Proceedings of Machine Learning Research vol 195:1–40, 2023 36th Annual Conference on Learning Theory

Intrinsic dimensionality and generalization properties of the $\mathcal{R}$-norm inductive bias

Navid Ardeshir NA2844@COLUMBIA.EDU
Dept. of Statistics, Columbia University

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Dept. of Computer Science, Columbia University

Clayton Sanford CLAYTON@CS.COLUMBIA.EDU
Dept. of Computer Science, Columbia University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study the structural and statistical properties of $\mathcal{R}$-norm minimizing interpolants of datasets labeled by specific target functions. The $\mathcal{R}$-norm is the basis of an inductive bias for two-layer neural networks, recently introduced to capture the functional effect of controlling the size of network weights, independently of the network width. We find that these interpolants are intrinsically multivariate functions, even when there are ridge functions that fit the data, and also that the $\mathcal{R}$-norm inductive bias is not sufficient for achieving statistically optimal generalization for certain learning problems. Altogether, these results shed new light on an inductive bias that is connected to practical neural network training.

Keywords: interpolation, variational norm, generalization, inductive bias

1. Introduction

The study of inductive biases in neural network learning is important for theoretical understanding and for developing practical guidance in network training. Recent theories of generalization rely on inductive biases of training algorithms to explain how neural nets that (nearly) interpolate training data can be accurate out-of-sample (Neyshabur et al., 2015; Zhang et al., 2021). When inductive biases are made explicit and their effects are elucidated, they can be incorporated into training procedures when deemed appropriate for a problem at hand.

In this paper, we study the inductive bias for two-layer neural networks implied by a variational norm called the $\mathcal{R}$*-norm*, introduced by Savarese et al. (2019) and Ongie et al. (2019) to capture the functional effect of controlling the size of network weights. (A definition is given in Section 2.2.) We focus on the *approximation* and *generalization* consequences of preferring networks with small $\mathcal{R}$-norm in the context of learning explicit target functions. It is well-known that the size of the weights can play a critical role in generalization properties of neural networks (Bartlett, 1996), and weight-decay regularization is a common practice in gradient-based training (Hinton, 1987; Hanson and Pratt, 1988). Thus, explicating the consequences of the $\mathcal{R}$-norm inductive bias may advance our understanding of generalization in practical settings.

We investigate the d-dimensional variational problem (VP), which seeks a neural net $g\colon \Omega \to \mathbb{R}$ of minimum $\mathcal{R}$-norm among those that perfectly fit a given labeled dataset $\{(x_i, y_i)\}_{i\in[n]} \subset \Omega \times \mathbb{R}$:

$$\inf_{g\colon \Omega\to\mathbb{R}} \|g\|_{\mathcal{R}} \quad \text{s.t.} \quad g(x_i) = y_i \qquad \forall i \in [n]; \tag{VP}$$

as well as a variant (ϵ-VP) that only requires g to uniformly approximate labels up to error $\epsilon \in (0, 1)$:

$$\inf_{g\colon \Omega\to\mathbb{R}} \|g\|_{\mathcal{R}} \quad \text{s.t.} \quad |g(x_i) - y_i| \leq \epsilon \quad \forall i \in [n]. \tag{ϵ-VP}$$

Here, $\Omega \subset \mathbb{R}^d$ is a d-dimensional domain of interest. We study structural and statistical properties of solutions to these problems for datasets labeled by specific target functions in high dimensions.

The recent introduction of the $\mathcal{R}$-norm and its connections to weight-decay regularization have catalyzed research on the foundational properties of solutions to (VP). In particular, solutions in the one-dimensional ($d = 1$) setting have been precisely characterized and their generalization properties are now well-understood by their connections to splines (Debarre et al., 2022; Savarese et al., 2019; Parhi and Nowak, 2021a; Hanin, 2021). However, far less is known about the solutions of $\mathcal{R}$-norm-minimizing interpolants for the general d-dimensional case.

Key message. Inductive biases based on certain variational norms, such as the $\mathcal{R}$-norm, are believed to offer a way around the curse of dimensionality suffered by kernel methods, because they are adaptive to low-dimensional structure. Researchers have pointed to this adaptivity property in non-parametric settings (Bach, 2017; Parhi and Nowak, 2021b) and specific learning tasks with low-dimensional structure (Wei et al., 2019) as mathematical evidence of the statistical advantage of neural networks over kernel methods. One may hypothesize that the $\mathcal{R}$-norm inductive bias achieves this advantage by favoring functions with low-dimensional structure. Indeed, many other forms of inductive bias used in statistics and machine learning are known to explicitly identify relevant low-dimensional structure (Candès et al., 2006; Donoho, 2006; Candès and Recht, 2009; Bhojanapalli et al., 2016; Barak et al., 2022; Damian et al., 2022; Frei et al., 2022; Mousavi-Hosseini et al., 2022; Galanti et al., 2022). Our results provide theoretical evidence that this is not always the case with the $\mathcal{R}$-norm inductive bias, in a very strong sense that becomes more pronounced in higher dimensions.

We show that, even in cases where the dataset can be perfectly fit by an intrinsically one-dimensional function, the solutions g to (VP) or (ϵ-VP) are not necessarily the piecewise-linear ridge functions described in previous works (Savarese et al., 2019; Hanin, 2021). Rather, the $\mathcal{R}$-norm is far better minimized by a *multi-directional*[1] neural network g that averages several ridge functions pointing in different directions, each of which approximates a small fraction of the data.

1.1. Our contributions

Our results are summarized by the following informal theorems concerning the structural and generalization properties of $\mathcal{R}$-norm interpolation. Together, they show that the $\mathcal{R}$-norm inductive bias (1) leads to interpolants that are qualitatively different from those that minimize width or intrinsic dimensionality of the learned network, and (2) is insufficient for obtaining optimal generalization for a well-studied learning problem.

Informal Theorem 1 ($\mathcal{R}$-norm minimizers of the parity dataset are not ridge functions) *Suppose the dataset $\{(x_i, y_i)\}_{i\in[n]} \subset \{\pm 1\}^d \times \{\pm 1\}$ used in (VP) and (ϵ-VP) is the complete dataset of 2^d examples labeled by the d-variable parity function.*

- *The optimal value of (VP) is $\Theta(d)$.*

1. By a multi-directional function, we mean a function that does not *only* depend on a one-dimensional projection of its input—i.e., a function that is not a ridge function (defined in Section 2.1).

- *The optimal value of* (ϵ-VP) *for any* $\epsilon \in [0, 1/2)$*—with the additional constraint that* g *be a ridge function—is* $\Theta(d^{3/2})$.

This result is presented formally in Section 3. In Section 3.1, we show that every ridge function satisfying the constraints of (ϵ-VP) has $\mathcal{R}$-norm at least $\Omega(d^{3/2})$; this bound is tight for ridge functions, as there is a matching upper bound. Using an averaging strategy, we show in Section 3.2 the existence of multi-directional interpolants g of the parity dataset with $\|g\|_{\mathcal{R}} = O(d)$, and we also establish the optimality of this construction in Section 3.3. These results characterize the optimal value of (VP) in terms of the dimension d, and also establish the $\mathcal{R}$-norm-suboptimality of ridge function interpolants. (In Section 5, we extend the averaging strategy to other types of target functions, expanding the scope of our structural findings.)

Informal Theorem 2 (Min-$\mathcal{R}$-norm interpolation is sub-optimal for learning parities) *Suppose the dataset* $\{(\mathbf{x}_i, \chi(\mathbf{x}_i))\}_{i\in[n]} \subset \{\pm 1\}^d \times \{\pm 1\}$ *used in* (VP) *is an i.i.d. sample, where* $\mathbf{x}_i \sim \mathrm{Unif}(\{\pm 1\}^d)$ *is labeled by the* d*-variable parity function* χ *for all* $i \in [n]$*. If the sample size is* $n = o(d^2/\sqrt{\log d})$*, then with probability at least* $1/2$*, every solution to* (VP) *has mean squared error at least* $1 - o(1)$ *for predicting* χ *over* $\mathrm{Unif}(\{\pm 1\}^d)$.

This result is presented formally in Section 4.1, and it is complemented by a sample complexity upper bound in Section 4.2. The results are stated for the parity function on all d variables, but the same holds for any parity function over $\Omega(d)$ variables. It is well-known that an i.i.d. sample of size $O(d)$ is sufficient for learning parity functions exactly (Helmbold et al., 1992; Fischer and Simon, 1992), and hence we conclude that the $\mathcal{R}$-norm inductive bias is insufficient for achieving the statistically optimal sample complexity for this learning problem.

1.2. Related work

Variational norms and inductive biases of optimization methods. Many variational norms (such as $\mathcal{R}$-norm) from functional analysis can be regarded as representational costs that induce topologies on the space of infinitely-wide neural networks with certain activation functions. Prior works have analyzed these norms for homogeneous activation functions like ReLU (e.g., Kurková and Sanguineti, 2001; Mhaskar, 2004; Bach, 2017; Savarese et al., 2019; Ongie et al., 2019); see Siegel and Xu (2021) and references therein for a comparison. In particular, the work of Ongie et al. (2019) provided analytical descriptions of $\mathcal{R}$-norm in terms of the Radon transform of the function itself. This work was extended to higher powers of ReLU by Parhi and Nowak (2021a).

The variational norms are also connected to the implicit biases of optimization methods for training neural networks. In the context of univariate functions, the dynamics of gradient descent was shown to be biased towards (adaptive) linear or cubic spline depending on the optimization regime (Williams et al., 2019; Shevchenko et al., 2021; Maennel et al., 2018), and these results have been partially extended to the multivariate case (Jin and Montúfar, 2020). For classification problems, the implicit bias of gradient descent was connected to a variational problem related to $\mathcal{R}$-norm with margin constraints on the data (Bach and Chizat, 2021).

Solutions to the variational problem. Debarre et al. (2022) and Hanin (2021) fully characterized the form of all solutions of (VP) for one-dimensional datasets (as discussed above). However, pinning down even a single solution for general multidimensional datasets appears to be difficult; Ergen and Pilanci (2021) was able to do so for rank-one datasets, where all the feature vectors lie on a line. The datasets we study do not satisfy the rank-one condition of Ergen and Pilanci (2021), and thus we require different techniques to analyze multi-directional functions.

Adaptivity. In the context of non-parametric regression, it is well-known that (deep) neural networks achieve minimax-optimal rates in the presence of low-dimensional structure in the target function (e.g., Schmidt-Hieber, 2020; Bauer and Kohler, 2019; Kohler and Krzyżak, 2005; Györfi et al., 2002). The convergence rates in these works depend only on the intrinsic dimension of the target function (and not the ambient dimension) and are achieved by optimally trading off accuracy and regularization in certain deep neural network architectures. Recent works (Klusowski and Barron, 2016; Parhi and Nowak, 2021b; Bach, 2017) consider two-layer neural networks with variational norm (similar to $\mathcal{R}$-norm) regularization, which also allows for adaptivity to low-dimensional structure. That is, a function $g\colon \mathbb{R}^d \to \mathbb{R}$ depending only on a k-dimensional projection of its input x, i.e., $g(x) = \phi(Ux)$ for some $U \in \mathbb{R}^{k\times d}$ (with orthonormal rows) and $\phi\colon \mathbb{R}^k \to \mathbb{R}$ has variational norm no greater than that of the corresponding low-dimensional function ϕ (Bach, 2017). In particular, Bach (2017) and Klusowski and Barron (2016) studied minimax rates under ridge target functions where $k = 1$. Our results on generalization are of a different flavor: rather than striking a careful balance between fitting and regularization to achieve minimax rates, we study the behavior of $\mathcal{R}$-norm-minimizing interpolation.

Regularization based on weight decay (equivalent to $\mathcal{R}$-norm for shallow networks) has also been used to obtain minimax rates for learning smooth target functions. Parhi and Nowak (2021b) do so by drawing analogies to spline theory, while Wang and Lin (2021) consider a connection to the Group Lasso. Zhang and Wang (2022) exploits depth to promote stronger sparsity regularizes. This is distinct from the low-dimensional structures studied in this work and mentioned above.

Learning ridge functions and parity functions with neural nets. Target functions that depend on low-dimensional projections of the input (of which ridge functions are the simplest case) have been long studied in statistics (see, e.g., Li, 2018), and learning such functions is one of the simplest problems where neural network training demonstrates adaptivity. Such demonstrations typically require going beyond the neural tangent kernel regime and have been used to explain the "feature learning" ability of neural networks (Frei et al., 2022; Damian et al., 2022; Mousavi-Hosseini et al., 2022; Bietti et al., 2022). Several recent works have considered the prospects of learning (sparse) parity functions by training neural nets with gradient-based algorithms (Abbe and Sandon, 2020; Daniely and Malach, 2020; Malach et al., 2021; Barak et al., 2022; Telgarsky, 2022). The positive results express parities as low-weight linear combinations of (random) ReLUs, which motivates our focus on the variational norm of approximating neural nets. Our sample complexity lower bound shows that, even if computational and optimization considerations are set aside, the inductive bias imposed by the $\mathcal{R}$-norm may lead to suboptimal statistical performance.

Averaging and ensembling. Neural networks have been interpreted as forms of averaging or ensemble methods to explain their statistical properties (e.g., Bartlett, 1996; Baldi and Sadowski, 2013; Gal and Ghahramani, 2016; Olson et al., 2018). Our use of averaging is different in that it serves as an approximation-theoretic mechanism for achieving smaller $\mathcal{R}$-norm.

Weight lower bounds for other explicit functions. Representation costs for two-layer neural networks to approximate other explicit functions have been explored in several prior works (Martens et al., 2013; Daniely, 2017; Safran and Shamir, 2017; Safran et al., 2019). These works establish exponential lower-bounds on the width of two-layer networks needed to approximate functions that are represented more compactly by three-layer networks. These results also imply lower-bounds on the size of second-layer weights in a two-layer network after fixing the width of the network. In contrast, our results hold for networks of unbounded width and for a target function that can be exactly represented by a two-layer networks of $\mathrm{poly}(d)$ width.

2. Preliminaries

2.1. Notation

In this work, we consider real-valued functions over the radius-$\sqrt{d}$ Euclidean ball $\Omega := \{x \in \mathbb{R}^d : \|x\|_2 \leq \sqrt{d}\}$. Let $\chi_S \colon \Omega \to \{\pm 1\}$ denote the multi-linear monomial $\chi_S(x) := \prod_{i \in S} x_i$ over variables indexed by $S \subseteq [d]$, and let $\chi := \chi_{[d]}$. On input $x \in \{\pm 1\}^d$, $\chi_S(x)$ computes the *parity* of $\{x_i : i \in S\}$. We say $g \colon \Omega \to \mathbb{R}$ is a *ridge function* if $g(x) = \phi(v^\intercal x)$ for some unit vector $v \in \mathbb{S}^{d-1}$ and function $\phi \colon [-\sqrt{d}, \sqrt{d}] \to \mathbb{R}$. A function ϕ is ρ*-periodic* if $\phi(z + \rho) = \phi(z)$ for all $z \in \mathbb{R}$.

We consider two-layer neural networks (of infinite and finite width) with ReLU activations $\varphi_{\mathrm{r}}(z) := \max\{0, z\}$. Let $\mathcal{M}$ denote the space of signed measures over $\mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}]$. For $\mu \in \mathcal{M}$, let $g_\mu : \Omega \to \mathbb{R}$ denote the infinite-width neural network given by

$$g_\mu(x) := \int_{\mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}]} \varphi_{\mathrm{r}}(w^\intercal x + b)\, \mu(\mathrm{d}w, \mathrm{d}b).$$

The total variation norm of μ is $|\mu| := \int_{\mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}]} |\mu|\,(\mathrm{d}w, \mathrm{d}b)$, where $|\mu|\,(\mathrm{d}w, \mathrm{d}b)$ is the corresponding total variation measure (somewhat abusing notation). The width-m neural network g_θ with parameters $\theta = (a^{(j)}, w^{(j)}, b^{(j)})_{j \in [m]} \in (\mathbb{R} \times \mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}])^m$ is given by

$$g_\theta(x) := \sum_{j=1}^{m} a^{(j)} \varphi_{\mathrm{r}}(w^{(j)\intercal} x + b^{(j)}).$$

We regard g_θ as an infinite-width neural network with the "sparse" atomic measure

$$\mu_\theta = \sum_{j=1}^{m} a^{(j)} \delta_{(w^{(j)}, b^{(j)})}.$$

Observe that $g_\theta = g_{\mu_\theta}$ and $|\mu_\theta| = \sum_{j=1}^m |a^{(j)}| = \|a\|_1$.

