 0$ and $\mathbb{E}[\tilde{X}_1^2] = \rho^2$. Then the marginal distribution of the random vector $(\tilde{X}_1,\ldots,\tilde{X}_r)$ on the subspace spanned by any $\mathbf{w}_1 \neq \mathbf{w}_2$ has a (two-dimensional) density $p(\cdot)$ satisfying*

$$\inf_{\mathbf{x}:\|\mathbf{x}\|\leq\frac{1}{9\rho}} p(\mathbf{x}) \geq \rho^2 2^{-16}.$$

Proof Let O denote the orthogonal transformation such that the change of variables $\tilde{\mathbf{x}} \mapsto O\tilde{\mathbf{x}}$ followed by a marginalization of the last $r-2$ coordinates results in the random two-dimensional vector having density p. We first claim that this density is log-concave; this holds since affine transformations and marginalizations preserve log-concavity (Saumard and Wellner, 2014). Next, we have that the distribution of this vector is given by the random variables

$$X_1 = \sum_{j=1}^{r} o_{1,j}\tilde{X}_j, \quad X_2 = \sum_{j=1}^{r} o_{2,j}\tilde{X}_j,$$

where $o_{i,j}$ is the (i,j)-th entry of O. Consider the density $\int_{\mathbb{R}^2} p(\mathbf{x})d\mathbf{x} = 1$. We perform a change of variables $x_1 \mapsto \frac{1}{\rho}y_1$ and $x_2 \mapsto \frac{1}{\rho}y_2$ which entails $d\mathbf{x} = \frac{1}{\rho^2}d\mathbf{y}$, and we have

$$1 = \int_{\mathbb{R}^2} p(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^2} \frac{1}{\rho^2}p\left(\frac{1}{\rho}\mathbf{y}\right)d\mathbf{y}.$$

That is, $\frac{1}{\rho^2}p\left(\frac{1}{\rho}\mathbf{x}\right)$ is the density function of the random vector $\left(\frac{1}{\rho}X_1,\frac{1}{\rho}X_2\right)$. We will show that this random vector is in isotropic position which means it has zero mean and identity covariance matrix: We clearly have for $i \in \{1,2\}$ that

$$\mathbb{E}\left[\frac{1}{\rho}X_i\right] = \sum_{j=1}^{r} \frac{o_{i,j}}{\rho}\mathbb{E}\left[\tilde{X}_j\right] = 0,$$

and to compute the covariance matrix of $\left(\frac{1}{\rho}X_1, \frac{1}{\rho}X_2\right)$, we begin with the diagonal entries to obtain

$$\mathbb{E}\left[\frac{1}{\rho^2}X_i^2\right] = \frac{1}{\rho^2}\left(\sum_{j=1}^{r} o_{i,j}^2 \mathbb{E}\left[\tilde{X}_j^2\right] + \sum_{j_1 \neq j_2} o_{i,j_1} o_{i,j_2} \mathbb{E}\left[\tilde{X}_{j_1}\tilde{X}_{j_2}\right]\right) = \frac{1}{\rho^2}\sum_{j=1}^{r} o_{i,j}^2 \rho^2 = 1,$$

where we used the facts that $\tilde{X}_j$ are i.i.d. and that the rows of O are orthogonal. To compute the off-diagonal entries, we have

$$\mathbb{E}\left[\frac{1}{\rho^2}X_1 X_2\right] = \frac{1}{\rho^2}\left(\sum_{j=1}^{r} o_{1,j} o_{2,j} \mathbb{E}\left[\tilde{X}_j^2\right] + \sum_{j_1 \neq j_2} o_{1,j_1} o_{2,j_2} \mathbb{E}\left[\tilde{X}_{j_1}\tilde{X}_{j_2}\right]\right) = \frac{1}{\rho^2}\sum_{j=1}^{r} o_{1,j} o_{2,j} \rho^2 = 0,$$

where again we used the facts that $\tilde{X}_j$ are i.i.d. and that the rows of O are orthogonal. We now have by Lovász and Vempala (2007, Thm. 5.14) that the density $\frac{1}{\rho^2}p\left(\frac{1}{\rho}\mathbf{x}\right)$ satisfies

$$\frac{1}{\rho^2}p\left(\frac{1}{\rho}\mathbf{v}\right) \geq 2^{-18\|\mathbf{v}\|}\frac{1}{\rho^2}p(\mathbf{0}) \geq 2^{-18\|\mathbf{v}\|-14}$$

for all $\mathbf{v}$ such that $\|\mathbf{v}\| \leq \frac{1}{9}$. The above implies that for all $\mathbf{x}$ with $\|\mathbf{x}\| \leq \frac{1}{9\rho}$, we have

$$p(\mathbf{x}) \geq \rho^2 2^{-18\rho\|\mathbf{x}\|-14} \geq \rho^2 2^{-16},$$

concluding the proof of the lemma. ∎

We are ready to prove Thm. 10.

Proof [of Thm. 10] We begin with evaluating the distribution of the output of an arbitrary neuron in the first hidden layer, given by $\mathrm{erf}(\langle U_j, X\rangle + B_j)$, where $X \sim \mathcal{D}_d$, $B_j \sim \mathcal{N}(0, 0.25)$ and $U_j \sim \mathcal{N}(0, 0.25)$. We first observe that the outputs of two neurons are not independent in general. Indeed, if $\mathcal{D}_d$ generated an instance with a very large norm, then $\mathrm{erf}(\langle U_j, X\rangle + B_j)$ will be very close to either -1 or 1, depending on the sign of $\langle U_j, X\rangle$. Given the information that a neuron outputs a value close to 1 in absolute value, it is far less likely for other neurons to attain values close to zero. Fortunately, by Lemma 20 we have that

$$(\langle U_j, X\rangle + B_j \mid \|X\| = x) \sim \mathcal{N}(0, 0.25(1 + x^2)).$$

In words, $\langle U_j, X\rangle + B_j$ conditioned on $\|X\|$ taking the value x follows a normal distribution with zero mean and variance $0.25(1 + x^2)$. This conditioning is useful, since Lemma 20 also implies that under this conditioning, the output of the neurons are mutually independent random variables, which would prove integral in our analysis. We now compute the distribution of the random variable $\tilde{X}_j = \mathrm{erf}(\langle U_j, X\rangle + B_j)$, also conditioned on $\|X\| = x$. We have for any $z \in (-1, 1)$

$$\mathbb{P}\left[\tilde{X}_j \leq z \mid \|X\| = x\right] = \mathbb{P}\left[\langle U_j, X\rangle + B_j \leq \mathrm{erf}^{-1}(z) \mid \|X\| = x\right]$$

$$= 0.5 + 0.5\,\mathrm{erf}\left(\frac{\mathrm{erf}^{-1}(z)}{0.5\sqrt{2 + 2x^2}}\right). \tag{39}$$

Differentiating with respect to z using the fact that $\frac{\partial}{\partial z} \operatorname{erf}^{-1}(z) = \frac{\sqrt{\pi}}{2} \exp\left([\operatorname{erf}^{-1}(z)]^2\right)$, we have that the density of $\tilde{X}_j$ given that $\|X\| = x$ is given by

$$f(z|x) = \frac{\exp\left(\operatorname{erf}^{-1}(z)^2 \left(1 - \frac{2}{1+x^2}\right)\right)}{\sqrt{2 + 2x^2}}. \tag{40}$$

Using the above, we can express the density of the random vector $(\tilde{X}_1, \ldots, \tilde{X}_r)$ as follows

$$\tilde{f}(\tilde{x}_1, \ldots, \tilde{x}_r) = \int_0^\infty \bar{f}(\tilde{x}_1, \ldots, \tilde{x}_r, x)dx = \int_0^\infty \prod_{j=1}^r f(\tilde{x}_j \mid x) \cdot \gamma_d(x)dx, \tag{41}$$

where $\bar{f}$ is the joint density of $\tilde{X}_1, \ldots, \tilde{X}_r$ and $\|X\|$, and we used the conditional mutual dependence of the $\tilde{X}_j$'s given that $\|X\| = x$ which Lemma 20 guarantees. We are now interested in lower bounding the above expression once it is marginalized to an arbitrary two-dimensional subspace. To this end, we may assume without loss of generality that $\mathbf{w}_1, \mathbf{w}_2$ are orthogonal and of unit length (e.g. by applying a Gram–Schmidt process which does not change their span). Letting $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_r)$, $\mathbf{o}_j \in \mathbb{R}^r$ denote the i-th row of an orthogonal matrix $O \in \mathbb{R}^{r \times r}$ and $A \subseteq \mathbb{R}^{r-2}$ denote the transformed integration domain $[-1, 1]^{r-2}$ after applying the transformation defined by the first $r - 2$ rows of O. We need to lower bound the following expression