Our constructions frequently use *sawtooth functions*, a family of ridge functions that are composed of $t+1$ repetitions of a triangular wave that draw inspiration from a construction of Yehudai and Shamir (2019, Proposition 4.2). For $t \in \{0, \ldots d\}$ with $t \equiv d \pmod 2$ and $w \in \{\pm 1\}^d$, let $s_{w,t}(x) := (-1)^{d-t} \chi(w) \phi_t(w^\intercal x)$ where $\phi_t : \mathbb{R} \to \mathbb{R}$ is a function that forms a piecewise affine interpolation between the points $(-t-1, 0)$, $\{(t - 2\tau, (-1)^\tau)\}_{\tau \in \{0, \ldots, t\}}$, $(t+1, 0)$, and $\phi_t(z) = 0$ for all $z \leq -t-1$ and $z \geq t+1$. We refer to t as the *width* of the sawtooth function $s_{w,t}$. Note that $s_{w,t}$ is $\sqrt{d}$-Lipschitz and $s_{w,t}(x) = \chi(x) \mathbb{1}\{|w^\intercal x| \leq t\}$ for all $x \in \{\pm 1\}^d$. Also, $s_{w,t}$ can be expressed as a neural network g_θ with width $m \leq O(t+1)$ and $|a^{(i)}| \leq O(\sqrt{d})$ for each $i \in [m]$.

Let $\nu := \mathrm{Unif}(\{\pm 1\}^d)$ denote the uniform distribution on $\{\pm 1\}^d$, and let $\boldsymbol{\nu}_n$ denote the empirical distribution on $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim_{\mathrm{iid}} \nu$. We use the following inner products and norms over the vector space of real-valued functions on $\{\pm 1\}^d$ with respect to a distribution ν_0 (such as ν or $\boldsymbol{\nu}_n$):

$$\langle g, h \rangle_{L^2(\nu_0)} := \mathop{\mathbb{E}}_{\mathbf{x} \sim \nu_0}[g(\mathbf{x}) h(\mathbf{x})], \quad \|g\|_{L^2(\nu_0)} := \langle g, g \rangle_{L^2(\nu_0)}^{1/2}, \quad \|g\|_{L^\infty(\nu_0)} := \max_{x \in \mathrm{supp}(\nu_0)} |g(x)|.$$

2.2. $\mathcal{R}$-norm and attainment of the infimum

We now recall the definition of the $\mathcal{R}$-norm of a function $g\colon \Omega \to \mathbb{R}$, presented here in a variational form as the minimum cost of representing g as an infinite-width neural network with a "skip-connection":

$$\|g\|_{\mathcal{R}} := \inf_{\mu\in\mathcal{M}, v\in\mathbb{R}^d, c\in\mathbb{R}} |\mu| \quad \text{s.t.} \quad g(x) = g_\mu(x) + v^\top x + c \ \ \forall x \in \Omega. \tag{$\mathcal{R}$-norm}$$

Indeed, $\|\cdot\|_{\mathcal{R}}$ is a semi-norm on the space of functions with finite $\mathcal{R}$-norm. It was initially introduced by Ongie et al. (2019) along with explicit characterizations in terms of the Radon transform. See the works of Ongie et al. (2019), Parhi and Nowak (2021a), and Siegel and Xu (2021) for more discussion about the $\mathcal{R}$-norm and its connections to other function spaces.

The appearance of the affine component $v^\top x + c$ in the definition of $\mathcal{R}$-norm has implications for how the bias terms are treated. Notice that a neuron $x \mapsto \varphi_{\mathrm{r}}(w^\top x + b)$ with bias $|b| \geq \sqrt{d}$ behaves as an affine function over the domain of interest Ω, so it can be absorbed into the "free" affine component (in the definition of $\mathcal{R}$-norm) so as to not be counted towards the $\mathcal{R}$-norm. Other works (e.g., Siegel and Xu, 2021) consider a different variational norm, $\|\cdot\|_{\mathscr{V}_2}$, which does not have "free" affine components, but instead permits biases b to be in the larger range $[-2\sqrt{d}, 2\sqrt{d}]$. These differences in the way affine components are accommodated do not lead to different function spaces (see Parhi and Nowak, 2021b, Theorem 6), and the results of this paper for $\mathcal{R}$-norm also hold for these other variational norms, as we demonstrate in Appendix C.

Although the $\mathcal{R}$-norm is defined in terms of an infimum, it has been shown by Parhi and Nowak (2021b, Lemma 2; see also Proposition 25 in Appendix B) that the infimum is always achieved by a particular signed measure $\mu \in \mathcal{M}$. Since the total variation norm is sparsity-inducing, the objective in (VP) favors finite-width networks. It can be shown, using an extension of Caratheodory's theorem (Rosset et al., 2007), that (VP) in fact always has a finite-width solution. That is, (VP) is solved by the sum of an affine function $x \mapsto v^\top x + c$ and a width-m neural network, for some $m \leq \max\{0, n - (d+1)\}$. This claim is formalized as Theorem 26 and proved in Appendix B. Thus, considering finite-width neural networks is sufficient to determine the value of (VP).[2]

The following lemma, which is a minor elaboration on Lemma 25 of Parhi and Nowak (2021a), relates the $\mathcal{R}$-norm of a finite-width network to the ℓ_1-norm of its top-layer weights.

Lemma 1 *Let $v \in \mathbb{R}^d$, $c \in \mathbb{R}$, and $\theta = (a^{(j)}, w^{(j)}, b^{(j)})_{j\in[m]} \in (\mathbb{R} \times \mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}])^m$ be the set of parameters of a finite neural network where $(w^{(i)}, b^{(i)}) \neq (w^{(j)}, b^{(j)})$ for all $i \neq j$.*

(i) The $\mathcal{R}$-norm of the sum of g_θ and an affine function $v^\top x + c$ satisfies

$$\|g_\theta(x) + v^\top x + c\|_{\mathcal{R}} \leq \|a\|_1 = |a^{(1)}| + \cdots + |a^{(m)}|. \tag{1}$$

(ii) Moreover, if the measure μ_θ is even in a distributional sense (i.e., $\mu_\theta(w, b) = \mu_\theta(-w, -b)$), then the inequality in (1) holds with equality.

Note that our assumption that μ_θ is even in Lemma 1(ii) precludes the case where $a^{(i)} = -a^{(j)}$ and $(w^{(i)}, b^{(i)}) = (-w^{(j)}, -b^{(j)})$ for some $i \neq j$. This is needed because if such a case were allowed, we would have $a^{(i)}\varphi_{\mathrm{r}}(w^{(i)\top}x + b^{(i)}) + a^{(j)}\varphi_{\mathrm{r}}(w^{(j)\top}x + b_j) = a^{(i)}(w^{(i)\top}x + b^{(i)})$ for all $x \in \Omega$—an affine function. After ruling out these cases, we can apply the argument of Parhi and Nowak (2021a) to prove Lemma 1(ii).

2. We note that the finite-width solution to (VP) is not necessarily unique; Hanin (2021) discusses this issue in the one-dimensional case ($d = 1$) under general data models.

2.3. $\mathcal{R}$-norm of ridge functions

Prior works illuminate precise formulations of the $\mathcal{R}$-norm, and characterize solutions to (VP), albeit only for the one-dimensional setting (Hanin, 2021; Savarese et al., 2019; Ergen and Pilanci, 2021). These results are nevertheless useful for analyzing ridge functions in d-dimensional space.

Theorem 2 *For any ridge function $g\colon \Omega \to \mathbb{R}$ of the form $g(x) = \phi(w^\intercal x)$ where $w \in \mathbb{S}^{d-1}$ and $\phi\colon [-\sqrt{d}, \sqrt{d}] \to \mathbb{R}$ is Lipschitz, we have*

$$\|g\|_{\mathcal{R}} = \|\phi'\|_{\mathrm{TV}} := \operatorname*{ess\,sup}_{-\sqrt{d} \le t_0 < t_1 < \cdots < t_r \le \sqrt{d};\, r \in \mathbb{N}} \sum_{i=1}^{r} \left|\phi'(t_i) - \phi'(t_{i-1})\right|,$$

where ϕ' is a right continuous derivative of ϕ.[3]

Remark 3 *If ϕ is twice differentiable, then $\|g\|_{\mathcal{R}} = \int_{-\sqrt{d}}^{\sqrt{d}} |\phi''(u)|\,\mathrm{d}u = \|\phi''\|_1$. Intuitively, this ℓ_1-norm penalty induces sparsity in the second derivative, leading to representations that use few neurons. In contrast, minimizing the ℓ_2-norm penalty $\|\phi''\|_2$ on the second derivative yields a cubic spline (Kimeldorf and Wahba, 1971).*

Proof. Without loss of generality, g only depends on the first coordinate x_1 due to the invariance of the $\mathcal{R}$-norm to rotation (cf. Proposition 11 of Ongie et al., 2019). The result then follows from Remark 4 of Parhi and Nowak (2021b). ■

This bound on the $\mathcal{R}$-norm for ridge functions (and univariate functions) is critical for analyses of the solutions to (VP) for $d = 1$ (Hanin, 2021; Savarese et al., 2019). It suggests a potential approach for our high-dimensional setting: project the dataset to every one-dimensional subspace, interpolate the data with a ridge function that points in that direction, directly compute the $\mathcal{R}$-norm of each using Theorem 2, and return the ridge function with the lowest $\mathcal{R}$-norm. In the sequel, we examine the optimality of this approach, and find that ridge functions *cannot* be optimal solutions to (VP), even when the dataset can be perfectly fit by a ridge function.

3. Solutions to the variational problem for parity are multi-directional

In this section, we study the $\mathcal{R}$-norm of neural networks that solve VP or ϵ-VP for the (full) *parity dataset* $\{(x, \chi(x)) : x \in \{\pm 1\}^d\}$, which has size $n = 2^d$. For simplicity, the labels are provided by the parity function χ over all d variables, although the same quantitative results (up to constant factor differences) hold for any χ_S with $|S| = \Theta(d)$.

The high level message is that, despite the fact that this dataset can be exactly fit using ridge functions, the solutions to (VP) and (ϵ-VP) are *not* ridge functions and instead must be multi-directional.

3.1. Every ridge parity interpolant has $\mathcal{R}$-norm $\Omega(d^{3/2})$

We first show that any ϵ-approximate interpolant of the parity dataset *that is also a ridge function* must have $\mathcal{R}$-norm $\Omega(d^{3/2})$. This lower-bound is established even for $\epsilon = 1/2$.

3. Take $\phi'(u) = \lim_{t \downarrow 0} \frac{\phi(u+t) - \phi(u)}{t}$; the limit exists almost everywhere by Rademacher's theorem.

Theorem 4 *For $d \geq 2$, let Ridge_d be the set of functions $g \colon \Omega \to \mathbb{R}$ such that $g(x) = \phi(w^\top x)$ for some $w \in \mathbb{S}^{d-1}$ and Lipschitz continuous $\phi \colon [-\sqrt{d}, \sqrt{d}] \to \mathbb{R}$. Then*

$$\inf\{\|g\|_{\mathcal{R}} : g \in \mathrm{Ridge}_d,\ \|g - \chi\|_{L^\infty(\nu)} \leq 1/2\} \geq d^{3/2}/(2\sqrt{2}).$$

The proof constructs a labeled dataset of $d + 1$ points, and shows that any ridge function $g(x) = \phi(w^\top x)$ that approximates that dataset must have many high-magnitude oscillations. These oscillations imply a lower bound on $\|\phi'\|_{\mathrm{TV}}$, which proves the claim by way of Theorem 2.

Proof. Take any $g \in \mathrm{Ridge}_d$ of the form $g(x) = \phi(w^\top x)$ for some function ϕ and vector w satisfying the approximation constraint $\|g - \chi\|_{L^\infty(\nu)} \leq 1/2$. Suppose for sake of contradiction that $w_i = 0$ for some $i \in [d]$. Then, there exists a pair of points $x, x' \in \{-1, 1\}^d$ that are identical except in the i-th positions, x_i and x'_i. Thus, $\chi(x) = -\chi(x')$, but $w^\top x = w^\top x'$ and hence $g(x) = g(x')$; this contradicts the approximation constraint. So, we may henceforth assume that $w_i \neq 0$ for all $i \in [d]$.

For each $i \in \{0, 1, \ldots, d\}$, define

$$x^{(i)} := (\mathrm{sign}(w_1), \ldots, \mathrm{sign}(w_i), -\mathrm{sign}(w_{i+1}), \ldots, -\mathrm{sign}(w_d)).$$

Because the parity of $x^{(i)}$ alternates with i, i.e., $\chi(x^{(i)}) \neq \chi(x^{(i+1)})$, $\mathrm{sign}(g(x^{(i)}))$ also alternates because g satisfies the approximation constraint. Furthermore, again due to the approximation constraint, we have $|g(x^{(i)}) - g(x^{(i+1)})| \geq 1$. We claim that, because ϕ interpolates $d+1$ well-separated data points $(w^\top x^{(i)}, \phi(w^\top x^{(i)}))$ that satisfy $w^\top x^{(i)} < w^\top x^{(i+1)}$ for all $i \in \{0, 1, \ldots, d-1\}$, there must be a large cost for representing ϕ using a neural network. By Theorem 2, it suffices to obtain a lower bound on $\|\phi'\|_{\mathrm{TV}}$, since this will imply a lower bound on $\|g\|_{\mathcal{R}}$.

By Lemma 28 (in Appendix D.1; essentially the mean value theorem), for every $i \in \{0, 1, \ldots, d-1\}$, there exists $A_i \subseteq [w^\top x^{(i)}, w^\top x^{(i+1)}]$ with Lebesgue measure $\mathrm{Leb}(A_i) > 0$ such that, for every $z^{(i)} \in A_i$, we have

$$|\phi'(z^{(i)})| \geq \frac{1}{2} \cdot \frac{|\phi(w^\top x^{(i+1)}) - \phi(w^\top x^{(i)})|}{w^\top x^{(i+1)} - w^\top x^{(i)}},$$

and $\mathrm{sign}\left(\phi'(z^{(i)})\right) = \mathrm{sign}\left(\phi(w^\top x^{(i+1)}) - \phi(w^\top x^{(i)})\right)$. The fact that the signs of $\phi(w^\top x^{(i)})$ alternate with i implies that the signs of $\phi'(z^{(i)})$ also alternate with i. We now lower-bound the total variation of ϕ' using the fact that $\prod_{i=1}^d \mathrm{Leb}(A_i) > 0$ and taking advantage of the alternating signs:

$$\begin{aligned}
2\left\|\phi'\right\|_{\mathrm{TV}} &= 2 \operatorname*{ess\,sup}_{-\sqrt{d} \leq t_0 < t_1 < \cdots < t_r \leq \sqrt{d};\, r \in \mathbb{N}} \sum_{i=1}^{r} \left|\phi'(t_i) - \phi'(t_{i-1})\right| \\
&\geq 2 \sum_{i=1}^{d-1} |\phi'(z^{(i)}) - \phi'(z^{(i-1)})| = 2 \sum_{i=1}^{d-1} (|\phi'(z^{(i)})| + |\phi'(z^{(i-1)})|) \geq 2 \sum_{i=0}^{d-1} |\phi'(z^{(i)})| \\
&\geq \sum_{i=0}^{d-1} \frac{|\phi(w^\top x^{(i+1)}) - \phi(w^\top x^{(i)})|}{w^\top x^{(i+1)} - w^\top x^{(i)}} \geq \sum_{i=0}^{d-1} \frac{1}{w^\top x^{(i+1)} - w^\top x^{(i)}} \\
&\geq \frac{d^2}{\sum_{i=0}^{d-1} w^\top x^{(i+1)} - w^\top x^{(i)}} = \frac{d^2}{w^\top x^{(d)} - w^\top x^{(0)}} \geq \frac{d^2}{\|w\|_2 \|x^{(d)} - x^{(0)}\|_2} = \frac{d^{3/2}}{\sqrt{2}}.
\end{aligned}$$

The second-to-last inequality is a consequence of Cauchy-Schwarz: for any $a_1, \ldots, a_d > 0$, $d^2 = (\sum_i \sqrt{a_i}/\sqrt{a_i})^2 \leq (\sum_i a_i)(\sum_i 1/a_i)$. Therefore, $\|g\|_{\mathcal{R}} = \|\phi'\|_{\mathrm{TV}} \geq d^{3/2}/(2\sqrt{2})$. ∎

The lower-bound in Theorem 4 is tight up to constants, because the sawtooth function $s_{\mathbf{1},d}$ satisfies the constraints of (VP) and has $\|s_{\mathbf{1},d}\|_{\mathcal{R}} = O(d^{3/2})$.

3.2. Existence of a multi-directional parity interpolant with $\mathcal{R}$-norm $O(d)$

We now show that the $\Omega(d^{3/2})$ $\mathcal{R}$-norm lower-bound from Theorem 4 for ridge functions can be avoided by neural networks that are not ridge functions. The main idea is to employ an *averaging strategy* that combines a collection of distinct ridge functions, each of which perfectly fits a small fraction of the parity dataset—those on the "equator" relative to the ridge direction—while ignoring the "outliers" in that direction. Because all points on the cube are "outliers" for some directions and on the "equator" for others, this strategy ultimately ensures that every example is perfectly fit.

Theorem 5 *For any even*[4] d*, there exists a neural network* $g : \Omega \to \mathbb{R}$ *having* $g(x) = \chi(x)$ *for all* $x \in \{\pm 1\}^d$ *such that* $\|g\|_{\mathcal{R}} \leq O(d)$.

Proof. Recall that the sawtooth function $s_{w,0} : \Omega \to \mathbb{R}$ satisfies $s_{w,0}(x) = \chi(x)\mathbb{1}\{w^\top x = 0\}$ for all $x \in \{\pm 1\}^d$. By construction, $s_{w,0}$ is a ridge function that is a single "bump" around zero in the direction of w, and $\|s_{w,0}\|_{\mathcal{R}} \leq O(\sqrt{d})$. Consider $\mathbf{w} \sim \mathrm{Unif}(\{\pm 1\}^d)$. By symmetry, $\mathbb{P}[\mathbf{w}^\top x = 0] = \mathbb{P}[\mathbf{w}^\top x' = 0]$ for all $x, x' \in \{\pm 1\}^d$, so

$$\mathbb{E}[s_{\mathbf{w},0}(x)] = \chi(x) \cdot \mathbb{P}[\mathbf{w}^\top x = 0] = \chi(x) \cdot 2^{-d} \cdot |\{v \in \{\pm 1\}^d : v^\top x = 0\}| = \chi(x) \cdot q,$$

where $q := \binom{d}{d/2}/2^d = \Theta(1/\sqrt{d})$. Define $g(x) := \frac{1}{q2^d}\sum_{w \in \{\pm 1\}^d} s_{w,0}(x)$. Then $g(x) = \frac{1}{q}\mathbb{E}[s_{\mathbf{w},0}(x)] = \chi(x)$ for each $x \in \{\pm 1\}^d$, i.e., g interpolates the parity dataset. Finally, we bound the $\mathcal{R}$-norm:

$$\|g\|_{\mathcal{R}} \leq \frac{1}{q2^d}\sum_{w \in \{\pm 1\}^d} \|s_{w,0}\|_{\mathcal{R}} \leq \frac{1}{q} \cdot O(\sqrt{d}) \leq O(d). \qquad \blacksquare$$

While Theorem 5 successfully exhibits a neural network g that perfectly fits the parity dataset with $\|g\|_{\mathcal{R}} = O(d)$, the width of g is $\Omega(2^d)$. We next show that by allowing non-zero $L^\infty(\nu)$ error in the approximation, we can achieve a construction with both $O(d)$ $\mathcal{R}$-norm and $\mathrm{poly}(d)$ width.

Theorem 6 *There is a universal constant* $c > 0$ *such that the following holds. For any even* d*, any* $\epsilon \in (0, 1)$*, and any even* $t \in \{0, 2, \ldots, d\}$*, there exists a function* $g \colon \Omega \to \mathbb{R}$ *that can be represented by a width-*m *neural network such that* $\|g - \chi\|_{L^\infty(\nu)} \leq \epsilon$*, where*

$$\begin{aligned} &m \leq O(d^{3/2}\sqrt{\log(1/\epsilon)}/\epsilon^2) && \textit{and} && \|g\|_{\mathcal{R}} \leq O(d\log(1/\epsilon)) && \textit{if } t \leq c\sqrt{d\log(1/\epsilon)}; \\ &m \leq O(d^2/(\epsilon t)) && \textit{and} && \|g\|_{\mathcal{R}} \leq O(t\sqrt{d}) && \textit{otherwise}. \end{aligned}$$

Moreover, g *can be expressed as a linear combination of width-*t *sawtooth functions.*

Remark 7 *Suppose* ϵ *is a constant. Using* $t = \Theta(d)$*, we obtain a neural network of width* $m = O(d)$ *and* $\|g\|_{\mathcal{R}} = O(d^{3/2})$*, matching the properties of the sawtooth (ridge function) interpolant* $s_{w,d}$*. Using* $t = \Theta(1)$*, we obtain a neural network of width* $m = O(d^{3/2})$ *and* $\|g\|_{\mathcal{R}} = O(d)$*, matching the properties of the interpolant from Theorem 5 but with almost exponentially smaller width.*

A more detailed version of Theorem 6 (which also specifies the *intrinsic dimensionality* of g) is stated and proved in Appendix D.2. The proof uses a similar technique as that of Theorem 5, but instead averages randomly sampled sawtooth functions $s_{\mathbf{w}^{(1)},t}, \ldots, s_{\mathbf{w}^{(k)},t}$ for $\mathbf{w}^{(j)} \sim \mathrm{Unif}(\{\pm 1\}^d)$ of width t. We show that for sufficiently large k, every $x \in \{\pm 1\}^d$ lies the in the "active" region of about the same number of sawtooth functions; this yields a good approximation of $\chi(x)$ for all x.

4. Our results also hold for odd d, but the proofs are more tedious.

3.3. Every parity interpolant has $\mathcal{R}$-norm $\Omega(d)$

Finally, we show that $\mathcal{R}$-norm upper-bounds from Theorems 5 and 6 are tight. That is, we show that every solution to (ϵ-VP) for the parity dataset has $\mathcal{R}$-norm $\Omega(d)$, even for constant ϵ. This is implied by the following stronger result, which requires only $L^2(\nu)$ approximation, as opposed to $L^\infty(\nu)$.

Theorem 8 *For any $d \geq 8$ and $\alpha \in (0,1)$, $\inf\{\|g\|_{\mathcal{R}} : \|g - \chi\|_{L^2(\nu)} \leq 1 - \alpha\} \geq \alpha d/8$.*

The core of the proof of Theorem 8 (given in Appendix D.3) is an upper-bound on the correlation of any fixed ReLU neuron with the parity function χ.

We note that a result analogous to Theorem 8 also holds for most *sampled parity datasets* (defined in Section 4). This result is stated and proved in Appendix D.4.

4. Generalization properties of solutions to the variational problem

In this section, we consider the generalization properties of a learning algorithm that returns a solution to (VP) for a *sampled parity dataset* $\{(\mathbf{x}_i, \chi(\mathbf{x}_i)) : i \in [n]\}$ for $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim_{\text{iid}} \nu$. (Again, for simplicity, we label data using χ, but the same results hold for any χ_S with $|S| = \Theta(d)$.)

We show that $n = o(d^2/\sqrt{\log d})$ results in a predictor with nearly trivial accuracy. Note that information-theoretically, $n \geq O(d)$ is sufficient for learning any parity function (Helmbold et al., 1992; Fischer and Simon, 1992). This means that the inductive bias based on $\mathcal{R}$-norm is not sufficient to achieve statistically optimal sample complexity for learning parity functions.

4.1. Poor generalization with $n \ll d^2/\sqrt{\log d}$ samples

We first give a lower bound on the sample size needed for non-trivial generalization for learning parity functions by solving (VP) with the sampled parity dataset.

Theorem 9 *If $n = o(d^2/\sqrt{\log d})$, then with probability at least $1/2$, every solution $\mathbf{g} \colon \Omega \to \mathbb{R}$ to (VP) for the sampled parity dataset has $\|\mathbf{g} - \chi\|_{L^2(\nu)} \geq 1 - o(1)$.*

Its proof relies on the following approximation lemma, which shows the existence of a low-$\mathcal{R}$-norm network $\mathbf{g}$ that perfectly fits all n samples. The lemma (which is proved in Appendix E.1) defines $\mathbf{g}$ with the same "cap construction" used in Theorem 1 of Bubeck et al. (2021).

Lemma 10 *There is an absolute constant $c > 0$ such that the following holds. If $n \leq cd^2$, and $\mathbf{x}_1, \ldots \mathbf{x}_n \sim_{\text{iid}} \nu$, then with probability at least $1/2$, there exists $\mathbf{g} \colon \Omega \to \mathbb{R}$ with $\mathbf{g}(\mathbf{x}_i) = \chi(\mathbf{x}_i)$ for all $i \in [n]$ and $\|\mathbf{g}\|_{\mathcal{R}} \leq 4n\sqrt{\ln d}/d$.*

We conclude that generalization fails in this low-sample regime because Theorem 8 shows that no network with sufficiently small $\mathcal{R}$-norm can correlate with parity.

Proof of Theorem 9. Let $\alpha := 64n\sqrt{\ln d}/d^2$, so $\alpha = o(1)$ by assumption on n. By Theorem 8, every $g \colon \Omega \to \mathbb{R}$ with $\|g - \chi\|_{L^2(\nu)} \leq 1 - \alpha$ has $\|g\|_{\mathcal{R}} \geq \alpha d/8 \geq 8n\sqrt{\ln d}/d$. However, by Lemma 10, with probability at least $1/2$, every solution $\mathbf{g}$ to (VP) for the dataset $\{(\mathbf{x}_i, \chi(\mathbf{x}_i))\}_{i \in [n]}$ has $\|\mathbf{g}\|_{\mathcal{R}} \leq 4n\sqrt{\ln d}/d$. In this event, the solutions $\mathbf{g}$ have $\|\mathbf{g} - \chi\|_{L^2(\nu)} \geq 1 - \alpha = 1 - o(1)$. ■

4.2. Good generalization with $n \gtrsim d^3$ samples

We complement the lower-bound in Theorem 9 with the following sample complexity upper-bound.

Theorem 11 *There is an absolute constant $C > 0$ such that the following holds. For any $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$, if $n \geq C(\log(1/\delta) + d^3/\epsilon^2)$, then with probability at least $1 - \delta$, every solution $\mathbf{g}\colon \Omega \to \mathbb{R}$ to (VP) for the sampled parity dataset satisfies $\|\chi - \mathrm{clip} \circ \mathbf{g}\|^2_{L^2(\nu)} \leq \epsilon$, where $\mathrm{clip}(t) := \min\{\max\{t, -1\}, 1\}$.*

For technical reasons, we only bound the $L^2(\nu)$ error of the natural truncation of a solution to (VP). The proof in Appendix E.2 is based on standard Rademacher complexity arguments.

We note that there is a gap between our lower bound (Theorem 9) and upper bound (Theorem 11): roughly a factor of $d\sqrt{\log d}$. We believe that this gap could be narrowed if one resolves the open question raised by Bubeck et al. (2021) about the minimum Lipschitz constant achievable by two-layer ReLU networks of width m networks that interpolate a sample of size n; Lemma 10 is derived from a theorem that produces networks with smoothness conjectured to be sub-optimal. Nevertheless, our lower bound in Theorem 9 is already high enough to establish the statistical sub-optimality of solutions to (VP).

5. Generality of the averaging technique for minimizing $\mathcal{R}$-norm

In this section, we show how the benefit of averaging goes beyond the parity dataset. We consider an *f-dataset* $\{(x, f(x)\}_{x \in \{\pm 1\}^d}$, a generalization of the parity dataset where $f(x) = \phi(v^\top x)$ is a ridge function with L-Lipschitz and ρ-periodic ϕ. For another dataset generated by oscillatory ridge functions, we prove the same contrast between minimum-$\mathcal{R}$-norm interpolation with and without ridge constraints, so long as the periodicity ρ is not too small (specifically, $\rho \geq 1/\sqrt{d}$). More concretely, suppose the dataset $\{(x_i, f(x_i))\}_{i \in [n]} \subset \{\pm 1\}^d \times \{\pm 1\}$ used in (VP) and (ϵ-VP) is the *f-dataset*, where $v \in \{\pm \frac{1}{\sqrt{d}}\}^d$ and ϕ is ρ-periodic and $\frac{1}{\rho}$-Lipschitz. Then we have the following:

- The optimal value of (ϵ-VP) for constant $\epsilon \in (0, 1/2)$ is $\tilde{O}(\sqrt{d}/\rho)$. (Theorem 12)
- The optimal value of (ϵ-VP) for constant $\epsilon \in (0, 1/2)$—with the additional constraint that g be a ridge function—is $\Omega(\sqrt{d}/\rho^2)$. (Theorem 14)

Because the parity dataset is an f-dataset with a $1/\sqrt{d}$-periodic and $\sqrt{d}$-Lipschitz choice of ϕ, the above results closely match those of Informal Theorem 1. We give both results, starting with an upper bound on the minimum-$\mathcal{R}$-norm approximate interpolant, which parallels Theorem 6.

Theorem 12 *Suppose $f\colon \Omega \to [-1, 1]$ is given by $f(x) = \phi(v^\top x)$ for some unit vector $v \in \mathbb{S}^{d-1}$ and some $\phi\colon [-\sqrt{d}, \sqrt{d}] \to [-1, 1]$ that is L-Lipschitz and ρ-periodic for $\rho \in [\|v\|_\infty, 1]$. Fix any $\epsilon \in (0, 1)$. There exists a function $g\colon \Omega \to \mathbb{R}$ represented by a width-m neural network such that:*

$$\|f - g\|_{L^\infty(\nu)} \leq \epsilon; \quad m \leq dL \,\mathrm{polylog}(1/\epsilon)\sqrt{\rho\|v\|_1}/\epsilon^2; \quad \|g\|_{\mathcal{R}} \leq L^2 \,\mathrm{polylog}(d/\epsilon)\rho\|v\|_1/\epsilon.$$

Remark 13 *Suppose $f(x) = \cos(\frac{2\pi}{\rho} v^\top x)$ for $v \in \{\pm \frac{1}{\sqrt{d}}\}^d$ and $\rho \in [\frac{1}{\sqrt{d}}, 1]$. Theorem 12 implies that there exists an ϵ-approximate interpolating neural network g of width $\tilde{O}(\frac{d^{5/4}}{\sqrt{\rho}\epsilon^2})$ and $\|g\|_{\mathcal{R}} = \tilde{O}(\frac{\sqrt{d}}{\rho\epsilon})$. If d is even and $\rho = 4/\sqrt{d}$, then $f(x) = \chi(x)$ for $x \in \{\pm 1\}^d$, and the width and $\mathcal{R}$-norm bounds of Theorem 6 for small t are approximately recovered.*

A detailed version of Theorem 12 appears in Appendix F.1. The construction is more delicate than that in Theorem 6 due to the potential lack of symmetries that had existed in the parity dataset.

We give the lower bound on the $\mathcal{R}$-norm of all approximately interpolanting ridge functions, whose proof in Appendix F.2 relies a reduction to the argument of Theorem 4.

Theorem 14 *Assume d is even. Let Ridge_d be the set of functions $g\colon \Omega \to \mathbb{R}$ such that $g(x) = \phi(w^\intercal x)$ for some $w \in \mathbb{S}^{d-1}$ and Lipschitz continuous $\phi\colon [-\sqrt{d}, \sqrt{d}] \to \mathbb{R}$. Let $\rho := 4q/\sqrt{d}$ for $q \in \{1, 2, \dots, \lfloor \sqrt{d}/4 \rfloor\}$ and $f(x) := \cos((2\pi/(\rho\sqrt{d}))\mathbf{1}^\intercal x)$. Then*

$$\inf\{\|g\|_{\mathcal{R}} : g \in \mathrm{Ridge}_d,\ \|g - f\|_{L^\infty(\nu)} \le 1/2\} = \Omega(\sqrt{d}/\rho^2).$$

Remark 15 *By contrasting the above result to the $\tilde{O}(\frac{\sqrt{d}}{\rho\epsilon})$ $\mathcal{R}$-norm of the averaging-based construction from Remark 13, ridge functions are suboptimal solutions to ϵ-VP for constant ϵ.*

Remark 16 *Lemma 38 (in Appendix F.1) implies the existence of a neural network $g_{\mathrm{Ridge}} \in \mathrm{Ridge}_d$ that point-wise approximates f (i.e., $\|g_{\mathrm{Ridge}} - f\|_{L^\infty(\nu)} \le \epsilon$) and has $\|g_{\mathrm{Ridge}}\|_{\mathcal{R}} = O(\frac{\sqrt{d}}{\rho^2\epsilon})$. Hence, the lower bound in Theorem 14 is tight when ϵ is constant.*

6. Conclusion and future work

In this work, we shed light on the $\mathcal{R}$-norm inductive bias for learning neural networks, but numerous questions remain. We are particularly interested in understanding the solutions to (VP) for other datasets, as well as the generality of the averaging techniques used in our constructions. Extensions of the $\mathcal{R}$-norm to deeper networks and analyzing solutions to (VP) for other high dimensional datasets could also be useful for proving depth-separation results that focus on variational norm, complementing existing works that focus on width (Telgarsky, 2016; Eldan and Shamir, 2016; Martens et al., 2013; Daniely, 2017; Safran and Shamir, 2017; Safran et al., 2019). Finally, our work suggests that minimizing $\mathcal{R}$-norm yields neural networks that are intrinsically high-dimensional, and we are interested in whether this phenomenon is borne out in architectures beyond two-layer fully-connected networks.