$$\inf_{\tilde{\mathbf{x}}: \|\tilde{\mathbf{x}}\| \le \frac{1}{9}} \inf_{O:O \text{ orthogonal}} \int_A \tilde{f}(\langle \mathbf{o}_1, \tilde{\mathbf{x}} \rangle, \ldots, \langle \mathbf{o}_r, \tilde{\mathbf{x}} \rangle)d\tilde{x}_1 \ldots d\tilde{x}_{r-2}$$

$$= \inf_{\tilde{\mathbf{x}}: \|\tilde{\mathbf{x}}\| \le \frac{1}{9}} \inf_{O:O \text{ orthogonal}} \int_A \int_0^\infty \prod_{j=1}^r f(\langle \mathbf{o}_j, \tilde{\mathbf{x}} \rangle \mid x) \cdot \gamma_d(x)dxd\tilde{x}_1 \ldots d\tilde{x}_{r-2}$$

$$\ge \inf_{\tilde{\mathbf{x}}: \|\tilde{\mathbf{x}}\| \le \frac{1}{9}} \inf_{O:O \text{ orthogonal}} \int_A \int_0^2 \prod_{j=1}^r f(\langle \mathbf{o}_j, \tilde{\mathbf{x}} \rangle \mid x) \cdot \gamma_d(x)dxd\tilde{x}_1 \ldots d\tilde{x}_{r-2}$$

$$\ge \inf_{\tilde{\mathbf{x}}: \|\tilde{\mathbf{x}}\| \le \frac{1}{9}} \inf_{O:O \text{ orthogonal}} \int_A \int_0^2 \prod_{j=1}^r f(\langle \mathbf{o}_j, \tilde{\mathbf{x}} \rangle \mid x = 1) \cdot \gamma_d(x)dxd\tilde{x}_1 \ldots d\tilde{x}_{r-2}$$

$$\ge C \inf_{\tilde{\mathbf{x}}: \|\tilde{\mathbf{x}}\| \le \frac{1}{9}} \inf_{O:O \text{ orthogonal}} \int_A \prod_{j=1}^r f(\langle \mathbf{o}_j, \tilde{\mathbf{x}} \rangle \mid x = 1)d\tilde{x}_1 \ldots d\tilde{x}_{r-2}, \tag{42}$$

where the equality is by Eq. (41), the first inequality is because the integrand is non-negative, the second inequality is by Lemma 23 since $\langle \mathbf{o}_j, \tilde{\mathbf{x}} \rangle \le \|\mathbf{o}_j\| \cdot \|\tilde{\mathbf{x}}\| \le 1/9$ from Cauchy-Schwartz, and the last inequality is due to $\mathcal{D}_d$ satisfying Assumption 6(1) with the constant $C > 0$. Turning to analyze the density function of the marginal given by

$$p(\tilde{x}_{r-1}, \tilde{x}_r) = \int_A \prod_{j=1}^r f(\langle \mathbf{o}_j, \tilde{\mathbf{x}} \rangle \mid x = 1)d\tilde{x}_1 \ldots d\tilde{x}_{r-2},$$

observe that it is the density function obtained by an orthogonal transformation and marginalization on r i.i.d. random variables with density $f(z \mid x = 1)$. Conveniently, we have from Eq. (40) that $f(z \mid x = 1) = 0.5$ for all $z \in [-0.5, 0.5]$. Namely, this is a uniform distribution on the interval

$[-0.5, 0.5]$, which has variance $\frac{1}{12}$ and is clearly log-concave in z by definition. Lemma 24 then yields that Eq. (42) is lower bounded by

$$C \cdot \frac{1}{12} \cdot 2^{-16} \geq C \cdot 10^{-6},$$

concluding the proof of Thm. 10. ∎

Appendix F. Gradient Descent Convergence Proofs

The proofs of Propositions (11,12) build on the following lemmas, where the first which is presented below, bounds the growth function of the hypothesis class of predictors that are intersections of halfspaces. This lemma will allow us to bound the empirical Rademacher complexity of certain function classes that we will encounter in the proofs of the propositions.

Lemma 25 *Let $r \geq 2$. Let $\mathcal{H}_{r,m}$ denote the hypothesis class of intersections of m halfspaces in r-dimensional Euclidean space (i.e. $\mathbf{x} \mapsto 1$ if and only if it is contained inside all m halfspaces). Then the growth function of $\mathcal{H}_{r,m}$ denoted $\tau_{\mathcal{H}_{r,m}}(\cdot)$ satisfies*

$$\tau_{\mathcal{H}_{r,m}}(n) \leq n^{m(r+1)}$$

for all $n > r + 1$.

Proof Let $\mathcal{H}_r$ denote the class of predictors in $\mathbb{R}^r$ parameterized by $\mathbf{w} \in \mathbb{R}^r$ and $b \in \mathbb{R}$ such that $h_{\mathbf{w},b}(\mathbf{x}) = \mathbb{1}\left\{\mathbf{w}^\top \mathbf{x} + b \geq 0\right\}$. It is a well-known fact that the VC-dimension of $\mathcal{H}_r$ is $r + 1$. By the Sauer-Shelah lemma (Shalev-Shwartz and Ben-David, 2014, Lemma 6.10) and the assumption $r \geq 2$, we have

$$\tau_{\mathcal{H}_r}(n) \leq \left(\frac{en}{r+1}\right)^{r+1} \leq n^{r+1}.$$

Next, observe that for any $h \in \mathcal{H}_{r,m}$ and a set $C = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n\}$, the number of different labellings on C produced by hypotheses in $\mathcal{H}_{r,m}$ is uniquely determined by a set of m halfspaces. Since each such halfspace can produce at most n^{r+1} different labellings on the set C, the size of the different labellings attained by hypotheses in $\mathcal{H}_{r,m}$ is therefore upper bounded by $\left(n^{r+1}\right)^m = n^{m(r+1)}$, concluding the proof of the lemma. ∎

The following lemma bounds the empirical Rademacher complexity of certain function classes that control the empirical risk we obtain over our data.

Lemma 26 *Fix $\mathbf{v} \in \mathbb{R}^r$, assume $n > r + 1$ and define the function classes*

$$\mathcal{F}_1 := \left\{\tilde{\mathbf{x}} \mapsto \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}} \geq 1\right\} \cdot \mathbb{1}\left\{\mathbf{v}^\top \tilde{\mathbf{x}} \leq 0\right\} + \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}} \leq 0\right\} \cdot \mathbb{1}\left\{\mathbf{v}^\top \tilde{\mathbf{x}} \geq 1\right\} : \mathbf{w} \in \mathbb{R}^r\right\},$$

$$\mathcal{F}_2 := \left\{\tilde{\mathbf{x}} \mapsto \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}} \in (0,1)\right\} \cdot \mathbb{1}\left\{\mathbf{v}^\top \tilde{\mathbf{x}} \in (0,1)\right\} \cdot \left(\frac{(\mathbf{w} - \mathbf{v})^\top}{\|\mathbf{w} - \mathbf{v}\|} \tilde{\mathbf{x}}\right)^2 : \mathbf{w} \in \mathbb{R}^r, \mathbf{w} \neq \mathbf{v}\right\}.$$

Let $f_{\mathbf{w}} \in \mathcal{F}_1$. Then for any distribution $\tilde{\mathcal{D}}$ we sample $\tilde{\mathbf{x}}_i$ from, we have with probability at least $1 - \delta$ that

$$\sup_{\mathbf{w} \in \mathbb{R}^r} \left| \frac{1}{n} \sum_{i=1}^{n} f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) - \mathbb{E}_{\tilde{\mathbf{x}}_i \sim \tilde{\mathcal{D}}} \left[f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) \right] \right| \leq 4 \sqrt{\frac{(8r+8) \log_2(n)}{n}} + \sqrt{\frac{2 \ln(2/\delta)}{n}}.$$

Further, if $f_{\mathbf{w}} \in \mathcal{F}_2$ then for any distribution $\tilde{\mathcal{D}}$ we sample $\tilde{\mathbf{x}}_i$ from, which satisfies $\|\tilde{\mathbf{x}}_i\| \leq \sqrt{r}$ almost surely, we have with probability at least $1 - \delta$ that

$$\sup_{\mathbf{w} \in \mathbb{R}^r} \left| \frac{1}{n} \sum_{i=1}^{n} f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) - \mathbb{E}_{\tilde{\mathbf{x}}_i \sim \tilde{\mathcal{D}}} \left[f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) \right] \right| \leq 4r \sqrt{\frac{(8r+8) \log_2(n)}{n}} + \sqrt{\frac{2 \ln(2/\delta)}{n}}.$$

Proof We shall use a Rademacher complexity argument to prove the lemma. Starting with $\mathcal{F}_1$, we have by standard Rademacher complexity arguments that

$$\sup_{\mathbf{w} \in \mathbb{R}^r} \left| \frac{1}{n} \sum_{i=1}^{n} f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) - \mathbb{E}_{\tilde{\mathbf{x}}_i \sim \tilde{\mathcal{D}}} \left[f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) \right] \right| \leq 2 R_n \left(\mathcal{F}_1 \left(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n \right) \right) + \sqrt{\frac{2 \ln(2/\delta)}{n}} \qquad (43)$$

(Boucheron et al., 2005, Thm. 3.2), where $R_n \left(\mathcal{F}_1 \left(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n \right) \right) := \mathbb{E} \left[\sup_{f_{\mathbf{w}} \in \mathcal{F}_1} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) \right| \right]$ is the empirical Rademacher complexity of $\mathcal{F}_1$, and the expectation is over $\xi_1, \ldots, \xi_n$ which are i.i.d. Rademacher random variables. Next, we have that the empirical Rademacher complexity of $\mathcal{F}_1$ can be upper bounded by

$$\sqrt{\frac{2 \log_2(\tau_{\mathcal{F}_1}(n))}{n}},$$

where $\tau_{\mathcal{F}_1}(\cdot)$ is the growth function of $\mathcal{F}_1$ (Boucheron et al., 2005, Eq. (6)). Since the class $\mathcal{F}_1$ is contained inside the class of intersections of 4 halfspaces, its growth function is upper bounded by the growth function of the class of such intersections, and thus by virtue of Lemma 25 the above is at most

$$\sqrt{\frac{(8r+8) \log_2(n)}{n}}.$$

Plugging this back in Eq. (43), we arrive at

$$\sup_{\mathbf{w} \in \mathbb{R}^r} \left| \frac{1}{n} \sum_{i=1}^{n} f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) - \mathbb{E}_{\tilde{\mathbf{x}}_i \sim \tilde{\mathcal{D}}} \left[f_{\mathbf{w}}(\tilde{\mathbf{x}}_i) \right] \right| \leq 2 \sqrt{\frac{(8r+8) \log_2(n)}{n}} + \sqrt{\frac{2 \ln(2/\delta)}{n}},$$

where the bound on $\mathcal{F}_1$ now follows from Shalev-Shwartz and Ben-David (2014, Lemma 26.6) after a simple scaling to accommodate for the difference between classifiers mapping to $\{0, 1\}$ and $\{-1, 1\}$.

Turning to bound the difference for $\mathcal{F}_2$, by Eq. (43) it suffices to upper bound the empirical Rademacher complexity of $\mathcal{F}_2$. For any $i \in [n]$, define the functions $\phi_i(x) := x \left(\frac{(\mathbf{w} - \mathbf{v})^\top}{\|\mathbf{w} - \mathbf{v}\|} \tilde{\mathbf{x}}_i \right)^2$ and for $\mathbf{a} \in \mathbb{R}^n$ let $\boldsymbol{\phi}(\mathbf{a}) = (\phi_1(a_1), \ldots, \phi_n(a_n))$. We now have that

$$R_n \left(\mathcal{F}_2 \left(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n \right) \right) \leq R_n \left(\boldsymbol{\phi} \circ \mathcal{F}_1 \left(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n \right) \right) \leq r \cdot R_n \left(\mathcal{F}_1 \left(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n \right) \right),$$

where the first inequality is by the definitions of $\mathcal{F}_1$ and $\mathcal{F}_2$ and the fact that changing the orientation (or omitting some) of the halfspaces defining $\mathcal{F}_1$ results in the same growth function bound, and

the second inequality follows from the contraction lemma (Shalev-Shwartz and Ben-David, 2014, Lemma 26.9) where the Lipschitz constant of ϕ_i is upper bounded by r for all i due to Cauchy-Schwartz and the fact that $\|\tilde{\mathbf{x}}_i\| \leq \sqrt{r}$ almost surely. The bound on $\mathcal{F}_2$ then follows immediately from our previously derived bound on $\mathcal{F}_1$. ∎

Lemma 27 *Suppose $\mathbf{0} \neq \mathbf{w}, \mathbf{v} \in \mathbb{R}^r$ such that $\theta := \theta_{\mathbf{w},\mathbf{v}} > 0$, $\|\mathbf{w} - \mathbf{v}\| \leq \|\mathbf{v}\| \leq r^{-0.25}$ and $r \geq 10^6$. Let $A_{1,3} = (-\infty, 0] \times [1, \infty)$ and $A_{3,1} = [1, \infty) \times (-\infty, 0]$. Then under Assumption 7, we have*

$$\mathbb{E}_{\tilde{\mathbf{x}}} \left[\mathbb{1} \left\{ \left(\mathbf{w}^\top \tilde{\mathbf{x}}, \mathbf{v}^\top \tilde{\mathbf{x}} \right) \in A_{1,3} \cup A_{3,1} \right\} \right] \leq \frac{1}{r}.$$

Proof Since the expectation of an indicator function is the probability of the event, we have

$$\mathbb{E}_{\tilde{\mathbf{x}}} \left[\mathbb{1} \left\{ \left(\mathbf{w}^\top \tilde{\mathbf{x}}, \mathbf{v}^\top \tilde{\mathbf{x}} \right) \in A_{1,3} \cup A_{3,1} \right\} \right] = \mathbb{P}_{\tilde{\mathbf{x}}} \left[\mathbb{1} \left\{ \left(\mathbf{w}^\top \tilde{\mathbf{x}}, \mathbf{v}^\top \tilde{\mathbf{x}} \right) \in A_{1,3} \cup A_{3,1} \right\} \right]. \tag{44}$$

In what follows, we consider the projections of $\mathbf{w}, \mathbf{v}$, denoted by $\hat{\mathbf{w}}, \hat{\mathbf{v}}$, onto the two-dimensional space spanned by them, so that $\hat{\mathbf{x}} \in \mathbb{R}^2$ is sampled from the marginal distribution $\tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}$. This is justified since the expression we are bounding in the lemma only depends on dot products of $\tilde{\mathbf{x}}$ with $\mathbf{w}, \mathbf{v}$ which do not change when we perform such a projection.

We shall now derive a bound on the above probability by bounding $\|\hat{\mathbf{x}}\|$ with high probability over the randomness in sampling it from the marginal distribution. Recall that by Lemma 20 we have that conditioned on $\|\tilde{\mathbf{x}}\|$, each coordinate x_j of $\hat{\mathbf{x}}$ is distributed according to $\sum_{i=1}^n o_{j,i} \tilde{X}_i$, where $\tilde{X}_i$ are i.i.d. with density given in Eq. (40) and $\mathbf{o}_j = (o_{j,1}, \ldots, o_{j,n})$, $j \in \{1, 2\}$, are orthogonal. Since $\tilde{X}_i$ is bounded in $[-1, 1]$ and $\mathbb{E}\left[\tilde{X}_i\right] = 0$ regardless of the realization of $\|\tilde{\mathbf{x}}\|$, we can use Hoeffding's inequality to deduce that for all $j \in \{1, 2\}$,

$$\mathbb{P}_{\tilde{\mathbf{x}} \sim \tilde{\mathcal{D}}} \left[|x_j| \geq \frac{r^{0.25}}{2\sqrt{2}\theta} \right] \leq 2 \exp\left(-\frac{\sqrt{r}}{4\theta^2} \right).$$

By a union bound and since the ball of radius $\frac{r^{0.25}}{2\theta}$ contains the square $\left[-\frac{r^{0.25}}{2\sqrt{2}\theta}, \frac{r^{0.25}}{2\sqrt{2}\theta} \right]^2$, we have

$$\mathbb{P}_{\hat{\mathbf{x}} \sim \tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}} \left[\|\hat{\mathbf{x}}\| \geq \frac{r^{0.25}}{2\theta} \right] \leq \mathbb{P}_{\hat{\mathbf{x}} \sim \tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}} \left[\hat{\mathbf{x}} \notin \left[-\frac{r^{0.25}}{2\sqrt{2}\theta}, \frac{r^{0.25}}{2\sqrt{2}\theta} \right]^2 \right] \leq 4 \exp\left(-\frac{\sqrt{r}}{4\theta^2} \right). \tag{45}$$

Now, define

$$\bar{\mathbf{w}} := \frac{\|\hat{\mathbf{v}}\|}{\|\hat{\mathbf{w}}\|} \hat{\mathbf{w}}, \qquad \bar{\mathbf{v}} := \frac{\|\hat{\mathbf{w}}\|}{\|\hat{\mathbf{v}}\|} \hat{\mathbf{v}}.$$