Acknowledgments

This work was supported in part by NSF grants CCF-1740833 and IIS-1563785, a JP Morgan Faculty Award, and an NSF GRFP.

References

Emmanuel Abbe and Colin Sandon. Poly-time universality and limitations of deep learning. *arXiv preprint arXiv:2001.02992*, 2020.

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

Francis Bach and Lenaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. *arXiv preprint arXiv:2110.08084*, 2021.

Pierre Baldi and Peter J Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems 26*, 2013.

Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. *arXiv preprint arXiv:2207.08799*, 2022.

Peter L Bartlett. For valid generalization the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems 9*, 1996.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *arXiv preprint arXiv:2210.15651*, 2022.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In *Conference on Learning Theory*, 2021.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, 2022.

Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, 2017.

Amit Daniely and Eran Malach. Learning parities with neural networks. In *Advances in Neural Information Processing Systems 33*, 2020.

Thomas Debarre, Quentin Denoyelle, Michael Unser, and Julien Fageot. Sparsest piecewise-linear regression of one-dimensional data. *Journal of Computational and Applied Mathematics*, 406: 114044, 2022.

David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Weinan E, Chao Ma, and Lei Wu. The Barron space and the flow-induced function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory*, 2016.

Tolga Ergen and Mert Pilanci. Convex geometry and duality of over-parameterized neural networks. *Journal of Machine Learning Research*, 22(212):1–63, 2021.

Paul Fischer and Hans-Ulrich Simon. On learning ring-sum-expansions. *SIAM Journal on Computing*, 21(1):181–192, 1992.

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

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Tomer Galanti, Zachary S. Siegel, Aparna Gupte, and Tomaso Poggio. SGD and weight decay provably induce a low-rank bias in neural networks. *arXiv preprint arXiv:2206.05794*, 2022.

László Györfi, Michael Köhler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Boris Hanin. Ridgeless interpolation with shallow ReLU networks in $1d$ is nearest neighbor curvature extrapolation and provably generalizes on Lipschitz functions. *arXiv preprint arXiv:2109.12960*, 2021.

Stephen Hanson and Lorien Pratt. Comparing biases for minimal network construction with back-propagation. In *Advances in Neural Information Processing Systems 1*, 1988.

David Helmbold, Robert Sloan, and Manfred K Warmuth. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

Geoffrey E Hinton. Learning translation invariant recognition in a massively parallel networks. In *International Conference on Parallel Architectures and Languages Europe*, 1987.

Hui Jin and Guido Montúfar. Implicit bias of gradient descent for mean squared error regression with wide neural networks. *arXiv preprint arXiv:2006.07356*, 2020.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, 2008.

George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Jason M Klusowski and Andrew R Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

Michael Kohler and Adam Krzyżak. Adaptive regression estimation with multilayer feedforward neural networks. *Nonparametric Statistics*, 17(8):891–913, 2005.

Vera Kurková and Marcello Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665, 2001.

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. CRC Press, 2018.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018.

Eran Malach, Pritish Kamath, Emmanuel Abbe, and Nathan Srebro. Quantifying the benefit of using differentiable learning over tangent kernels. *arXiv preprint arXiv:2103.01210*, 2021.

James Martens, Arkadev Chattopadhya, Toni Pitassi, and Richard Zemel. On the representational efficiency of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems 26*, 2013.

Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

Hrushikesh Narhar Mhaskar. On the tractability of multivariate integration and approximation by neural networks. *Journal of Complexity*, 20(4):561–590, 2004.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR Workshop*, 2015.

Matthew Olson, Abraham Wyner, and Richard Berk. Modern neural networks generalize on small data sets. In *Advances in Neural Information Processing Systems 31*, 2018.

Patrick E. O'Neil. Hyperplane cuts of an n-cube. *Discrete Mathematics*, 1(2):193–195, 1971.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *International Conference on Learning Representations*, 2019.

Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021a.

Rahul Parhi and Robert D Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *arXiv preprint arXiv:2109.08844*, 2021b.

Saharon Rosset, Grzegorz Swirszcz, Nathan Srebro, and Ji Zhu. ℓ_1 regularization in infinite dimensional feature spaces. In *Conference on Learning Theory*, 2007.

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

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: What is actually being separated? In *Conference on Learning Theory*, 2019.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Alexander Shevchenko, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis of piecewise linear solutions for wide ReLU networks. *arXiv preprint arXiv:2111.02278*, 2021.

Jonathan W Siegel and Jinchao Xu. Characterization of the variation spaces corresponding to shallow neural networks. *arXiv preprint arXiv:2106.15002*, 2021.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on Learning Theory*, 2016.

Matus Telgarsky. Feature selection with gradient descent on two-layer networks in low-rotation regimes. *arXiv preprint arXiv:2208.02789*, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Huiyuan Wang and Wei Lin. Harmless overparametrization in two-layer neural networks. *arXiv preprint arXiv:2106.04795*, 2021.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate ReLU networks. In *Advances in Neural Information Processing Systems 32*, 2019.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems 32*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Kaiqi Zhang and Yu-Xiang Wang. Deep learning meets nonparametric regression: Are weight-decayed dnns locally adaptive? *arXiv preprint arXiv:2204.09664*, 2022.

Appendix A. Additional preliminaries

A.1. Additional definitions and notations

We say that $g : \Omega \to \mathbb{R}$ is *k-index* if there exists a matrix $U \in \mathbb{R}^{k \times d}$ and $\phi : \mathbb{R}^k \to \mathbb{R}$ such that $g(x) = \phi(Ux)$ for all $x \in \Omega$. A ridge function is 1-index.

For a matrix $M \in \mathbb{R}^{m \times n}$, we denote the i-th largest singular value of M by $\sigma_i(M)$ for $i = 1, \ldots, \min\{m, n\}$.

A random variable $\mathbf{u}$ is *c-subgaussian* if $\|\mathbf{u}\|_{\psi_2} := \inf\{t \geq 0 : \mathbb{E}[\exp(\mathbf{u}^2/t^2)] \leq 2\} \leq c$, and a random vector $\mathbf{v}$ is σ^2*-subgaussian* if every one-dimensional projection of $\mathbf{v}$ is c-subgaussian.

The *bias-corrected network* $\bar{g}_\mu$ obtained from the (infinite-width) neural network g_μ is given by $\bar{g}_\mu(x) := g_\mu(x) - g_\mu(0)$; equivalently, $\bar{g}_\mu(x) = \int (\varphi_{\mathrm{r}}(w^\intercal x + b) - \varphi_{\mathrm{r}}(b))\, \mu(\mathrm{d}w, \mathrm{d}b)$.

The asymptotics implied in the Landau notation (big-O, big-Ω, etc.) regard all quantities as potentially increasing functions (e.g., t) or decreasing functions (e.g., ϵ, δ, α, ρ) of the dimension d. The soft-O notation $\tilde{O}(\cdot)$ (only used informally) suppresses terms that are poly-logarithmic in those that appear. Some of our theorems and lemmas contain an "if clause" that uses Landau notation, such as "if $n \geq O(d^2)$, [...]". The interpretation of such a clause is: "there exists $n_0(d) \in O(d^2)$ such that if $n \geq n_0(d)$, [...]". (And, of course, an analogous interpretation should be used when "$O(d^2)$" is replaced by other expressions using Landau notation.)

A.2. Concentration inequalities

Our proofs make extensive use of textbook probability concentration inequalities. We provide those results below.

Lemma 17 (Hoeffding's inequality; Theorem 2.8 in Boucheron et al., 2013) *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be independent, mean-zero random variables such that $\mathbf{u}_i$ takes value in $[a_i, b_i]$ almost surely for all $i \in [n]$. Then, for any $t > 0$,*

$$\mathbb{P}\left[\sum_{i=1}^n \mathbf{u}_i \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 18 (Multiplicative Chernoff bound; Theorem 4.4 in Mitzenmacher and Upfal, 2017) *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be independent Bernoulli random variables with $\mathbb{P}[\mathbf{u}_i = 1] = p \in [0, 1]$ for all $i \in [n]$. Then, for any $\eta \in (0, 1]$,*

$$\mathbb{P}\left[\sum_{i=1}^n \mathbf{u}_i \geq (1 + \eta)p\right] \leq \exp\left(-\frac{p\eta^2}{3}\right).$$

Lemma 19 (Bernstein's inequality; Corollary 2.11 in Boucheron et al., 2013) *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be independent, mean-zero random variables with $\mathbf{u}_i \leq K$ almost surely for all $i \in [n]$, and let $v := \sum_{i=1}^n \mathbb{E}[\mathbf{u}_i^2]$. Then, for any $t > 0$,*

$$\mathbb{P}\left[\sum_{i=1}^n \mathbf{u}_i \geq t\right] \leq \exp\left(-\frac{t^2}{2(v + Kt/3)}\right).$$

Lemma 20 (McDiarmid's inequality; Theorem 6.2 in Boucheron et al., 2013) *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be independent random variables, and let f be a measurable function. Suppose, for each $i \in [n]$, the value of $f(u_1, \ldots, u_n)$ can change by at most $c_i \geq 0$ by changing the value of u_i. Then, for any $t > 0$,*

$$\mathbb{P}\left[f(\mathbf{u}_1, \ldots, \mathbf{u}_n) - \mathbb{E}\left[f(\mathbf{u}_1, \ldots, \mathbf{u}_n)\right] \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma 21 (Properties of subgaussian random variables) *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be independent random variables with $\|\mathbf{u}_i\|_{\psi_2} < \infty$ for all $i \in [n]$. There is a universal constant $C > 0$ such that the following hold.*

(i) (Concentration; Section 2.5.2 in Vershynin, 2018) For any $t > 0$, $\mathbb{P}\left[|\mathbf{u}_1| \geq t\right] \leq 2\exp(-t^2/(C\|\mathbf{u}_1\|_{\psi_2}))$.

(ii) (Maximum; Exercise 2.5.10 in Vershynin, 2018) $\mathbb{E}\left[\max_{i\in[n]} |\mathbf{u}_i|\right] \leq C\sqrt{\ln n}\max_{i\in[n]}\|\mathbf{u}_i\|_{\psi_2}$.

(iii) (Averaging: Proposition 2.6.1 in Vershynin, 2018) $\|\sum_{i=1}^n \mathbf{u}_i\|_{\psi_2}^2 \leq C\sum_{i=1}^n \|\mathbf{u}_i\|_{\psi_2}^2$.

(iv) (Centering; Lemma 2.6.8 in Vershynin, 2018) $\|\mathbf{u}_1 - \mathbb{E}\left[\mathbf{u}_1\right]\|_{\psi_2} \leq C\|\mathbf{u}_1\|_{\psi_2}$.

(v) (Lipschitzness) For any 1-Lipschitz function $\phi \colon \mathbb{R} \to \mathbb{R}$ with $\phi(0) = 0$, $\|\phi(\mathbf{u}_1)\|_{\psi_2} \leq \|\mathbf{u}_1\|_{\psi_2}$.

Proof. The only property not already proved in (Vershynin, 2018) is (v). Since ϕ is 1-Lipschitz and $\phi(0) = 0$,

$$|\phi(\mathbf{u}_1)| = |\phi(\mathbf{u}_1) - \phi(0)| \leq |\mathbf{u}_1 - 0| = |\mathbf{u}_1|.$$

Hence

$$\mathbb{E}\left[\exp(\phi(\mathbf{u}_1)^2/\|\mathbf{u}_1\|_{\psi_2}^2)\right] \leq \mathbb{E}\left[\exp(\mathbf{u}_1^2/\|\mathbf{u}_1\|_{\psi_2}^2)\right] \leq 2,$$

which implies $\|\phi(\mathbf{u}_1)\|_{\psi_2} \leq \|\mathbf{u}_1\|_{\psi_2}$. ■

Lemma 22 (Singular values of random matrices; Theorem 4.6.1 in Vershynin, 2018) *There is a positive constant $C > 0$ such that the following holds. Let $\mathbf{A}$ be a $m \times n$ random matrix whose rows are independent, mean-zero, v-subgaussian random vectors in $\mathbb{R}^n$. For any $t \geq 0$, with probability at least $1 - 2e^{-t}$, the singular values $\sigma_1(\mathbf{A}), \sigma_2(\mathbf{A}), \ldots, \sigma_n(\mathbf{A})$ of $\mathbf{A}$ satisfy*

$$\sqrt{m} - Cv(\sqrt{n} + \sqrt{t}) \leq \sigma_i(\mathbf{A}) \leq \sqrt{m} + Cv(\sqrt{n} + \sqrt{t}) \quad \text{for all } i \in [n].$$

Lemma 23 (Moment generating function for $\mathrm{Unif}(\{\pm 1\})$; Lemma 2.2 in Boucheron et al., 2013) *If $\mathbf{u} \sim \mathrm{Unif}(\{\pm 1\})$, then $\mathbb{E}[\exp(t\mathbf{u})] \leq \exp(t^2/2)$ for all $t \in \mathbb{R}$.*

A.3. Covering numbers

Let $\mathcal{N}(\epsilon, A, \gamma)$ denote the *covering number* over a metric space A with metric $\gamma : A \times A \to \mathbb{R}_+$. That is, $\mathcal{N}(\epsilon, A, d) = \min|\mathcal{N}_\epsilon|$ for $\mathcal{N}_\epsilon \subset A$ such that for all $x \in A$, there exists $x' \in \mathcal{N}_\epsilon$ with $\gamma(x, x') \leq \epsilon$. Note that $\mathcal{N}(\epsilon, [-1, 1], |\cdot|) \leq \frac{2}{\epsilon}$.

Lemma 24 (Covering numbers of $\mathbb{S}^{d-1}$; Corollary 4.2.13 in Vershynin, 2018) *For any $\epsilon > 0$, $\mathcal{N}(\epsilon, \mathbb{S}^{d-1}, \|\cdot\|_2) \leq (\frac{2}{\epsilon} + 1)^d$. If $\epsilon \in [0, 1]$, then $\mathcal{N}(\epsilon, \mathbb{S}^{d-1}, \|\cdot\|_2) \leq (\frac{3}{\epsilon})^d$.*

Appendix B. Additional properties of $\mathcal{R}$-norm

B.1. Existence and sparsity of the solution to VP

Proposition 25 (Lemma 2 in Parhi and Nowak, 2021b) *For any $g : \Omega \to \mathbb{R}$ with $\|g\|_{\mathcal{R}} < \infty$, there exists an even Radon measure*[5] *μ over $\mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}]$, and $v \in \mathbb{R}^d, c \in \mathbb{R}$, such that g admits an integral of the form*

$$g(x) = \int_{\mathbb{S}^{d-1}\times[-\sqrt{d}\times\sqrt{d}]} \varphi_{\mathrm{r}}(w^\intercal x + b)\,\mu(\mathrm{d}w, \mathrm{d}b) + v^\intercal x + c \quad \forall x \in \Omega.$$

Moreover, μ attains the ($\mathcal{R}$-norm), i.e., $\|g\|_{\mathcal{R}} = |\mu|$.

The following theorem of Parhi and Nowak (2021b) formalizes the fact that the $\mathcal{R}$-norm-minimizing interpolant of a d-dimensional dataset can be represented as a finite-width neural network.