Since $\theta := \theta_{\mathbf{w},\mathbf{v}} = \theta_{\bar{\mathbf{w}},\hat{\mathbf{v}}} = \theta_{\hat{\mathbf{w}},\bar{\mathbf{v}}}$, we have

$$\|\bar{\mathbf{w}} - \hat{\mathbf{v}}\|^2 = \|\bar{\mathbf{w}}\|^2 + \|\hat{\mathbf{v}}\|^2 - 2\|\bar{\mathbf{w}}\| \|\hat{\mathbf{v}}\| \cos(\theta) = 2\|\hat{\mathbf{v}}\|^2 (1 - \cos(\theta)) \leq \frac{\theta^2}{\sqrt{r}}, \tag{46}$$

where in the inequality we used $\|\hat{\mathbf{v}}\|^2 \leq 1/\sqrt{r}$ and $1 - \cos(\theta) \leq \frac{\theta^2}{2}$ which holds for all $\theta \in \mathbb{R}$ by a Taylor expansion. From a similar argument we have

$$\|\hat{\mathbf{w}} - \bar{\mathbf{v}}\| \leq 2\|\hat{\mathbf{w}}\|^2 (1 - \cos(\theta)) \leq \frac{4\theta^2}{\sqrt{r}}, \tag{47}$$

where we used the triangle inequality and our assumption to obtain $\|\hat{\mathbf{w}}\| \leq \|\hat{\mathbf{w}} - \hat{\mathbf{v}}\| + \|\hat{\mathbf{v}}\| \leq 2\|\hat{\mathbf{v}}\|$. Using the above shorthands, we also have that

$$\left(\hat{\mathbf{w}}^\top \hat{\mathbf{x}}, \hat{\mathbf{v}}^\top \hat{\mathbf{x}}\right) \in A_{1,3} \iff \left(\bar{\mathbf{w}}^\top \hat{\mathbf{x}}, \hat{\mathbf{v}}^\top \hat{\mathbf{x}}\right) \in A_{1,3},$$

and

$$\left(\hat{\mathbf{w}}^\top \hat{\mathbf{x}}, \hat{\mathbf{v}}^\top \hat{\mathbf{x}}\right) \in A_{3,1} \iff \left(\hat{\mathbf{w}}^\top \hat{\mathbf{x}}, \bar{\mathbf{v}}^\top \hat{\mathbf{x}}\right) \in A_{3,1}.$$

This holds true since $x \in (-\infty, 0]$ if and only if $ax \in (-\infty, 0]$ for all $a > 0$. A necessary condition for $\left(\bar{\mathbf{w}}^\top \hat{\mathbf{x}}, \hat{\mathbf{v}}^\top \hat{\mathbf{x}}\right) \in A_{1,3}$ to hold is that

$$\left|\bar{\mathbf{w}}^\top \hat{\mathbf{x}} - \hat{\mathbf{v}}^\top \hat{\mathbf{x}}\right| \geq 1,$$

since otherwise, if the two dot products are less than that apart then the inclusion of either in one of the intervals $(-\infty, 0], [1, \infty)$ excludes the other from the second interval. We now compute using Cauchy-Schwartz and the square roots of Eqs. (46,47) to obtain

$$\left|\bar{\mathbf{w}}^\top \hat{\mathbf{x}} - \hat{\mathbf{v}}^\top \hat{\mathbf{x}}\right| \leq \|\bar{\mathbf{w}} - \hat{\mathbf{v}}\| \cdot \|\hat{\mathbf{x}}\| \leq r^{-0.25} \theta \|\hat{\mathbf{x}}\|,$$

and

$$\left|\hat{\mathbf{w}}^\top \hat{\mathbf{x}} - \bar{\mathbf{v}}^\top \hat{\mathbf{x}}\right| \leq \|\hat{\mathbf{w}} - \bar{\mathbf{v}}\| \cdot \|\hat{\mathbf{x}}\| \leq 2r^{-0.25} \theta \|\hat{\mathbf{x}}\|,$$

That is, $\|\hat{\mathbf{x}}\| < 0.5 r^{0.25}/\theta$ implies $\left(\hat{\mathbf{w}}^\top \hat{\mathbf{x}}, \hat{\mathbf{v}}^\top \hat{\mathbf{x}}\right) \notin A_{1,3} \cup A_{3,1}$. Along with Eq. (45), this means that

$$\mathbb{P}_{\hat{\mathbf{x}} \sim \tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}} \left[\left(\hat{\mathbf{w}}^\top \hat{\mathbf{x}}, \hat{\mathbf{v}}^\top \hat{\mathbf{x}}\right) \in A_{1,3} \cup A_{3,1}\right] \leq \mathbb{P}_{\hat{\mathbf{x}} \sim \tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}} \left[\|\hat{\mathbf{x}}\| \geq \frac{r^{0.25}}{2\theta}\right] \leq 4 \exp\left(-\frac{\sqrt{r}}{4\theta^2}\right).$$

Plugging the above in Eq. (44) and returning to r-dimensional space, we have arrived at

$$\mathbb{E}_{\tilde{\mathbf{x}}} \left[\mathbb{1}\left\{\left(\mathbf{w}^\top \tilde{\mathbf{x}}, \mathbf{v}^\top \tilde{\mathbf{x}}\right) \in A_{1,3} \cup A_{3,1}\right\}\right] \leq 4 \exp\left(-\frac{\sqrt{r}}{4\theta^2}\right) \leq 4 \exp\left(-\frac{\sqrt{r}}{4\pi^2}\right) \leq \frac{1}{r},$$

where the last inequality follows from the inequality $4 \exp(-\sqrt{x}/4\pi^2) \leq 1/x$ which holds for all $x \geq 10^6$. ∎

Having established the required tools to prove Propositions (11,12), we now turn to do so.

F.1. Proof of Proposition 11

Proof First note that the proposition statement is trivially true if $\mathbf{w} = \mathbf{v}$. Assume $\mathbf{w} \neq \mathbf{v}$ for the remainder of the proof and define $\epsilon_i := \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i\right]_+^1 - y_i$. We compute

$$\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{v} \rangle = \frac{1}{n} \sum_{i=1}^n \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i\right]_+^1 - y_i \right) \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i\right)$$

$$= \frac{1}{n} \sum_{i=1}^n \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i\right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i\right]_+^1 + \epsilon_i \right) \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i\right)$$

$$= \frac{1}{n} \sum_{i=1}^n \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i\right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i\right]_+^1 \right) \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i\right)$$

$$+ \frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i\right).$$

By Eq. (7), the above is at least

$$\frac{1}{2n} \sum_{i=1}^n \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i\right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i\right]_+^1 \right) \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i\right).$$

Since $[\cdot]_+^1$ is non-decreasing, $\left([a]_+^1 - [b]_+^1\right) \cdot (a - b) \geq 0$ for all $a, b \in \mathbb{R}$, and therefore each summand in the above is non-negative and we can omit summands in which $\mathbf{v}^\top \hat{\mathbf{x}}_i \notin (0,1)$, to lower bound the above by

$$\frac{1}{2n} \sum_{i=1}^n \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i\right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i\right]_+^1 \right) \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \mathbb{1}\left\{\mathbf{v}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i\right)$$

$$= \frac{1}{2n} \|\mathbf{w} - \mathbf{v}\|^2 \sum_{i=1}^n \left(\frac{(\mathbf{w} - \mathbf{v})^\top}{\|\mathbf{w} - \mathbf{v}\|} \tilde{\mathbf{x}}_i \right)^2 \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \mathbb{1}\left\{\mathbf{v}^\top \tilde{\mathbf{x}}_i \in (0,1)\right\}.$$

By Lemma 26, with probability at least $1 - \delta$, the above is lower bounded uniformly for all $\mathbf{w}$ by

$$\frac{1}{2} \|\mathbf{w} - \mathbf{v}\|^2 \left(\mathbb{E}_{\tilde{\mathbf{x}}} \left[\left(\frac{(\mathbf{w} - \mathbf{v})^\top}{\|\mathbf{w} - \mathbf{v}\|} \tilde{\mathbf{x}}_i \right)^2 \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}} \in (0,1)\right\} \cdot \mathbb{1}\left\{\mathbf{v}^\top \tilde{\mathbf{x}} \in (0,1)\right\} \right] \right.$$

$$\left. - 4r\sqrt{\frac{(8r+8)\log_2(n)}{n}} - \sqrt{\frac{2\log(2/\delta)}{n}} \right). \tag{48}$$