Theorem 26 *For any dataset $\{(x_i, y_i)\}_{i\in[n]}$ from $\Omega \times \mathbb{R}$, the infimum in (VP) is achieved by the sum of an affine function $x \mapsto v^\intercal x + c$ and a finite-width neural network g of the form*

$$g(x) = \sum_{j=1}^{m} a^{(j)}\varphi_{\mathrm{r}}(w^{(j)}x + b^{(j)}) + v^\intercal x + c,$$

with $m \le \max\{0, n-(d+1)\}$ and $(w_j, b_j) \in \mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}]$ for all $i \in [m]$.

Proof. By Theorem 5 of Parhi and Nowak (2021b) (see also the proof of Theorem 1 of Parhi and Nowak (2021a) which covers the interpolation form of the optimization problem), there exists a neural network $x \mapsto \sum_{j=1}^{m'} a^{(j)}\varphi_{\mathrm{r}}(w^{(j)\intercal}x + b^{(j)})$ of width $m' \le n-(d+1)$, and an affine function $x \mapsto v^{(0)\intercal}x + c^{(0)}$, such that their sum achieves the infimum in (VP). We can divide neurons of the neural network into two sets based on whether their corresponding bias term is smaller or larger than $\sqrt{d}$ in absolute value. Since every $x \in \Omega$ satisfies $\|x\|_2 \le \sqrt{d}$, without loss of generality (by possibly flipping the sign of some $a^{(j)}$ and $w^{(j)}$), assume the first m neurons satisfy $\left|b^{(j)}\right| \le \sqrt{d}$ and the rest satisfy $b^{(j)} > \sqrt{d}$. Then we have

$$\begin{aligned} g(x) &= \sum_{j=1}^{m} a^{(j)}\varphi_{\mathrm{r}}(w^{(j)\intercal}x + b^{(j)}) + \sum_{j=m+1}^{m'} a^{(j)}\varphi_{\mathrm{r}}(w^{(j)\intercal}x + b^{(j)}) + v^{(0)\intercal}x + c^{(0)} \\ &= \sum_{j=1}^{m} a^{(j)}\varphi_{\mathrm{r}}(w^{(j)\intercal}x + b^{(j)}) + \sum_{j=m+1}^{m'} a^{(j)}(w^{(j)\intercal}x + b^{(j)}) + v^{(0)\intercal}x + c^{(0)} \\ &= \sum_{j=1}^{m} a^{(j)}\varphi_{\mathrm{r}}(w^{(j)\intercal}x + b^{(j)}) + \underbrace{\left(v^{(0)} + \sum_{j=m+1}^{m'} a^{(j)}w^{(j)}\right)^\intercal}_{=:v} x + \underbrace{\sum_{j=m+1}^{m'} b^{(j)} + c^{(0)}}_{=:c}. \end{aligned}$$

Therefore, g has the desired form with $m \le m' \le n-(d+1)$. ■

5. Evenness of μ should interpreted in the distributional sense, but it roughly means $\mu(w, b) = \mu(-w, -b)$ when μ has a density.

Appendix C. Extending our results to a different variational norm

This paper considers the approximation and generalization implications of bounding the complexity of shallow neural networks with the $\mathcal{R}$-norm. However, $\mathcal{R}$-norm is not the only weight-based complexity measurement, and other works employ slightly different norms for similar purposes. This appendix demonstrates that our results are not peculiarities of our formulation of $\mathcal{R}$-norm and extend to other variational norms. One alternative—which we refer to as the $\mathcal{V}_2$-*norm*—omits the linear component of the neural network whose measure determines the $\mathcal{R}$-norm and instead permits ReLU neurons whose thresholds lie outside the domain Ω.

We first introduce notation for an infinite-width neural network that permits such thresholds. Let $\mathcal{M}'$ denote the space of probability measures over $\mathbb{S}^{d-1} \times [-2\sqrt{d}, 2\sqrt{d}]$. For some measure $\tilde{\mu} \in \mathcal{M}'$, let $\tilde{g}_\mu : \Omega \to \mathbb{R}$ be an infinite-width neural network with

$$\tilde{g}_{\tilde{\mu}}(x) = \int_{\mathbb{S}^{d-1} \times [-2\sqrt{d}, 2\sqrt{d}]} \varphi_{\mathrm{r}}(w^\intercal x + b)\, \tilde{\mu}(\mathrm{d}w, \mathrm{d}b),$$

which has total variation norm $|\tilde{\mu}| = \int_{\mathbb{S}^{d-1} \times [-2\sqrt{d}, 2\sqrt{d}]} |\tilde{\mu}|\,(\mathrm{d}w, \mathrm{d}b)$. Now, we introduce the $\mathcal{V}_2$-norm for some $g : \Omega \to \mathbb{R}$:

$$\|g\|_{\mathscr{V}_2} = \inf_{\tilde{\mu} \in \mathcal{M}'} |\tilde{\mu}| \quad \text{s.t.} \quad g(x) = \tilde{g}_{\tilde{\mu}}(x), \quad \forall x \in \Omega. \tag{$\mathcal{V}_2$-norm}$$

In the same spirit as Lemma 1, for a discrete network $g(x) = g_\theta(x)$ with $\theta = (a^{(j)}, w^{(j)}, b^{(j)})_{j \in [m]} \in (\mathbb{R} \times \mathbb{S}^{d-1} \times [-2\sqrt{d}, 2\sqrt{d}])^m$, we have $\|g\|_{\mathscr{V}_2} \leq \|a\|_1$.

Our definition of the $\mathcal{V}_2$-norm was introduced by Siegel and Xu (2021) as the norm corresponding to their variation space $\mathbb{P}_1$ with constants $c_1 = -2\sqrt{d}$ and $c_2 = 2\sqrt{d}$. They relate the $\mathcal{V}_2$-norm to the *Barron norm* of E et al. (2019) and the Radon norm of Ongie et al. (2019). We show that the $\mathcal{V}_2$-norm and the $\mathcal{R}$-norm are closely related and that all of our bounds apply equivalently to the $\mathcal{V}_2$-norm. We first place upper and lower bounds on the $\|g\|_{\mathscr{V}_2}$ in terms of $\|g\|_{\mathcal{R}}$ and then explain why each of our results transfers to this new variational norm.

Theorem 27 *Suppose $g : \Omega \to \mathbb{R}$ has $\|g\|_{\mathcal{R}} < \infty$. Then, $\|g\|_{\mathcal{R}} \leq \|g\|_{\mathscr{V}_2}$. If g is bounded near the origin (i.e., $|g(x)| \leq K$ for all x with $\|x\|_2 \leq 1$), then $\|g\|_{\mathscr{V}_2} \leq 12\,\|g\|_{\mathcal{R}} + 18K$.*

As a result, all of our results that apply to the $\mathcal{R}$-norm translate modulo constants to the $\mathcal{V}_2$-norm. Because $\|g\|_{\mathscr{V}_2} \geq \|g\|_{\mathcal{R}}$ always holds, every theorem that places lower bounds on an $\mathcal{R}$-norm exactly translates to $\|g\|_{\mathscr{V}_2}$, including Theorems 4, 8, 32, and 11. The upper-bounds hold up to constants by observing that every target function we consider is bounded by some K on Ω.

- Because every sawtooth $s_{w,t}$ is bounded by 1, the averages of sawtooths g in Theorems 5 and 6 are bounded by $K = \frac{1}{q} = O(\sqrt{d})$. Hence, $\|g\|_{\mathscr{V}_2} = O(d)$, just like $\|g\|_{\mathcal{R}}$.
- For the "cap" construction $\mathbf{g}$ of Theorem 9, there are $k = O(n \log(d)/d)$ neurons, none of which are active at the origin. Their biases are negative and—under the "good event"—their weight norms are $O(1/\log d)$. Thus, no neuron can output a value greater than $O(1/\log d)$, so even if all k neurons activate, every x with $\|x\|_2 \leq 1$ has $|\mathbf{g}(x)| = O(n/d)$, which is dominated by the $\mathcal{R}$-norm of $O(n\sqrt{\log d}/d)$.

- The construction g of Theorem 12 computes an average of functions bounded on $[-1,1]$. Therefore, g is bounded by 1, and its $\mathcal{V}_2$-norm is no more than its $\mathcal{R}$-norm.

Proof of Theorem 27. We show separately that $\|g\|_{\mathscr{V}_2} \geq \|g\|_{\mathcal{R}}$, and then that $\|g\|_{\mathscr{V}_2} \leq 12\,\|g\|_{\mathcal{R}} + 18K$ under the additional hypothesis that $|g(x)| \leq K$ for all $x \in \Omega$ such that $\|x\|_2 \leq 1$.

Lower bound on $\mathcal{V}_2$-norm: Fix any $\xi > 0$. By the definition of $\mathcal{V}_2$-norm, there exists $\tilde{\mu} \in \mathcal{M}'$ such that $g(x) = \tilde{g}_{\tilde{\mu}}(x)$ for all $x \in \Omega$ and $|\tilde{\mu}| \leq \|g\|_{\mathscr{V}_2} + \xi$.[6] We show that there exists g_μ (where μ is $\tilde{\mu}$ with the support of b restricted to $[-\sqrt{d}, \sqrt{d}]$), v, and c such that $\tilde{g}_{\tilde{\mu}}(x) = g_\mu(x) + v^\intercal x + c$ for all $x \in \Omega$. Observe that for any $x \in \Omega$, $w^\intercal x + b > 0$ if $b > \sqrt{d}$ and $w^\intercal x + b < 0$ if $b < -\sqrt{d}$.

$$
\begin{aligned}
\tilde{g}_{\tilde{\mu}}(x) &= \int_{\mathbb{S}^{d-1}\times[-2\sqrt{d},-\sqrt{d}]} \varphi_{\mathrm{r}}(w^\intercal x + b)\tilde{\mu}(\mathrm{d}w,\mathrm{d}b) + \int_{\mathbb{S}^{d-1}\times[-\sqrt{d},\sqrt{d}]} \varphi_{\mathrm{r}}(w^\intercal x + b)\tilde{\mu}(\mathrm{d}w,\mathrm{d}b) \\
&\quad + \int_{\mathbb{S}^{d-1}\times[\sqrt{d},2\sqrt{d}]} \varphi_{\mathrm{r}}(w^\intercal x + b)\tilde{\mu}(\mathrm{d}w,\mathrm{d}b) \\
&= 0 + \int_{\mathbb{S}^{d-1}\times[-\sqrt{d},\sqrt{d}]} \varphi_{\mathrm{r}}(w^\intercal x + b)\mu(\mathrm{d}w,\mathrm{d}b) + \int_{\mathbb{S}^{d-1}\times[\sqrt{d},2\sqrt{d}]} (w^\intercal x + b)\tilde{\mu}(\mathrm{d}w,\mathrm{d}b) \\
&= g_\mu(x) + \sum_{i=1}^{d} x_i \underbrace{\int_{\mathbb{S}^{d-1}\times[\sqrt{d},2\sqrt{d}]} w_i \tilde{\mu}(\mathrm{d}w,\mathrm{d}b)}_{:=v_i} + \underbrace{\int_{\mathbb{S}^{d-1}\times[\sqrt{d},2\sqrt{d}]} b\tilde{\mu}(\mathrm{d}w,\mathrm{d}b)}_{:=c} \\
&= g_\mu(x) + v^\intercal x + c.
\end{aligned}
$$

As a result, $\|g\|_{\mathcal{R}} \leq |\mu| \leq |\tilde{\mu}| \leq \|g\|_{\mathscr{V}_2} + \xi$. Because the argument holds simultaneously for all $\xi > 0$, we conclude that $\|g\|_{\mathcal{R}} \leq \|g\|_{\mathscr{V}_2}$.

Upper bound on $\mathcal{V}_2$-norm: By Proposition 25, there exist $\mu \in \mathcal{M}'$, $v \in \mathbb{R}^d$, and $c \in \mathbb{R}$ such that $g(x) = g_\mu(x) + v^\intercal x + c$ for all $x \in \Omega$ and $|\mu| = \|g\|_{\mathcal{R}}$. We construct $\tilde{\mu} \in \mathcal{M}'$ such that $g_\mu(x) + v^\intercal x + c = \tilde{g}_{\tilde{\mu}}(x)$ for all $x \in \Omega$:[7]

$$
\tilde{\mu}(w,b) = \begin{cases} \mu(w,b) & \text{if } b \in [-\sqrt{d}, \sqrt{d}]; \\ \left(-3\,\|v\|_2 + \frac{2c}{\sqrt{d}}\right)\delta\left((w,b) - \left(v, 2\sqrt{d}\right)\right) & \\ \quad + \left(4\,\|v\|_2 - \frac{2c}{\sqrt{d}}\right)\delta\left((w,b) - \left(v, \frac{3}{2}\sqrt{d}\right)\right) & \text{otherwise.} \end{cases}
$$

Fix any $x \in \Omega$. Then:

$$
\begin{aligned}
\tilde{g}_{\tilde{\mu}}(x) - g_\mu(x) &= \left(-3\,\|v\|_2 + \frac{2c}{\sqrt{d}}\right)\varphi_{\mathrm{r}}\left(\frac{v^\intercal}{\|v\|_2}x + 2\sqrt{d}\right) \\
&\quad + \left(4\,\|v\|_2 - \frac{2c}{\sqrt{d}}\right)\varphi_{\mathrm{r}}\left(\frac{v^\intercal}{\|v\|_2}x + \frac{3}{2}\sqrt{d}\right) \\
&= \left(-3\,\|v\|_2 + \frac{2c}{\sqrt{d}}\right)\left(\frac{v^\intercal}{\|v\|_2}x + 2\sqrt{d}\right) + \left(4\,\|v\|_2 - \frac{2c}{\sqrt{d}}\right)\left(\frac{v^\intercal}{\|v\|_2}x + \frac{3}{2}\sqrt{d}\right) \\
&= v^\intercal x + c.
\end{aligned}
$$

6. This relies on the assumption that $\|g\|_{\mathscr{V}_2} < \infty$, but if it is not, then the claim trivially follows because $\|g\|_{\mathcal{R}} < \infty$.
7. In the event that $v = \mathbf{0}$, we use $\frac{v}{\|v\|_2} := \frac{1}{\sqrt{d}}\mathbf{1} \in \mathbb{S}^{d-1}$.

Therefore, $|\tilde{\mu}| \leq |\mu| + |-3\,\|v\|_2 + \frac{2c}{\sqrt{d}}| + |4\,\|v\|_2 - \frac{2c}{\sqrt{d}}| \leq |\mu| + 7\|v\|_2 + \frac{4|c|}{\sqrt{d}}$. It suffices to bound $\|v\|_2$ and $|c|$.

- Let $x_0 := \frac{v}{\|v\|_2}$. By boundedness, the triangle inequality, and several applications of Holder's inequality:

$$\begin{aligned}
|g(x_0) - g(0)| &\geq |v^\mathsf{T} x_0| - |g_\mu(x_0) - g_\mu(0)| \\
&= \|v\|_2 - \left| \int_{\mathbb{S}^{d-1}\times[-\sqrt{d},\sqrt{d}]} (\varphi_\mathrm{r}(w^\mathsf{T} x_0 + b) - \varphi_\mathrm{r}(b))\mu(\mathrm{d}w, \mathrm{d}b) \right| \\
&\geq \|v\|_2 - \int_{\mathbb{S}^{d-1}\times[-\sqrt{d},\sqrt{d}]} |\varphi_\mathrm{r}(w^\mathsf{T} x_0 + b) - \varphi_\mathrm{r}(b)|\,|\mu|\,(\mathrm{d}w, \mathrm{d}b) \\
&\geq \|v\|_2 - \|x_0\|_2\,|\mu| = \|v\|_2 - |\mu|\,.
\end{aligned}$$

  Hence, $\|v\|_2 \leq |g(x_0) - g(0)| + |\mu| \leq 2K + |\mu|$.

- We similarly employ our bound on $g(0)$:

$$K \geq |g(0)| \geq |c| - \left| \int_{\mathbb{S}^{d-1}\times[-\sqrt{d},\sqrt{d}]} \varphi_\mathrm{r}(b)\mu(\mathrm{d}w, \mathrm{d}b) \right| \geq |c| - |\mu|\,\sqrt{d}.$$

  As a result, $|c| \leq K + |\mu|\,\sqrt{d}$.

Therefore, $\|g\|_{\mathscr{V}_2} \leq |\tilde{\mu}| \leq 12\,|\mu| + 18K \leq 12\,\|g\|_{\mathcal{R}} + 18K$. ■

Appendix D. Proofs for Section 3

D.1. Proofs for Section 3.1

The proof of Theorem 4 relies on the following lemma, which is essentially a robust version of the mean value theorem.