Let P denote the orthogonal projection matrix projecting onto the subspace spanned by $\mathbf{w}, \mathbf{v}$. Since the expectation above is non-negative, we can lower bound it by

$$\mathbb{E}_{\tilde{\mathbf{x}}} \left[\left(\frac{(\mathbf{w} - \mathbf{v})^\top}{\|\mathbf{w} - \mathbf{v}\|} \tilde{\mathbf{x}} \right)^2 \cdot \mathbb{1}\left\{\|P\tilde{\mathbf{x}}\| \leq \alpha\right\} \cdot \mathbb{1}\left\{\mathbf{w}^\top \tilde{\mathbf{x}} \in (0,1)\right\} \cdot \mathbb{1}\left\{\mathbf{v}^\top \tilde{\mathbf{x}} \in (0,1)\right\} \right]. \tag{49}$$

We now argue that $\|\mathbf{w} - \mathbf{v}\|^2 \leq \|\mathbf{v}\|^2$ implies that $\theta_{\mathbf{w}, \mathbf{v}} \leq \frac{\pi}{2}$. This follows from

$$\|\mathbf{w}\|^2 - 2\langle \mathbf{w}, \mathbf{v} \rangle + \|\mathbf{v}\|^2 = \|\mathbf{w} - \mathbf{v}\|^2 \leq \|\mathbf{v}\|^2,$$

from which we have $\langle \mathbf{w}, \mathbf{v} \rangle \geq \frac{\|\mathbf{w}\|^2}{2}$, i.e. the dot product between $\mathbf{w}$ and $\mathbf{v}$ is strictly positive and therefore

$$\theta_{\mathbf{w},\mathbf{v}} \leq \frac{\pi}{2}. \tag{50}$$

Next, we have from the triangle inequality and our assumption that $\|\mathbf{w}\| \leq \|\mathbf{w} - \mathbf{v}\| + \|\mathbf{v}\| \leq 2\|\mathbf{v}\| \leq 1$. Since P is an orthogonal projection that keeps $\mathbf{w}, \mathbf{v}$ unchanged, we have that $\mathbf{w}^\top \tilde{\mathbf{x}} = P\mathbf{w}^\top \tilde{\mathbf{x}} = \mathbf{w}^\top P\tilde{\mathbf{x}}$ and $\mathbf{v}^\top \tilde{\mathbf{x}} = P\mathbf{v}^\top \tilde{\mathbf{x}} = \mathbf{v}^\top P\tilde{\mathbf{x}}$. We thus have for any $\tilde{\mathbf{x}}$ in Eq. (49) by using Cauchy-Schwartz and Thm. 10 which guarantees $\alpha \leq 1$, that $|\mathbf{w}^\top \tilde{\mathbf{x}}| \leq \|\mathbf{w}\| \cdot \|P\tilde{\mathbf{x}}\| \leq \alpha \leq \frac{1}{9}$. Letting $\hat{\mathbf{w}}, \hat{\mathbf{v}} \in \mathbb{R}^2$ denote the projections of $\mathbf{w}, \mathbf{v}$ onto the two-dimensional subspace spanned by them, we have that Eq. (49) equals

$$\mathbb{E}_{\tilde{\mathbf{x}}} \left[\left(\frac{(\mathbf{w} - \mathbf{v})^\top}{\|\mathbf{w} - \mathbf{v}\|} \tilde{\mathbf{x}} \right)^2 \cdot \mathbb{1}\left\{ \|P\tilde{\mathbf{x}}\| \leq \alpha \right\} \cdot \mathbb{1}\left\{ \mathbf{w}^\top \tilde{\mathbf{x}} > 0 \right\} \cdot \mathbb{1}\left\{ \mathbf{v}^\top \tilde{\mathbf{x}} > 0 \right\} \right]$$

$$= \mathbb{E}_{\hat{\mathbf{x}} \sim \tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}} \left[\left(\frac{(\hat{\mathbf{w}} - \hat{\mathbf{v}})^\top}{\|\hat{\mathbf{w}} - \hat{\mathbf{v}}\|} \hat{\mathbf{x}} \right)^2 \cdot \mathbb{1}\left\{ \|\hat{\mathbf{x}}\| \leq \alpha \right\} \cdot \mathbb{1}\left\{ \hat{\mathbf{w}}^\top \hat{\mathbf{x}} > 0 \right\} \cdot \mathbb{1}\left\{ \hat{\mathbf{v}}^\top \hat{\mathbf{x}} > 0 \right\} \right] \tag{51}$$

$$\geq \inf_{\mathbf{u} \in \mathbb{R}^2 : \|\mathbf{u}\|=1} \mathbb{E}_{\hat{\mathbf{x}} \sim \tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}} \left[\left(\mathbf{u}^\top \hat{\mathbf{x}} \right)^2 \cdot \mathbb{1}\left\{ \|\hat{\mathbf{x}}\| \leq \alpha \right\} \cdot \mathbb{1}\left\{ \hat{\mathbf{w}}^\top \hat{\mathbf{x}} > 0 \right\} \cdot \mathbb{1}\left\{ \hat{\mathbf{v}}^\top \hat{\mathbf{x}} > 0 \right\} \right]$$

$$\geq \beta \inf_{\mathbf{u} \in \mathbb{R}^2 : \|\mathbf{u}\|=1} \int_{\mathbb{R}^2} \left(\mathbf{u}^\top \hat{\mathbf{x}} \right)^2 \cdot \mathbb{1}\left\{ \|\hat{\mathbf{x}}\| \leq \alpha \right\} \cdot \mathbb{1}\left\{ \hat{\mathbf{w}}^\top \hat{\mathbf{x}} > 0 \right\} \cdot \mathbb{1}\left\{ \hat{\mathbf{v}}^\top \hat{\mathbf{x}} > 0 \right\} d\hat{\mathbf{x}} \tag{52}$$

$$\geq \frac{\alpha^4 \beta}{8\sqrt{2}} \sin^3 \left(\frac{\pi}{8} \right) \geq \frac{\alpha^4 \beta}{210}. \tag{53}$$

In the above, Eq. (51) follows from the fact that all the dot products are in the subspace spanned by $\mathbf{w}, \mathbf{v}$, so we can take the expectation over the marginal distribution $\tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}$; Eq. (52) follows from Thm. 10, since the marginal density of $\tilde{\mathcal{D}}_{\mathbf{w},\mathbf{v}}$ is lower bounded by β for all $\hat{\mathbf{x}}$ with norm at most $1/9$; and the inequalities in Eq. (53) are by Yehudai and Shamir (2020, Lemma B.1) applied to the vectors $\mathbf{w}$ and $\mathbf{v}$ and by Eq. (50). Plugging the lower bound on the expectation we obtained above in Eq. (48), the proposition follows. ∎

F.2. Proof of Proposition 12

Proof We begin with upper bounding the loss of making predictions using $\mathbf{w}$ when the target values are determined by $\mathbf{v}$:

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2, \tag{54}$$

in terms of the following surrogate loss

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2 \cdot \mathbb{1}\left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0, 1) \right\}. \tag{55}$$

To this end, we partition the sums into 9 parts, depending on the values the dot products take in the following manner: Define the intervals $I_1' = (-\infty, 0]$, $I_2' = (0, 1)$ and $I_3' = [1, \infty)$, and for

$j_1, j_2 \in \{1, 2, 3\}$ define the sets $A_{j_1, j_2} = I'_{j_1} \times I'_{j_2}$. We now wish to upper bound

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2 = \frac{1}{n} \sum_{j_1, j_2 \in \{1, 2, 3\}} \sum_{i: \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i, \mathbf{v}^\top \tilde{\mathbf{x}}_i \right) \in A_{j_1, j_2}} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2 .$$

We will do so by analyzing three different cases:

- We begin with the observation that the sums are identical over the sets $A_{2,1}, A_{2,2}, A_{2,3}$ and $A_{1,1}, A_{3,3}$. This is because for instances $\tilde{\mathbf{x}}_i$ with dot products in $A_{1,1}, A_{3,3}$, we have that $\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 = \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1$, and therefore all such summands are identically zero in both Eq. (54) and Eq. (55); and for instances with dot products in $A_{2,1}, A_{2,2}, A_{2,3}$ we have that $\mathbb{1} \left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0,1) \right\} = 1$, so each summand in both equations is identical.