Lemma 28 *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous on the interval $[t_1, t_2]$ with $\phi(t_1) \neq \phi(t_2)$. Define*

$$A := \Big\{ t \in [t_1, t_2] : \phi'(t) \text{ exists}, \\ \left|\phi'(t)\right| \geq \frac{1}{2} \cdot \frac{|\phi(t_2) - \phi(t_1)|}{t_2 - t_1},\ \operatorname{sign}\left(\phi'(t)\right) = \operatorname{sign}\left(\phi(t_2) - \phi(t_1)\right) \Big\}.$$

(The factor $1/2$ in the definition of A can be replaced by any constant in $(0, 1)$.) Then, $\mathrm{Leb}(A) > 0$, where Leb is the Lebesgue measure.

Proof. Recall that ϕ' denotes the right continuous derivative of g (or the right-hand Dini derivative) which is guaranteed to exist except on a null set by Rademacher's theorem. Let $s := \operatorname{sign}\left(\phi(t_2) - \phi(t_1)\right)$. By the assumption $\phi(t_1) \neq \phi(t_2)$ and the Fundamental Theorem of Calculus, we have

$$0 < |\phi(t_2) - \phi(t_1)| = s\left(\phi(t_2) - \phi(t_1)\right) = \int_{t_2}^{t_1} s\phi'(z)\,\mathrm{d}z \leq (t_2 - t_1) \operatorname*{ess\,sup}_{z\in[t_1,t_2]} s\phi'(z).$$

Recall that, by definition,

$$\operatorname*{ess\,sup}_{z\in[t_1,t_2]} s\phi'(z) = \inf\left\{a : \mathrm{Leb}(\{z \in [t_1,t_2] : \phi'(z) \text{ exists},\ s\phi'(z) > a\}) = 0\right\},$$

and thus

$$B := \left\{z \in [t_1,t_2] : \phi'(z) \text{ exists},\ s(\phi(t_2)-\phi(t_1)) \le 2\cdot(t_2-t_1)s\phi'(z)\right\}$$

satisfies $\mathrm{Leb}(B) > 0$. Observe that $B = A$, so $\mathrm{Leb}(A) > 0$, concluding the proof. ∎

D.2. Proofs for Section 3.2

Theorem 29 (Detailed version of Theorem 6) *There exists a universal constant C such that for any even d, even sawtooth width $t \in \{0, 2, \dots d\}$, and accuracy $\epsilon \in (0, \frac{1}{2})$, there exists a k-index (see Appendix A.1) width-m neural network g with $\|g-\chi\|_{L^\infty(\nu)} \le \epsilon$ such that:*

1. *If $t \le C\sqrt{d \ln \frac{1}{\epsilon}}$, then $k = O(\frac{d^{3/2}}{t+1}\cdot\frac{\log^{1/2}\epsilon}{\epsilon^2})$, $m = O(d^{3/2}\frac{\log^{1/2}\epsilon}{\epsilon^2})$, and $\|g\|_{\mathcal{R}} = O(d\log\frac{1}{\epsilon})$.*
2. *Otherwise, $k = O(\frac{d^2}{\epsilon t^2})$, $m = O(\frac{d^2}{\epsilon t})$, and $\|g\|_{\mathcal{R}} = O(t\sqrt{d})$.*

Proof. For $\mathbf{w}^{(1)},\dots,\mathbf{w}^{(k)} \sim_{\mathrm{iid}} \mathrm{Unif}(\{\pm1\}^d)$ and $q := \mathbb{P}\left[\left|\mathbf{w}^{(1)\intercal}x\right| \le t\right]$ (for any $x \in \{\pm1\}^d$), let

$$\mathbf{g}(x) := \frac{1}{kq}\sum_{j=1}^{k} s_{\mathbf{w}^{(j)},t}(x).$$

Because $\mathbb{E}\left[s_{\mathbf{w},t}(x)\right] = q\cdot\chi(x)$, we have $\mathbb{E}\left[\mathbf{g}(x)\right] = \chi(x)$. Following the arguments in the proof of Theorem 5, we have $\|\mathbf{g}\|_{\mathcal{R}} = O(\frac{(t+1)\sqrt{d}}{q})$, and $\mathbf{g}$ has width $O(k(t+1))$.

It remains to place a lower bound on q and to show that with non-zero probability, $\mathbf{g}$ uniformly approximates χ. By applying a union bound, it suffices to show that $\mathbb{P}\left[|\mathbf{g}(x)-\chi(x)| \ge \epsilon\right] \le \frac{1}{2^{d+1}}$ for any fixed $x \in \{\pm1\}^d$.

For fixed x, let $\mathbf{r}(x) := |\{j \in [k] : |x^\intercal\mathbf{w}^{(j)}| > t\}|$. We upper-bound the accuracy of approximation of $\chi(x)$ by $\mathbf{g}(x)$ in terms of $\mathbf{r}(x)$:

$$\begin{aligned}|\mathbf{g}(x)-\chi(x)| &= \left|\frac{1}{qk}\sum_{j=1}^{k}\mathbb{1}\{|x^\intercal\mathbf{w}^{(j)}| \le t\} - q\right| \\ &= \left|\frac{(k-\mathbf{r}(x))(1-q)}{qk} - \frac{\mathbf{r}(x)q}{qk}\right| = \left|\frac{1-q}{q} - \frac{\mathbf{r}(x)}{qk}\right|.\end{aligned}$$

Define $T := 2\lceil\sqrt{(d/2)\ln(8/\epsilon)}\rceil$, and note that $\mathbb{P}\left[|\mathbf{w}^\intercal x| \ge T\right] \le \frac{\epsilon}{4}$. The proof naturally divides into two cases, depending on the value of t.

Case 1: $t \le T$. We first lower-bound q. Because $\mathbf{w}^\intercal x$ is a shifted symmetric binomial distribution around $\mathbf{w}^\intercal x = 0$, if $|t'| \ge |t|$ and $t' \equiv t \pmod 2$, then $\mathbb{P}\left[\mathbf{w}^\intercal x = t'\right] \le \mathbb{P}\left[\mathbf{w}^\intercal x = t\right]$. Then, for any $t \le T$:

$$q = \sum_{\tau=-t/2}^{t/2} \mathbb{P}\left[\mathbf{w}^\intercal x = 2\tau\right] = (t+1) \cdot \frac{1}{t+1} \sum_{\tau=-t/2}^{t/2} \mathbb{P}\left[\mathbf{w}^\intercal x = 2\tau\right]$$
$$\ge (t+1) \cdot \frac{1}{T+1} \sum_{\tau=-T/2}^{T/2} \mathbb{P}\left[\mathbf{w}^\intercal x = 2\tau\right] = \frac{t+1}{T+1} \mathbb{P}\left[|\mathbf{w}^\intercal x| \le T\right]$$
$$\ge \frac{(t+1)(1-\frac{\epsilon}{2})}{2\sqrt{d \ln \frac{4}{\epsilon}}} \ge \frac{t+1}{4\sqrt{d \ln \frac{4}{\epsilon}}}.$$

Now, we bound $\mathbf{r}(x)$ by Bernstein's inequality (Lemma 19) by taking $k \ge \frac{Cd^{3/2}\sqrt{\ln \frac{1}{\epsilon}}}{\epsilon^2 (t+1)}$:

$$\mathbb{P}\left[|\mathbf{g}(x) - \chi(x)| > \epsilon\right] = \mathbb{P}\left[|\mathbf{r}(x) - \mathbb{E}[\mathbf{r}(x)]| > \epsilon q k\right]$$
$$\le 2\exp\left(-\frac{\epsilon^2 q^2 k^2}{2(kq(1-q) + \epsilon q k/3)}\right)$$
$$\le 2\exp\left(-\frac{\epsilon^2 k (t+1)}{8(1+\epsilon/3)\sqrt{d \ln \frac{4}{\epsilon}}}\right) \le \frac{1}{2^{d+1}}.$$

Case 2: $t \ge T$**.** By Hoeffding's inequality (Lemma 17) and the assumption on t, we have $q \ge 1 - 2\exp(-\frac{2t^2}{d}) \ge 1 - \epsilon/4 \ge 3/4$. Observe that $\mathbb{E}\left[\mathbf{r}(x)\right] = (1-q)k \le \frac{\epsilon k}{4}$.

We show that $|\mathbf{g}(x) - \chi(x)| = |\frac{1-q}{q} - \frac{\mathbf{r}(x)}{qk}| \le \epsilon$ by showing that $\mathbf{r}(x) \le (1-q)k + \epsilon q k$ and $\mathbf{r}(x) \ge (1-q)k - \epsilon q k$. Because $1 - q \le \frac{\epsilon}{4}$ and $q \ge \frac{3}{4}$, $(1-q)k - \epsilon q k \le -\frac{\epsilon}{2}k$, so the second inequality is always satisfied because $\mathbf{r}(x) \ge 0$. For the former inequality, it suffices to show that $\mathbf{r}(x) \le \frac{3\epsilon k}{4}$ with probability at least $1 - 2^{-(d+1)}$. We take $k \ge \frac{Cd^2}{\epsilon t^2}$, which implies $k \ge C(\frac{d}{2\epsilon} + \frac{de^{-2t^2/d}}{2\epsilon^2}) \ge C(\frac{d}{2\epsilon} + \frac{d(1-q)}{4\epsilon^2})$ by the bounds on t and q. Then, by Bernstein's inequality (Lemma 19), we have

$$\mathbb{P}\left[|\mathbf{g}(x) - \chi(x)| > \epsilon\right] = \mathbb{P}\left[|\mathbf{r}(x) - \mathbb{E}[\mathbf{r}(x)]| > \frac{3\epsilon k}{4}\right]$$
$$\le 2\exp\left(-\frac{9\epsilon^2 k^2/16}{2(kq(1-q) + \epsilon k/4)}\right) \le \frac{1}{2^{d+1}},$$

so the claim follows. ■

D.3. Proofs for Section 3.3

Theorem 8 *For any $d \ge 8$ and $\alpha \in (0,1)$, $\inf\{\|g\|_{\mathcal{R}} : \|g - \chi\|_{L^2(\nu)} \le 1 - \alpha\} \ge \alpha d/8$.*

Proof. Consider any measure μ over $\mathbb{S}^{d-1} \times [-2\sqrt{d}, 2\sqrt{d}]$, $v \in \mathbb{R}^d$, and $c \in \mathbb{R}$ such that $g(x) = g_\mu(x) + v^\intercal x + c = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \varphi_{\mathrm{r}}(w^\intercal x + b)\mu(\mathrm{d}w, \mathrm{d}b)$ for all $\|x\|_2 \le d$. We prove the claim by showing that $|\mu| \ge \frac{\alpha d}{8}$ for any such μ.

By Fact 1, $\|g-\chi\|_{L^2(\nu)} \leq 1-\alpha$ implies that $\langle g, \chi\rangle \geq \alpha$. We show that this inner product bound is only possible if $|\mu|$ is sufficiently large. By Lemma 30, any fixed neuron $r_{w,b}(x) := \varphi_{\mathrm{r}}(w^\mathsf{T} x + b)$ has $|\langle r_{w,b}, \chi\rangle| \leq \frac{8}{d}$. Because the inner-product over $\{\pm 1\}^d$ is a discrete sum and χ is orthogonal to any affine function (such as $x \mapsto v^\mathsf{T} x + c$), we can upper-bound the ability of g to correlate with χ as follows:

$$\begin{aligned}
\langle g, \chi\rangle_{L^2(\nu)} &= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \langle r_{w,b}, \chi\rangle_{L^2(\nu)}\, \mu(\mathrm{d}w, \mathrm{d}b) + \mathbb{E}\left[(v^\mathsf{T}\mathbf{x} + c)\chi(\mathbf{x})\right] \\
&\leq \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left|\langle r_{w,b}, \chi\rangle_{L^2(\nu)}\right| |\mu(\mathrm{d}w, \mathrm{d}b)| \\
&\leq \frac{8}{d}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} |\mu(\mathrm{d}w, \mathrm{d}b)| = \frac{8\,|\mu|}{d}.
\end{aligned}$$

Thus, $\langle g_\mu, \chi\rangle \geq \alpha$ only if $|\mu| \geq \frac{\alpha d}{8}$. ■

Fact 1 *For any measure ν_0 over $\{\pm 1\}^d$, $g \in L^2(\nu_0)$, $h : \{\pm 1\}^d \to \{\pm 1\}$, and $\alpha \in (0, 1)$, if $\|g - h\|_{L^2(\nu_0)} \leq 1 - \alpha$, then $\langle g, h\rangle_{L^2(\nu_0)} \geq \alpha$.*

Proof. The claim is a consequence of the fact $\langle h, h\rangle_{L^2(\nu_0)} = 1$ and Cauchy-Schwarz:

$$\begin{aligned}
\langle g, h\rangle_{L^2(\nu_0)} &= \langle h, h\rangle_{L^2(\nu_0)} + \langle g - h, h\rangle_{L^2(\nu_0)} \\
&= 1 + \langle g - h, h\rangle_{L^2(\nu_0)} \\
&\geq 1 - \|g - h\|_{L^2(\nu_0)} \\
&\geq 1 - (1 - \alpha) = \alpha.
\end{aligned}$$
■

Lemma 30 *For $d \geq 8$, $w \in \mathbb{R}^d$ with $\|w\|_2 \leq 1$, and $b \in [-2\sqrt{d}, 2\sqrt{d}]$, the neuron $r_{w,b}(x) := \varphi_{\mathrm{r}}(w^\mathsf{T} x + b)$ satisfies $|\langle r_{w,b}, \chi\rangle| \leq \frac{8}{d}$.*

Remark 31 *Lemma 30 is asymptotically tight. For even d, consider the "single-blade" sawtooth function*

$$s_{\mathbf{1},0}(x) = \sqrt{d}(r_{\mathbf{1}/\sqrt{d},1/\sqrt{d}}(x) - 2r_{\mathbf{1}/\sqrt{d},0}(x) + r_{\mathbf{1}/\sqrt{d},-1/\sqrt{d}}(x))$$

that satisfies $s_{\mathbf{1},0}(x) = \chi(x)\mathbb{1}\{\mathbf{1}^\mathsf{T} x = 0\}$. Then,

$$\langle s_{\mathbf{1},0}, \chi\rangle_{L^2(\nu)} = \frac{1}{2^d}\binom{d}{d/2} \geq \frac{1}{\sqrt{2d}},$$

and thus there exists b with $|\langle r_{\mathbf{1}/\sqrt{d},b}, \chi\rangle_{L^2(\nu)}| \geq \frac{1}{4\sqrt{2d}}$.

Proof. We directly bound the inner product by showing that we can bound a discrete second derivative. For any $x \in \{\pm 1\}^d$, let $x^j \in \{\pm 1\}^d$ denote x with a flipped jth bit. That is,

$x_i^j = (-1)^{\mathbb{1}\{i=j\}} x_i$. Observe that $\chi(x) = -\chi(x^j)$.

$$\begin{aligned}
|\langle r_{w,b}, \chi\rangle| &= \frac{1}{2^d}\left|\sum_x r_{w,b}(x)\chi(x)\right| \\
&= \frac{1}{4\cdot 2^d}\left|\sum_x (r_{w,b}(x)\chi(x) + r_{w,b}(x^j)\chi(x^j) + r_{w,b}(x^{j'})\chi(x^{j'}) + r_{w,b}(x^{j,j'})\chi(x^{j,j'}))\right| \\
&= \frac{1}{4\cdot 2^d}\left|\sum_x \chi(x)(r_{w,b}(x) - r_{w,b}(x^j) - r_{w,b}(x^{j'}) + r_{w,b}(x^{j,j'}))\right| \\
&\le \frac{1}{4\cdot 2^d}\sum_x |r_{w,b}(x) - r_{w,b}(x^j) - r_{w,b}(x^{j'}) + r_{w,b}(x^{j,j'})|.
\end{aligned}$$

We say that (x, x^j) is *cut* and denote $(x, x^j) \in C_j$ if x and x^j lie on the opposite side of the "hinge" of the neuron $r_{w,b}$, that is $\operatorname{sign}(w^\top x - b) \neq \operatorname{sign}(w^\top x^j - b)$. Let $S_{x,j,j'} = \{x, x^j, x^{j'}, x^{j,j'}\}$ represent a "square" in $\{\pm 1\}^d$, and let $S_{x,j,j'} \in C_{j,j'}$ if any of its edges (x, x^j), $(x, x^{j'})$, $(x^j, x^{j,j'})$, $(x^{j'}, x^{j,j'})$ are cut. We bound the term inside the sum by considering two cases.