- Turning to analyze the sums for $A_{1,2}, A_{3,2}$, we use Thm. 9 with a confidence of $\delta/2$ to obtain

$$\left| \left\{ i : \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i, \mathbf{v}^\top \tilde{\mathbf{x}}_i \right) \in A_{1,2} \cup A_{3,2} \right\} \right| \leq \frac{3600 q(d) + 1000 c_2 \sqrt{\ln (16n/\delta)}}{r^{0.25}} n.$$

  This is true since $\mathbf{v}^\top \tilde{\mathbf{x}}_i \in (0,1)$ implies that $\mathbf{v}$ misclassifies $\tilde{\mathbf{x}}_i$ which can only happen on the above fraction of the data instances. This allows us to derive the following upper bound

$$\frac{1}{n} \sum_{i: \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i, \mathbf{v}^\top \tilde{\mathbf{x}}_i \right) \in A_{1,2} \cup A_{3,2}} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2 \leq \frac{1}{n} \sum_{i: \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i, \mathbf{v}^\top \tilde{\mathbf{x}}_i \right) \in A_{1,2} \cup A_{3,2}} 1$$

$$\leq \frac{3600 q(d) + 1000 c_2 \sqrt{\ln (16n/\delta)}}{r^{0.25}} .$$

- Lastly, assuming that $\theta_{\mathbf{w}, \mathbf{v}} > 0$ we have uniformly for any $\mathbf{w} \in \mathbb{R}^r$ with probability at least $1 - \frac{\delta}{2}$, the sums over the sets $A_{1,3}, A_{3,1}$ are at most

$$\frac{1}{n} \sum_{i: \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i, \mathbf{v}^\top \tilde{\mathbf{x}}_i \right) \in A_{1,3} \cup A_{3,1}} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left\{ i : \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i, \mathbf{v}^\top \tilde{\mathbf{x}}_i \right) \in A_{1,3} \cup A_{3,1} \right\}$$

$$\leq \mathbb{E}_{\tilde{\mathbf{x}}} \left[\mathbb{1} \left\{ \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i, \mathbf{v}^\top \tilde{\mathbf{x}}_i \right) \in A_{1,3} \cup A_{3,1} \right\} \right] + 4 \sqrt{\frac{(8r + 8) \log_2 (n)}{n}} + \sqrt{\frac{2 \ln (4/\delta)}{n}}$$

$$\leq \frac{1}{r} + 4 \sqrt{\frac{(8r + 8) \log_2 (n)}{n}} + \sqrt{\frac{2 \ln (4/\delta)}{n}} .$$

  where the equality is due to $\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 = 1$ for any $\tilde{\mathbf{x}}_i$ with dot products in $A_{1,3} \cup A_{3,1}$, the first inequality is due to Lemma 26 using a confidence of $\delta/2$, and the last inequality is due to Lemma 27.

We can now combine the above three cases using a union bound to deduce that with probability at least $1 - \delta$

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2$$

$$- \frac{1}{r} - 4\sqrt{\frac{(8r+8)\log_2(n)}{n}} - \sqrt{\frac{2\log(4/\delta)}{n}} - \frac{3600q(d) + 1000c_2\sqrt{\ln(16n/\delta)}}{r^{0.25}}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2 \cdot \mathbb{1}\left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0, 1) \right\}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right) \cdot \mathbb{1}\left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0, 1) \right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i \right), \tag{56}$$

where in the last inequality we used that fact that $[\cdot]_+^1$ is 1-Lipschitz. We remark that in the case where $\theta_{\mathbf{w},\mathbf{v}} = 0$, the only possible non-empty dot product sets are $A_{1,1}, A_{2,2}, A_{3,3}$ and possibly also at most one of either $A_{2,3}, A_{3,2}$, therefore the upper bound we derived above is also valid in the case where $\theta_{\mathbf{w},\mathbf{v}} = 0$. This holds true since we can simply skip the third case which analyzes the sums with dot products in $A_{1,3}, A_{3,1}$ which requires that $\theta_{\mathbf{w},\mathbf{v}} > 0$ to invoke Lemma 27.

Turning to upper bound the expression

$$\frac{2}{n} \sum_{i=1}^{n} \left(\left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 - y_i \right) \cdot \mathbb{1}\left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0, 1) \right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i \right),$$

we have from Thm. 9 that the above loss can be decomposed into the loss over four intervals $I_1, \ldots, I_4$ and upper bounded as follows (note that we previously assumed that the theorem implications hold so we don't need to further evaluate the confidence over which they do)

$$\frac{2}{n} \sum_{j=1}^{4} \sum_{i:\|\mathbf{x}_i\| \in I_j} \left(\left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 - y_i \right) \cdot \mathbb{1}\left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0, 1) \right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i \right)$$

$$= \frac{2}{n} \sum_{i:\|\mathbf{x}_i\| \in I_2 \cup I_4} \left(\left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 - y_i \right) \cdot \mathbb{1}\left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0, 1) \right\} \cdot \left(\mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i \right)$$

$$\leq \frac{2}{n} \sum_{i:\|\mathbf{x}_i\| \in I_2 \cup I_4} 4 \leq 8 \frac{3600q(d) + 1000c_2\sqrt{\ln(16n/\delta)}}{r^{0.25}},$$

where the first inequality is due to $\left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 - y_i \leq 1$, $\mathbb{1}\left\{ \mathbf{w}^\top \tilde{\mathbf{x}}_i \in (0, 1) \right\} \leq 1$ and $\left| \mathbf{w}^\top \tilde{\mathbf{x}}_i - \mathbf{v}^\top \tilde{\mathbf{x}}_i \right| \leq 4$ for all i such that $\|\mathbf{x}_i\| \in I_2 \cup I_4$. Combining the above with Eq. (56) and the inequality in the proposition statement, we have arrived at the following bound

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 \right)^2 \leq 9 \cdot \frac{3600q(d) + 1000c_2\sqrt{\ln(16n/\delta)}}{r^{0.25}}$$

$$+ \frac{1}{r} + 4\sqrt{\frac{(8r+8)\log_2(n)}{n}} + \sqrt{\frac{2\ln(4/\delta)}{n}}. \tag{57}$$

We now use Thm. 9 one last time to upper bound the squared loss of $\mathbf{v}$ and obtain

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 - y_i \right)^2 \leq \frac{1}{n} \sum_{i:\|\mathbf{x}_i\| \in I_2 \cup I_4} 1 \leq \frac{3600 q(d) + 1000 c_2 \sqrt{\ln(16n/\delta)}}{r^{0.25}}.$$

Combining the above with Eq. (57) using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ with $a = \left[\mathbf{w}^\top \tilde{\mathbf{x}}_i \right]_+^1 - \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1$ and $b = \left[\mathbf{v}^\top \tilde{\mathbf{x}}_i \right]_+^1 - y_i$, the proposition follows. ∎

Appendix G. Proofs of Thm. 8 and Thm. 13

G.1. Proof of Thm. 13

Proof Define the function class

$$\mathcal{F} := \left\{ \tilde{\mathbf{x}} \mapsto \mathbf{w}^\top \tilde{\mathbf{x}} : \|\mathbf{w}\| \leq 1 \right\},$$

and define the loss

$$\ell(y, y') = \left([y]_+^1 - [y']_+^1 \right)^2.$$

Using standard Rademacher complexity arguments (Shalev-Shwartz and Ben-David, 2014, Thm. 26.5) and the fact that ℓ is upper bounded by 1, we have with probability at least $1 - \delta$ that

$$F(\mathbf{w}) \leq \hat{F}(\mathbf{w}) + 2R_n \left(\ell \circ \mathcal{F}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n) \right) + 4\sqrt{\frac{2 \ln(4/\delta)}{n}}.$$

where $R_n \left(\mathcal{F}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n) \right) := \mathbb{E} \left[\sup_{f_\mathbf{w} \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i f_\mathbf{w}(\tilde{\mathbf{x}}_i) \right| \right]$ is the empirical Rademacher complexity of $\mathcal{F}$, and the expectation is over $\xi_1, \ldots, \xi_n$ which are i.i.d. Rademacher random variables. To bound the empirical Rademacher complexity, we have from Shalev-Shwartz and Ben-David (2014, Lemma 26.10) that

$$R_n \left(\mathcal{F}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n) \right) \leq \frac{\max_i \|\tilde{\mathbf{x}}_i\|}{\sqrt{n}} \leq \sqrt{\frac{r}{n}}.$$

From the above, the contraction lemma (Shalev-Shwartz and Ben-David, 2014, Lemma 26.9) and since ℓ is 2-Lipschitz, we conclude

$$F(\mathbf{w}) \leq \hat{F}(\mathbf{w}) + 4\sqrt{\frac{r}{n}} + 4\sqrt{\frac{2 \ln(4/\delta)}{n}}.$$

∎

G.2. Proof of Thm. 8

Proof Before we delve into the proof of the theorem, we will first establish several technical inequalities that are implied by the assumptions in the theorem statement and that will be used throughout the proof. Let $\alpha, \beta \in (0, 1)$ denote the constants guaranteed by applying Thm. 10 which is justified by Assumption 6, and denote

$$p := \frac{1800q(d) + 500c_2\sqrt{\ln(64n/\delta)}}{r^{0.25}}.$$

Then we have that

$$r \geq \max\left\{12000^4, 500^4 \max\{c_1, 10\}^4 \ln^2(64n/\delta)\right\}, \tag{58}$$

$$\sqrt{\frac{\ln(64/\delta)}{2n}} \leq p \leq \frac{\epsilon}{16}, \tag{59}$$

$$4r\sqrt{\frac{(8r+8)\log_2(n)}{n}} + \sqrt{\frac{2\ln(8/\delta)}{n}} \leq \frac{\alpha^4\beta}{420}, \tag{60}$$

$$4\sqrt{\frac{r}{n}} + 4\sqrt{\frac{2\ln(16/\delta)}{n}} \leq \frac{\epsilon}{2}, \tag{61}$$

$$40p + \frac{2}{r} + 8\sqrt{\frac{(8r+8)\log_2(n)}{n}} + 2\sqrt{\frac{2\ln(16/\delta)}{n}} \leq \frac{\epsilon}{2}. \tag{62}$$

We first note that Eq. (62) implies that $40p \leq \epsilon/2$ and in particular $p \leq \epsilon/16$, which establishes the second inequality in Eq. (59). For the first inequality in Eq. (59), we compute

$$\sqrt{\frac{\ln(64/\delta)}{2n}} \leq \frac{1}{r^{0.25}} \cdot \frac{\sqrt{\ln(64/\delta)}}{r^{0.25}} \leq \frac{1800q(d) + 500c_2\sqrt{\ln(64n/\delta)}}{r^{0.25}} = p,$$

where in the above, the first inequality is due to $2n \geq r$ and the second inequality is due to our assumption in the theorem statement which implies $r^{0.25} \geq \max\{q(d), c_2\}$ if $c > 0$ is sufficiently large. Next, we compute

$$\frac{\ln(64n/\delta)}{\sqrt{r}} = \frac{1}{r^{0.4}}\left(\frac{\ln(64c)}{r^{0.1}} + \frac{\ln(r^3\log_2^2(r))}{r^{0.1}}\right) + \frac{\ln(1/\delta)}{\sqrt{r}}.$$

In the above we first have from the assumption in the theorem statement that $r^{-0.4} \leq c^{-0.4}\epsilon^2$. We can therefore bound the parenthesized expression by an absolute constant to establish that it can be made arbitrarily small. Upper bound the term

$$\frac{\ln(64c)}{r^{0.1}} \leq \frac{\ln(64c)}{c^{0.1}}\sqrt{\epsilon} \leq 6,$$

which follows from $\sup_x \ln(64x)/x^{0.25} \leq 6$ and $\epsilon \leq 1$. For the second term in the parentheses, we compute

$$\frac{\ln(r^3\log_2^2(r))}{r^{0.1}} = \frac{3\ln(r)}{r^{0.1}} + \frac{2\ln(\log_2(r))}{r^{0.1}} \leq 12 + 3 = 15,$$

which follows from $\sup_x 3\ln(x)/x^{0.1} \le 12$ and $\sup_x 2\ln(\log_2(x))/x^{0.1} \le 3$. The bound $\ln(1/\delta)/\sqrt{r} \le \frac{\epsilon^2}{\sqrt{c}}$ is immediate from the assumption in the theorem statement that $r \ge c\epsilon^{-5}\ln^2(1/\delta) \ge c\epsilon^{-4}\ln^2(1/\delta)$. Combining the bounds on the three terms we have shown that

$$\frac{\ln(64n/\delta)}{\sqrt{r}} \le \epsilon^2 \left(\frac{21}{c^{0.4}} + \frac{1}{\sqrt{c}}\right). \tag{63}$$

Squaring the above, rearranging and scaling c to be sufficiently large (while recalling that $\epsilon \le 1$) implies Eq. (58).

To establish the remaining inequalities in Eqs. (60-62), we would first need to upper bound the following two terms:

$$2\sqrt{\frac{2\ln(16/\delta)}{n}} \le \frac{2}{r}\sqrt{\frac{2\ln(16/\delta)}{r}} \le \frac{2\epsilon}{c}, \tag{64}$$

where in the above, the first inequality is by $n \ge r^3$ for $c > 1$ and the second inequality is by our assumption that $r \ge 2\ln(16/\delta)$ when c is sufficiently large, and that $r \ge c\epsilon^{-5} \ge c\epsilon^{-1}$. Next, we upper bound the term

$$4r\sqrt{\frac{(8r+8)\log_2(n)}{n}} = 4\sqrt{\frac{(8r+8)\log_2\left(cr^3\log_2^2(r)\right)}{cr\log_2^2(r)}}$$

$$\le \frac{16}{\sqrt{c}}\sqrt{\frac{\log_2\left(cr^3\log_2^2(r)\right)}{\log_2^2(r)}}$$

$$= \frac{16}{\sqrt{c}}\sqrt{\frac{\log_2^2(c) + 3\log_2(r) + 2\log_2(\log_2(r))}{\log_2^2(r)}}.$$

By the change of variables $x = \log_2(r)$ we have that the above equals

$$\frac{16}{\sqrt{c}}\sqrt{\frac{\log_2^2(c) + 3x + 2\log_2(x)}{x^2}} \le \frac{16}{\sqrt{c}}\sqrt{\log_2^2(c) + 3 + 1} \le \frac{16}{c^{0.25}}\sqrt{5+3+1} \le \frac{48}{c^{0.25}},$$

where the first inequality follows by our assumption $r \ge 2$ which implies $\log_2(r) \ge 1$ and thus $3/x \le 3$, and since $\sup_x 2\log_2(x)/x^2 \le 1$; and the second inequality follows from $\sup_x \log_2^2(x)/\sqrt{x} \le 5$. Combining this with Eq. (64), we can thus take $c > 0$ sufficiently large so that the left-hand side of Eq. (60) is smaller than any constant, thus satisfying the inequality in the aforementioned equation. Moreover, in the above we have also shown (when combined with Eq. (64) again) that

$$8\sqrt{\frac{(8r+8)\log_2(n)}{n}} + 2\sqrt{\frac{2\ln(16/\delta)}{n}} \le \frac{96}{c^{0.25}r} + \frac{2\epsilon}{c} \le \frac{96\epsilon}{c^{1.25}} + \frac{2\epsilon}{c},$$

where the second inequality uses our assumption in the theorem statement that $r \ge c\epsilon^{-1}$. Thus, for sufficiently large $c > 0$ we have that the above is at most $\epsilon/4$ which upper bounds the last two summands in the left-hand side of Eq. (62) and also implies Eq. (61). It only remains to show that $40p + 2/r \le \epsilon/4$ to establish the last remaining inequality in Eq. (62). To this end, compute

$$p = \frac{1800q(d) + 500c_2\sqrt{\ln(64n/\delta)}}{r^{0.25}} = \frac{1800q(d)}{r^{0.25}} + 500c_2\sqrt{\frac{\ln(64n/\delta)}{\sqrt{r}}}.$$

Bounding the term $1800q(d)/r^{0.25}$, we have from our assumption in the theorem statement $r \geq c\epsilon^{-4}q^4(d)$ that this is at most $1800c^{-0.25}\epsilon$. Using Eq. (63) to bound the term inside the square root, we obtain

$$p \leq 1800c^{-0.25}\epsilon + 500c_2\sqrt{\frac{21}{c^{0.4}} + \frac{1}{\sqrt{c}}} \cdot \epsilon,$$

which for a sufficiently large $c > 0$ is smaller than $\epsilon/320$ and therefore $40p \leq \epsilon/8$. The $2/r$ term is also at most $\epsilon/8$ by our assumption, concluding the derivation of the inequalities in Eqs. (58-62).

Turning to the proof of the theorem, we can apply Thms. (9,13) and Propositions (11,12) with a confidence parameter of $\delta/4$ in each, by using Eqs. (58,59), and further assume that their implications hold throughout the remainder of the proof. Note that this happens with probability of at least $1 - \delta$ by a union bound.