1. If $S_{x,j,j'} \notin C_{j,j'}$, then $|r_{w,b}(x) - r_{w,b}(x^j) - r_{w,b}(x^{j'}) + r_{w,b}(x^{j,j'})| = 0$.
2. Otherwise, the quantity is bounded by Lipschitzness:

$$\begin{aligned}
|r_{w,b}(x) - r_{w,b}(x^j) - r_{w,b}(x^{j'}) + r_{w,b}(x^{j,j'})| &\le |r_{w,b}(x) - r_{w,b}(x^j)| + |r_{w,b}(x^{j'}) - r_{w,b}(x^{j,j'})| \\
&\le |w^\top x - w^\top x^j| + |w^\top x^{j'} - w^\top x^{j,j'}| = 4|w_j|.
\end{aligned}$$

Therefore, $|\langle r_{w,b}, \chi\rangle| \le \min_{j\neq j'} \frac{1}{2^d}\left|C_{j,j'}\right| |w_j|$. It remains to bound $\left|C_{j,j'}\right|$ and $|w_j|$ for some j and j'. By employing a bound on the total number of cut edges (O'Neil, 1971):

$$\frac{1}{d}\sum_{j=1}^{d} |C_j| \le \frac{1}{2d}\cdot\left\lceil\frac{d}{2}\right\rceil\binom{d}{\lfloor d/2\rfloor} \le \frac{2^d}{2\sqrt{d}}.$$

As a result, at most $\frac{d}{2}$ choices of j satisfy $|C_j| \ge 2^d/\sqrt{d}$. Because $\|w\|_2 \le 1$, at most $\frac{d}{4}$ coordinates j have $|w_j| \ge 2/\sqrt{d}$. Thus, there exist at *least* $\frac{d}{4}$ coordinates j satisfying both $|C_j| \le 2^d/\sqrt{d}$ and $|w_j| \le 2/\sqrt{d}$. Assuming $d \ge 8$, let j, j' be two of those coordinates. Since $|C_{j,j'}| \le 2|C_j| + 2|C_{j'}|$, we conclude that $|C_{j,j'}| \le 4\cdot 2^d/\sqrt{d}$, which gives the desired bound on the inner product. ■

D.4. $\mathcal{R}$-norm lower bound for sampled parity datasets

Theorem 32 *Fix any $\delta \in (0,1)$ and $\alpha = \omega(1/d)$, and assume $n \ge O(d^3(\log d + \log(1/\delta)))$. With probability at least $1-\delta$, $\inf\{\|g\|_{\mathcal{R}} : \|g - \chi\|_{L^2(\boldsymbol{\nu}_n)} \le 1 - \alpha\} \ge \Omega(\alpha d)$.*

Proof. Let g be a function with finite $\mathcal{R}$-norm which satisfies the $L^2(\boldsymbol{\nu}_n)$ approximability condition, which admits an integral representation due to Proposition 25. That is,

$$g(x) = \int_{\mathbb{S}^{d-1}\times[-\sqrt{d},\sqrt{d}]} (\varphi_{\mathrm{r}}(w^\top x + b) - \varphi_{\mathrm{r}}(b))\mu(\mathrm{d}w, \mathrm{d}b) + c + v^\top x \quad \forall x \in \Omega$$

for some measure μ and $v \in \mathbb{R}^d, c = g(0)$. Moreover, g can be represented compactly as $g(x) = \bar{g}_\mu(x) + v^\intercal x + c$ where $\bar{g}_\mu(x) = g_\mu(x) - g_\mu(0)$.

By Fact 1, $\|g - \chi\|_{L^2(\boldsymbol{\nu}_n)} \leq 1 - \alpha$ only if $\langle g, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} \geq \alpha$. We use this correlation to prove lower bounds on $|\mu|$ (the total variation of measure μ). At a high level, we upper-bound

$$\langle g, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} = \langle \bar{g}_\mu, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} + \langle v^\intercal x + c, \chi(x) \rangle_{L^2(\boldsymbol{\nu}_n)}$$

in terms of $|\mu|$ by relating quantities in $L^2(\boldsymbol{\nu}_n)$ with their $L^2(\nu)$ counterparts. We show that each component of the sum is small for sufficiently large n and d.

We first bound the correlation of the linear combination of neurons with parity, proving upper bounds on $\langle g, \chi \rangle_{L^2(\boldsymbol{\nu}_n)}$. We denote $\bar{r}_{w,b}(x) = \varphi_{\mathrm{r}}(w^\intercal x + b) - \varphi_{\mathrm{r}}(b)$ to be the adjusted ReLU. By the triangle inequality,

$$\begin{aligned} \langle \bar{g}_\mu, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} &\leq \int_{\mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}]} \left| \langle \bar{r}_{w,b}, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} \right| |\mu| \,(dw, db) \\ &\leq |\mu| \sup_{w \in \mathbb{S}^{d-1}, b \in [-\sqrt{d}, \sqrt{d}]} \left| \langle \bar{r}_{w,b}, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} \right|. \end{aligned}$$

Lemmas 30 and 35 together bound the correlation of any neuron $\bar{r}_{w,b}$ with χ. That is, for any $w \in \mathbb{S}^{d-1}$ and $b \in [-\sqrt{d}, \sqrt{d}]$, with probability at least $1 - \delta/3$:

$$\left| \langle \bar{r}_{w,b}, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} \right| \leq \left| \langle \bar{r}_{w,b}, \chi \rangle_{L^2(\nu)} \right| + C_1 \sqrt{\frac{d(\ln n + \ln(3/\delta))}{n}} \leq \frac{8}{d} + 2C_1 \sqrt{\frac{d \ln n}{n}} \leq \frac{C_2}{d},$$

where C_1 is the constant from Lemma 35 and $n > C(d^3(\log d + \log(1/\delta)))$ by assumption.

We now show that the linear components cannot be substantially correlated with the parity function and bound $\langle v^\intercal x + c, \chi \rangle_{L^2(\boldsymbol{\nu}_n)}$. Because no linear term correlates with the full parity dataset, Lemma 34 provides an upper bound on the inner product between the linear perturbation and sampled parity dataset and implies the following bound with probability at least $1 - \delta/3$:

$$\begin{aligned} \left| \langle v^\intercal x + c, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} \right| &\leq 8 \max\{|\mu|, 1\} \sup_{\substack{|c| \leq 1 \\ \|v\|_2 \leq 1}} \left| \langle v^\intercal x + c, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} \right| \\ &= 8 \max\{|\mu|, 1\} \left(\left| \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \right| + \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \mathbf{x}_i \right\|_2 \right). \end{aligned}$$

By Lemma 33 and our assumptions on n, we bound the two data-dependent terms with probability at least $1 - \frac{\delta}{3}$ for some absolute constant C_2:

$$\begin{aligned} \left| \langle v^\intercal x + c, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} \right| &\leq 8 \max\{|\mu|, 1\} \left(\sqrt{\frac{2 \ln(12/\delta)}{n}} + 2\sqrt{\frac{d}{n}} \right) \\ &\leq \frac{C_2}{d} \max\{|\mu|, 1\}. \end{aligned}$$

Combining both bounds, we have with probability at least $1 - \delta$,

$$\alpha \leq \langle \bar{g}_\mu(x) + v^\intercal x + c, \chi \rangle_{L^2(\boldsymbol{\nu}_n)} \leq \frac{C_2}{d} \left(|\mu| + \max\{|\mu|, 1\} \right) \leq \frac{2C_2}{d} \max\{|\mu|, 1\}.$$

Therefore, we conclude

$$|\mu| \geq \frac{\alpha d}{2C_2} - 1. \qquad \blacksquare$$

Lemma 33 *Fix any $\delta \in (0, 1)$. Assume $n \geq O(\log(1/\delta))$ and $n = \omega(d)$, let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[n]}$ be the sampled parity dataset (where $\mathbf{y}_i = \chi(\mathbf{x}_i)$ for all $i \in [n]$), and let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the data matrix containing all samples. All of the following hold with probability $1 - \delta$:*

(i) $\left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i\right| \leq \sqrt{\frac{2\ln(4/\delta)}{n}}$;

(ii) $\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\right\|_2 \leq 2\sqrt{\frac{d}{n}}$;

(iii) $\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i\mathbf{x}_i\right\|_2 \leq 2\sqrt{\frac{d}{n}}$; *and*

(iv) $\frac{3}{4}\sqrt{n} \leq \sigma_d(\mathbf{X}) \leq \sigma_1(\mathbf{X}) \leq 2\sqrt{n}$.

Proof. Claim *(i)* holds with probability at least $1 - \frac{\delta}{2}$ as a result of a standard application of Hoeffding's inequality (Lemma 17) to a sum of Rademacher random variables.

Claim *(iv)* also holds with probability at least $1 - \frac{\delta}{2}$, since Lemma 22 and the assumptions on n imply that

$$\sigma_1(\mathbf{X}) \leq \sqrt{n} + C\left(\sqrt{d} + \sqrt{\ln\frac{2}{\delta}}\right) \leq 2\sqrt{n}$$

and

$$\sigma_d(\mathbf{X}) \geq \sqrt{n} - C\left(\sqrt{d} - \sqrt{\ln\frac{2}{\delta}}\right) \geq \frac{3}{4}\sqrt{n}.$$

Claims *(ii)* and *(iii)* follow from the singular value bounds on $\mathbf{X}$.

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\right\|_2 \leq \frac{1}{n}\sqrt{\operatorname{tr}(\mathbf{X}^\intercal\mathbf{X})} \leq \frac{1}{n}\cdot 2\sqrt{nd} = 2\sqrt{\frac{d}{n}};$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i\mathbf{x}_i\right\|_2 \leq \frac{1}{n}\sqrt{\mathbf{y}^\intercal\mathbf{X}\mathbf{X}^\intercal\mathbf{y}} \leq \frac{1}{n}\cdot\sigma_1(\mathbf{X})\sqrt{d} \leq 2\sqrt{\frac{d}{n}}. \qquad \blacksquare$$

Lemma 34 *Fix any $\delta \in (0, 1)$. Assume $n \geq O(\log(1/\delta))$ and $n = \omega(d)$. With probability at least $1 - \delta$ over the random measure $\boldsymbol{\nu}_n$, if $\mu \in \mathcal{M}$ satisfies $\|g_\mu(\mathbf{x}) + c + v^\intercal\mathbf{x} - \chi(\mathbf{x})\|_{L^2(\boldsymbol{\nu}_n)} \leq 1$, then $\max\{|c + g_\mu(0)|, \|v\|_2\} \leq 8\max\{|\mu|, 1\}$.*

Proof. We draw inspiration from the fact that the full parity dataset is orthogonal to any linear term and can never be well-approximated with large linear components. In other words, the square loss on approximating the full parity dataset with a linear function is minimized by the constant-zero function and strictly worsens as the linear terms increase. That is, orthogonality ensures that $\|c+v^\intercal x-\chi(x)\|_{L^2(\nu)}^2 = 1+|c|^2+\|v\|_2^2$. Thus, having an upper bound on the squared error imposes similar upper bounds on the norms of the linear terms. We make a similar argument for the *sampled* parity dataset, where we replace ν with $\boldsymbol{\nu}_n$.

Without loss of generality, we incorporate $g_\mu(0)$ into c and define $\bar{g}_\mu(x) = g_\mu(x) - g_\mu(0)$ which can be also represented as $\bar{g}_\mu(x) = \int \bar{r}_{w,b}(x)\mu(dw, db)$ where $\bar{r}_{w,b} = \varphi_{\mathrm{r}}(w^\intercal x + b) - \varphi_{\mathrm{r}}(b)$. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the collection of samples $\mathbf{x}_i$ and let $\mathbf{y}_i = \chi(\mathbf{x}_i)$. We bound the squared loss of the linear component $v^\intercal x + c$, ignoring the neural network $\bar{g}_\mu$:

$$\begin{aligned}\|c + v^\intercal x - \chi(x)\|^2_{L^2(\boldsymbol{\nu}_n)} &= 1 + c^2 + v^\intercal\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\intercal\right)v - \frac{2}{n}v^\intercal\left(\sum_{i=1}^n(\mathbf{y}_i - c)\mathbf{x}_i\right) - \frac{2c}{n}\sum_{i=1}^n \mathbf{y}_i \\ &\geq 1 + c^2 + \frac{1}{n}\|v\|_2^2\,\sigma_d(\mathbf{X})^2 \\ &\quad - 2\|v\|_2\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{x}_i\right\|_2 + |c|\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\right\|_2\right) - 2|c|\left|\frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\right|.\end{aligned}$$

With probability $1 - \delta$, all events of Lemma 33 hold, and we use them to lower-bound the squared loss.

$$\begin{aligned}\|c + v^\intercal x - \chi(x)\|^2_{L^2(\boldsymbol{\nu}_n)} &\geq 1 + c^2 + \frac{9}{16}\|v\|_2^2 - 4\sqrt{\frac{d}{n}}(1 + |c|)\|v\|_2 - \frac{2\sqrt{2\ln(8/\delta)}}{\sqrt{n}}|c| \\ &\geq \frac{1}{4}\max\{|c|, \|v\|_2\}^2.\end{aligned}$$

where we have used the assumptions on n and the AM/GM inequality. We now provide upper bounds on the square loss based on measure μ using the triangle inequality:

$$\|c + v^\intercal x - \chi(x)\|_{L^2(\boldsymbol{\nu}_n)} \leq \|\bar{g}_\mu\|_{L^2(\boldsymbol{\nu}_n)} + \|\bar{g}_\mu(x) + c + v^\intercal x - \chi(x)\|_{L^2(\boldsymbol{\nu}_n)} \leq \|\bar{g}_\mu\|_{L^2(\boldsymbol{\nu}_n)} + 1.$$

We may now connect $L^2(\boldsymbol{\nu}_n)$ norm of $\bar{g}_\mu$ to its variational norm. We bound the output of $\bar{g}_\mu$ on a single input $\mathbf{x}_i$ by employing Cauchy-Schwarz:

$$\begin{aligned}\bar{g}_\mu(\mathbf{x}_i)^2 &\leq \left(\int |\bar{r}_{w,b}(\mathbf{x}_i)|\,|\mu|\,(\mathrm{d}w, \mathrm{d}b)\right)^2 \\ &\leq |\mu| \int \bar{r}_{w,b}(\mathbf{x}_i)^2\,|\mu|\,(\mathrm{d}w, \mathrm{d}b).\end{aligned}$$

We sum over all i to bound the norm of $\bar{g}_\mu$:

$$\begin{aligned}\|\bar{g}_\mu(x)\|^2_{L^2(\boldsymbol{\nu}_n)} &\leq |\mu| \int \|\bar{r}_{w,b}(x)\|^2_{L^2(\boldsymbol{\nu}_n)}\,|\mu|\,(\mathrm{d}w, \mathrm{d}b) \leq |\mu|^2 \sup_{w\in\mathbb{S}^{d-1}, |b|\leq\sqrt{d}} \|\bar{r}_{w,b}\|^2_{L^2(\boldsymbol{\nu}_n)} \\ &\leq |\mu|^2 \sup_{w\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n |w^\intercal \mathbf{x}_i|^2 = |\mu|^2\,\frac{\sigma_1(\mathbf{X})^2}{n} \leq 4\,|\mu|^2.\end{aligned}$$

The second inequality relies on the Lipschitzness of φ_{r}. Combining all the above,

$$\frac{1}{2}\max\{|c|, \|v\|_2\} \leq \|c + v^\intercal x - \chi(x)\|_{L^2(\boldsymbol{\nu}_n)} \leq 1 + \|g_\mu\|_{L^2(\boldsymbol{\nu}_n)} < 2 + 2\,|\mu|. \qquad \blacksquare$$

Lemma 35 *For $\bar{r}_{w,b}(x) = \varphi_{\mathrm{r}}(w^\intercal x + b) - \varphi_{\mathrm{r}}(b)$ and $n \geq d$, there exists an absolute constant C such that for any $\delta \in (0,1)$ with probability at least $1-\delta$,*

$$\left|\langle \bar{r}_{w,b}, \chi\rangle_{L^2(\boldsymbol{\nu}_n)}\right| \leq \left|\langle \bar{r}_{w,b}, \chi\rangle_{L^2(\nu)}\right| + C\sqrt{\frac{d(\ln n + \ln(1/\delta))}{n}},$$

for all $w \in \mathbb{S}^{d-1}, b \in [-\sqrt{d}, \sqrt{d}]$.