We first show that for some $t \in \{0, 1, \ldots, T\}$, GD finds a point $\mathbf{w}_t$ such that

$$\hat{F}(\mathbf{w}_t) \leq \frac{\epsilon}{2}. \tag{65}$$

If the assumption in Proposition 11 is violated by $\mathbf{w}_0$, then we have from Proposition 12 and Eq. (62) that $\hat{F}(\mathbf{w}_0) \leq \frac{\epsilon}{2}$. Otherwise, we have from Thm. 10 that $\nu = \frac{\alpha^4\beta}{840} \leq 1$, and thus from Proposition 11 and Eq. (60) we get

$$\langle \nabla F(\mathbf{w}_0), \mathbf{w}_0 - \mathbf{v} \rangle \geq \nu \|\mathbf{w}_0 - \mathbf{v}\|^2,$$

With the above inequality, we can follow a similar approach as in Yehudai and Shamir (2020, Thm. 5.3(2)) and deduce that for $t = 0$,

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{v}\|^2 &= \|\mathbf{w}_t - \eta\nabla F(\mathbf{w}_t) - \mathbf{v}\|^2 \\
&= \|\mathbf{w}_t - \mathbf{v}\|^2 - 2\eta \langle \nabla F(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v} \rangle + \eta^2 \|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{v}\|^2 (1 - \eta\nu) + \eta^2 \|\nabla F(\mathbf{w}_t)\|^2.
\end{aligned}$$

We shall now derive a bound on the norm of the gradient in the above expression as follows

$$\begin{aligned}
\|\nabla F(\mathbf{w}_t)\| &= \left\| \frac{2}{n} \sum_{i=1}^n \left(\left[\mathbf{w}_t^\top \tilde{\mathbf{x}}_i\right]_+^1 - y_i \right) \cdot \mathbb{1}\left\{\mathbf{w}_t^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \cdot \tilde{\mathbf{x}}_i \right\| \\
&\leq \frac{2}{n} \sum_{i=1}^n \left| \left(\left[\mathbf{w}_t^\top \tilde{\mathbf{x}}_i\right]_+^1 - y_i \right) \cdot \mathbb{1}\left\{\mathbf{w}_t^\top \tilde{\mathbf{x}}_i \in (0,1)\right\} \right| \cdot \|\tilde{\mathbf{x}}_i\| \leq 2\sqrt{r}.
\end{aligned}$$

By our assumption $\eta < \frac{\nu}{8r}$ we have

$$1 - \eta\nu + 4\eta^2 r < 1 - \frac{\eta\nu}{2} < 1,$$

which implies

$$\|\mathbf{w}_{t+1} - \mathbf{v}\|^2 \leq (1 - \eta\nu + 4\eta^2 r) \|\mathbf{w}_t - \mathbf{v}\|^2 \leq \left(1 - \frac{\eta\nu}{2}\right) \|\mathbf{w}_t - \mathbf{v}\|^2. \tag{66}$$

In particular $\|\mathbf{w}_{t+1} - \mathbf{v}\|^2 \leq \|\mathbf{w}_t - \mathbf{v}\|^2 < \|\mathbf{v}\|^2$, and therefore

$$\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_{t+1} - \mathbf{v}\| + \|\mathbf{v}\| \leq 2\|\mathbf{v}\| \leq 2r^{-0.25}.$$

Now, if the inequality in Eq. (7) is violated, then by Proposition 12 and Eq. (62) we have that $\hat{F}(\mathbf{w}_1) \leq \frac{\epsilon}{2}$ and thus we have shown that Eq. (65) holds for $t = 1$. Otherwise, we have that all the conditions in Proposition 11 hold for $\mathbf{w}_{t+1}$, and we can therefore apply both propositions iteratively until Eq. (7) is violated and Eq. (65) holds for some $t < T$, or until T iterations of GD have been performed. In the latter case, we can apply Eq. (66) iteratively to obtain

$$\|\mathbf{w}_T - \mathbf{v}\|^2 \leq \left(1 - \frac{\eta\nu}{2}\right)^T \|\mathbf{w}_0 - \mathbf{v}\|^2 \leq \exp\left(-\ln\left(\frac{r}{8\epsilon}\right)\right) \|\mathbf{w}_0 - \mathbf{v}\|^2 \leq \frac{\epsilon}{8r},$$

where the second inequality is due to $T = 2\eta^{-1}\nu^{-1}\ln(r/8\epsilon)$ and the inequality $(1 - 1/x)^x \leq \exp(-1)$ which holds for all $x \geq 1$, and the last inequality is due to $\|\mathbf{w}_t\|, \|\mathbf{v}\| \leq 0.5$ and the triangle inequality. Using the above, we can compute

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left[\mathbf{w}_T^\top\tilde{\mathbf{x}}_i\right]_+^1 - \left[\mathbf{v}^\top\tilde{\mathbf{x}}_i\right]_+^1\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}_T^\top\tilde{\mathbf{x}}_i - \mathbf{v}^\top\tilde{\mathbf{x}}_i\right)^2$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{w}_T - \mathbf{v}\|^2 \cdot \|\tilde{\mathbf{x}}_i\|^2 \leq \frac{\epsilon}{8},$$

where the first inequality uses the fact that $[\cdot]_+^1$ is 1-Lipschitz and the second inequality is by Cauchy-Schwartz. We can now bound the squared loss of $\mathbf{v}$ by using Thm. 9 to get

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left[\mathbf{v}^\top\tilde{\mathbf{x}}_i\right]_+^1 - y_i\right)^2 \leq \frac{1}{n}\sum_{i:\|\mathbf{x}_i\|\in I_2\cup I_4}1 \leq \frac{3600q(d) + 1000c_2\sqrt{\ln(32n/\delta)}}{r^{0.25}} \leq \frac{\epsilon}{8},$$

where the last inequality is implied by Eq. (59). Combining the last two displayed inequalities using the inequality $(a + b)^2 \leq 2a^2 + 2b^2$ with $a = \left[\mathbf{w}_T^\top\tilde{\mathbf{x}}_i\right]_+^1 - \left[\mathbf{v}^\top\tilde{\mathbf{x}}_i\right]_+^1$ and $b = \left[\mathbf{v}^\top\tilde{\mathbf{x}}_i\right]_+^1 - y_i$, we have that

$$\hat{F}(\mathbf{w}_T) \leq \frac{\epsilon}{2}.$$

We finish the proof by showing that this results in a generalization bound. Using Thm. 13 and Eqs. (61,65), we have for some $t \in \{0, 1, \ldots, T\}$ that

$$F(\mathbf{w}_t) \leq \hat{F}(\mathbf{w}_t) + 4\sqrt{\frac{r}{n}} + 4\sqrt{\frac{2\ln(16/\delta)}{n}} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

The proof of the theorem is complete. ∎

Appendix H. Learning Rate Stability is Insufficient for Efficient Learning

In this appendix, we show that although the stability of the learning rate is sufficient to guarantee the asymptotic convergence of GD, it is not sufficient for convergence in polynomial time. For an L-smooth function f, it is well-known that GD with a fixed and stable step size η, meaning that $\eta < \frac{2}{L}$, converges to a local minimum asymptotically (e.g. Nesterov et al. (2018, Thm. 2.1.14)). Moreover, there are examples for which no non-trivial guarantee can be given if $\eta \geq \frac{2}{L}$. We now provide an example of a 2-smooth function where $T \geq 4$ iterations of GD with a step size of $\eta > \frac{2}{L} - \frac{1}{T}$ result in only constant progress towards the global minimum.

Consider the objective $f(x) = x^2$ with the global minimum $x^* = 0$. Then the gradient step update for $\eta > \frac{2}{L} - \frac{1}{T} = 1 - \frac{1}{T} \geq \frac{3}{4}$ is given by

$$x_{t+1} = x_t - 2\eta x_t = (1 - 2\eta)x_t.$$

Iteratively applying the above recursion we obtain

$$|x_T - x^*| = |x_T| = \left|(1 - 2\eta)^T x_0\right| = (2\eta - 1)^T |x_0| \geq \left(1 - \frac{2}{T}\right)^T |x_0| \geq \frac{1}{16} |x_0|.$$

In the above, the third equality is due to $\eta \geq \frac{3}{4}$ which guarantees that $2\eta - 1 > 0$, the first inequality follows from the fact that $(2\eta - 1)^T$ is increasing for all $\eta > 0.5$, and the last inequality is due to $(1 - 2/x)^x \geq \frac{1}{16}$ for all $x \geq 4$.

It follows from the above that GD with a step size larger than $\frac{2}{L} - \frac{1}{T}$ cannot get to less than a constant fraction of its initialization distance from the global minimizer. Thus, even if we allow the step size to be just a small margin away from the edge of stability $2/L$, convergence cannot be guaranteed in general. This results in the necessary condition that $\eta \leq \frac{2}{L} - \frac{1}{T}$ which is a stronger assumption than merely requiring that the learning rate is stable. We also remark that while the example given in this appendix is of a strongly convex function, our analysis holds locally and therefore also applies in the non-convex regime, given that the objective possesses a local minimum with a finite smoothness parameter L in its neighborhood.