Proof. Observe that the inner product over the sampled parity dataset is an unbiased estimate of the inner product over the full parity dataset,

$$\mathbb{E}\left[\langle \bar{r}_{w,b}, \chi\rangle_{L^2(\boldsymbol{\nu}_n)}\right] = \langle \bar{r}_{w,b}, \chi\rangle_{L^2(\nu)}.$$

Let $\mathbf{Z}_{w,b}$ denote the deviation from the mean, i.e.

$$\mathbf{Z}_{w,b} = \langle \bar{r}_{w,b}, \chi\rangle_{L^2(\boldsymbol{\nu}_n)} - \langle \bar{r}_{w,b}, \chi\rangle_{L^2(\nu)}.$$

We use standard concentration of measure techniques for the following steps:

1. $\mathbf{Z}_{w,b}$ is Lipschitz in terms of its parameterization (w,b) in the sense that $|\mathbf{Z}_{w_1,b_1} - \mathbf{Z}_{w_2,b_2}| \leq 4\sqrt{d}\gamma((w_1,b_1),(w_2,b_2))$, where γ is a distance defined later on.
2. $\mathbf{Z}_{w,b}$ is $O(\frac{1}{\sqrt{n}})$-subgaussian for fixed w, b.
3. $\mathbb{E}\left[\sup_{w\in\mathbb{S}^{d-1}, b\in[-\sqrt{d},\sqrt{d}]} |\mathbf{Z}_{w,b}|\right] = O(\sqrt{\frac{d}{n}})$ using a covering argument.
4. The maximum of $|\mathbf{Z}_{w,b}|$ is close to its expectation due to the bounded difference inequality.

(Step 1) Using the fact that φ_{r} is 1-Lipschitz and triangle inequality,

$$\begin{aligned}
|\mathbf{Z}_{w_1,b_1} - \mathbf{Z}_{w_2,b_2}| &\leq \left|\langle \bar{r}_{w_1,b_1}, \chi\rangle_{L^2(\boldsymbol{\nu}_n)} - \langle \bar{r}_{w_2,b_2}, \chi\rangle_{L^2(\boldsymbol{\nu}_n)}\right| + \left|\langle \bar{r}_{w_1,b_1}, \chi\rangle_{L^2(\nu)} - \langle \bar{r}_{w_2,b_2}, \chi\rangle_{L^2(\nu)}\right| \\
&\leq 2\left\|\bar{r}_{w_1,b_1} - \bar{r}_{w_2,b_2}\right\|_{L^\infty(\nu)} \\
&\leq 2(\left\|\bar{r}_{w_1,b_1} - \bar{r}_{w_2,b_1}\right\|_{L^\infty(\nu)} + \left\|\bar{r}_{w_2,b_1} - \bar{r}_{w_2,b_2}\right\|_{L^\infty(\nu)}) \\
&\leq 2\left(\max_{x\in\{\pm1\}^d} (w_1 - w_2)^\intercal x + 2\left|b_1 - b_2\right|\right) \\
&\leq 4\sqrt{d}\left(\left\|w_1 - w_2\right\|_2 + \frac{|b_1 - b_2|}{\sqrt{d}}\right) =: 4\sqrt{d}\gamma((w_1,b_1),(w_2,b_2)).
\end{aligned}$$

Thus $\mathbf{Z}_{w,b}$ is $4\sqrt{d}$-Lipschitz with respect to γ.

(Step 2) We bound the subgaussianity of $\mathbf{Z}_{w,b}$.

$$\begin{aligned}
\left\|\mathbf{Z}_{w,b}\right\|_{\psi_2} \leq C_1\left\|\langle \bar{r}_{w,b}, \chi\rangle_{L^2(\boldsymbol{\nu}_n)}\right\|_{\psi_2} &= C_1\left\|\sum_{i=1}^n \mathbf{y}_i(\varphi_{\mathrm{r}}(w^\intercal \mathbf{x}_i + b) - \varphi_{\mathrm{r}}(b))\right\|_{\psi_2} \\
&\leq \frac{C_2}{\sqrt{n}}\left\|\mathbf{y}_1(\varphi_{\mathrm{r}}(w^\intercal \mathbf{x}_1 + b) - \varphi_{\mathrm{r}}(b))\right\|_{\psi_2} \\
&\leq \frac{C_2}{\sqrt{n}}\left\|\varphi_{\mathrm{r}}(w^\intercal \mathbf{x}_1 + b) - \varphi_{\mathrm{r}}(b)\right\|_{\psi_2} \\
&\leq \frac{C_2}{\sqrt{n}}\left\|w^\intercal \mathbf{x}_1\right\|_{\psi_2} \leq \frac{2C_2}{\sqrt{n}}
\end{aligned}$$

The first, second, and fourth inequalities rely on the centering, averaging, and Lipschitzness properties of subgaussian random variables in Lemma 21. The third inequality follows from $|\mathbf{y}_1| = 1$, and the final is due to the 2-subgaussianity of a vector with i.i.d. Rademacher components.

(Step 3) Let $\mathcal{N}_\epsilon$ be an ϵ-covering of $\mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}]$ with respect to γ. We bound its size using the standard ϵ-net result in Lemma 24 for $\epsilon \le 2$.

$$\begin{aligned}\mathcal{N}\left(\epsilon, \mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}], \gamma\right) &\le \mathcal{N}\left(\frac{\epsilon}{2}, \mathbb{S}^{d-1}, \|\cdot\|_2\right) \times \mathcal{N}\left(\frac{\epsilon}{2}, [-1,1], |\cdot|\right) \\ &\le \left(\frac{6}{\epsilon}\right)^d \cdot \frac{4}{\epsilon} \le \left(\frac{6}{\epsilon}\right)^{d+1}.\end{aligned}$$

We bound the expected maximum deviation over all w and b by employing a bound on the expected maximum of subgaussian random variables (Lemma 21), applying the covering numbers argument, letting $\pi(w,b) = \arg\min_{(w',b') \in \mathcal{N}_\epsilon} \gamma((w,b),(w',b'))$, and setting $\epsilon := 1/\sqrt{n}$.

$$\begin{aligned}\mathbb{E}\left[\sup_{w \in \mathbb{S}^{d-1}, b \in [-\sqrt{d}, \sqrt{d}]} |\mathbf{Z}_{w,b}|\right] &\le \mathbb{E}\left[\sup_{w,b} \left|\mathbf{Z}_{w,b} - \mathbf{Z}_{\pi(w,b)}\right|\right] + \mathbb{E}\left[\sup_{(w,b) \in \mathcal{N}_\epsilon} |\mathbf{Z}_{w,b}|\right] \\ &\le 4\sqrt{d}\epsilon + \frac{2C_2}{\sqrt{n}} \sqrt{\ln \mathcal{N}\left(\epsilon, \mathbb{S}^{d-1} \times [-\sqrt{d}, \sqrt{d}], \gamma\right)} \\ &\le 4\sqrt{d}\epsilon + 2C_2 \sqrt{\frac{d+1}{n} \ln \frac{6}{\epsilon}} \le C_3 \sqrt{\frac{d \ln n}{n}}.\end{aligned}$$

(Step 4) We conclude by showing that $\sup_{w,b} |\mathbf{Z}_{w,b}|$ is close to its expectation with high probability due to the McDiarmid's inequality (Lemma 20). Consider a perturbation where $\mathbf{x}_i$ is replaced by some $\mathbf{x}_i' \in \{\pm 1\}^d$ with $\mathbf{y}_i' = \chi(\mathbf{x}_i')$, and let $\mathbf{Z}_{w,b}^i$ denote the resulting deviation term.

$$\begin{aligned}\left|\sup_{w,b} |\mathbf{Z}_{w,b}| - \sup_{w,b} \left|\mathbf{Z}_{w,b}^i\right|\right| &\le \sup_{w,b} \left|\mathbf{Z}_{w,b} - \mathbf{Z}_{w,b}^i\right| = \frac{1}{n} \sup_{w,b} \left|\mathbf{y}_i \bar{r}_{w,b}(\mathbf{x}_i) - \mathbf{y}_i' \bar{r}_{w,b}(\mathbf{x}_i')\right| \\ &\le \frac{1}{n} \sup_{w,b} \left[\left|\bar{r}_{w,b}(\mathbf{x}_i) - \bar{r}_{w,b}(\mathbf{x}_i')\right| + \left|(\mathbf{y}_i - \mathbf{y}_i')\bar{r}_{w,b}(\mathbf{x}_i)\right|\right] \\ &\le \frac{1}{n}\left[\left\|\mathbf{x}_i - \mathbf{x}_i'\right\|_2 + 2\left\|\mathbf{x}_i\right\|_2\right] \le \frac{4\sqrt{d}}{n}\end{aligned}$$

Hence, with probability at least $1 - \delta$:

$$\sup_{w,b} |\mathbf{Z}_{w,b}| \le \sqrt{\frac{8d \ln 1/\delta}{n}} + \mathbb{E}\left[\sup_{w,b} |\mathbf{Z}_{w,b}|\right] \le C_4 \sqrt{\frac{d(\ln n + \ln 1/\delta)}{n}}.$$

The bound in the lemma statement immediately follows. ■

Appendix E. Proofs for Section 4

E.1. Proofs for Section 4.1

Lemma 36 *Fix $S \subseteq [d]$ with $|S| \ge 3$, and let $\mathbf{x} \sim \mathrm{Unif}(\{\pm 1\}^d)$. Conditional on the value of $\chi_S(\mathbf{x})$, the random vector $\mathbf{x}$ is mean-zero, isotropic, and satisfies*

$$\mathbb{E}\left[\exp(u^\intercal \mathbf{x}) \mid \chi_S(\mathbf{x})\right] \le \exp(\|u\|_2^2)$$

for all $u = (u_1, \dots, u_d) \in \mathbb{R}^d$.

Proof. The assumption $|S| \geq 3$ implies that, conditioned on $\chi_S(\mathbf{x})$, the $\{\mathbf{x}_i\}_{i \in [d]}$ are mean-zero and pairwise uncorrelated. So it remains to show that, for any vector $u = (u_1, \dots, u_d) \in \mathbb{R}^d$,

$$\mathbb{E}\left[\exp(u^\intercal \mathbf{x}) \mid \chi_S(\mathbf{x})\right] \leq \exp(\|u\|_2^2).$$

So fix u, and fix any $i \in S$. Let u_{-i} (respectively, $\mathbf{x}_{-i}$) be the vector obtained from u (respectively, $\mathbf{x}$) by removing the i-th entry. Observe that $\mathbf{x}_{-i} \mid \chi_S(\mathbf{x}) \sim \mathrm{Unif}(\{\pm 1\}^{d-1})$, and also that $\mathbf{x}_i \mid \chi_S(\mathbf{x}) \sim \mathrm{Unif}(\{\pm 1\})$. (But, of course, $\mathbf{x}_{-i}$ and $\mathbf{x}_i$ are not conditionally independent given $\chi_S(\mathbf{x})$.) Therefore, using Cauchy-Schwarz,

$$\begin{aligned}
\mathbb{E}\left[\exp(u^\intercal \mathbf{x}) \mid \chi_S(\mathbf{x})\right] &= \mathbb{E}\left[\exp(u_{-i}^\intercal \mathbf{x}_{-i}) \exp(u_i \mathbf{x}_i) \mid \chi_S(\mathbf{x})\right] \\
&\leq \sqrt{\mathbb{E}\left[\exp(2u_{-i}^\intercal \mathbf{x}_{-i}) \mid \chi_S(\mathbf{x})\right]} \sqrt{\mathbb{E}\left[\exp(2u_i \mathbf{x}_i) \mid \chi_S(\mathbf{x})\right]} \\
&\leq \sqrt{\exp(\|2u_{-i}\|_2^2/2)} \sqrt{\exp((2u_i)^2/2)} \\
&= \exp(\|u\|_2^2).
\end{aligned}$$

Above, the second inequality uses the moment generating function bound from Lemma 23, as well as the conditional independence of $\{\mathbf{x}_j : j \neq i\}$ given $\chi(\mathbf{x})$. ■

Lemma 10 *There is an absolute constant $c > 0$ such that the following holds. If $n \leq cd^2$, and $\mathbf{x}_1, \dots \mathbf{x}_n \sim_{\mathrm{iid}} \nu$, then with probability at least $1/2$, there exists $\mathbf{g} \colon \Omega \to \mathbb{R}$ with $\mathbf{g}(\mathbf{x}_i) = \chi(\mathbf{x}_i)$ for all $i \in [n]$ and $\|\mathbf{g}\|_{\mathcal{R}} \leq 4n\sqrt{\ln d}/d$.*

Proof. Throughout, we take $C > 0$ to be a suitably large constant, and we assume $n \leq d^2/C$. The construction of $\mathbf{g} \colon \Omega \to \mathbb{R}$ is based on typical statistical behavior of the random examples $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)$, where $\mathbf{y}_i := \chi_S(\mathbf{x}_i)$ for each $i \in [n]$. We may assume that $n \geq d$, since otherwise the examples can be perfectly fit with a linear function $\mathbf{g}$, and this function has $\|\mathbf{g}\|_{\mathcal{R}} = 0$. So, combining the assumption $n \geq d$ with the assumption $n \leq d^2/C$ implies that $d \geq C$. Observe that $\mathbf{y}_1, \dots, \mathbf{y}_n$ are i.i.d. $\mathrm{Unif}(\{\pm 1\})$ random variables. Since $n \geq d \geq C$, it follows by standard binomial tail bounds that with probability at least $5/6$ over the realizations of $\mathbf{y}_1, \dots, \mathbf{y}_n$, the number of $\mathbf{y}_i$ that are equal to 1 is at least $n/3$, and also that the number of $\mathbf{y}_i$ that are equal to -1 is also at least $n/3$. We henceforth condition on this "good event" (which depends only on $\mathbf{y}_1, \dots, \mathbf{y}_n$).

To help define our construction of $\mathbf{g} \colon \Omega \to \mathbb{R}$ and set up the rest of the analysis, we partition $[n]$ into disjoint groups $G_1, G_2, \dots, G_m$ so that for each $j \in [m]$, (i) the size $n_j := |G_j|$ of the j-th group is between $c_1 d/\ln d$ and $2c_1 d/\ln d$, and (ii) all $\mathbf{y}_i$ for $i \in G_j$ are the same (i.e., all $+1$ or all -1). Here, with foresight, we set $c_1 := 1/256$; by using $d \geq C$, we ensure that each group is non-empty, and also that $n_j < d$. The feasibility of this partitioning is ensured because, in the "good event" (and using $d \geq C$), the number of $i \in [n]$ with $\mathbf{y}_i = 1$ is at least $n/3 \geq d/3 \geq c_1 d/\ln d$, and same for the number of $i \in [n]$ with $\mathbf{y}_i = -1$. Let $z^{(j)}$ denote the common $\mathbf{y}_i$ value for all $i \in G_j$. Finally, note that the number of groups m satisfies $m \leq n\ln(d)/(c_1 d)$.

We now define our construction of $\mathbf{g} \colon \Omega \to \mathbb{R}$. Let $\mathbf{A}_j$ denote the random $n_j \times d$ matrix whose rows are the $\mathbf{x}_i^\intercal$ for $i \in G_j$, and define the random vector $\mathbf{w}^{(j)} := \mathbf{A}_j^\dagger (z^{(j)} \mathbf{1})$. Observe that $\mathbf{w}^{(j)}$ is a least squares solution to the system of linear equations $\{\mathbf{x}_i^\top w = \mathbf{y}_i : i \in G_j\}$, since $\mathbf{y}_i = z^{(j)}$ for all $i \in G_j$. We define $\mathbf{g}$ as follows:

$$\mathbf{g}(x) = \sum_{j=1}^{m} z^{(j)} \varphi_{\mathrm{r}}(2z^{(j)}\mathbf{w}^{(j)\top}x - 1).$$

To analyze our construction, we consider the realizations of $\mathbf{x}_1, \dots, \mathbf{x}_n$, and establish some basic properties that hold with sufficiently high probability (conditional on the "good event"). Note that within a group G_j, the $\{\mathbf{x}_i\}_{i\in G_j}$ are (conditionally) iid, and the realizations across groups are also (conditionally) independent.

We claim that with probability at least $5/6$ (conditional on the "good event"),

- (P1) $\mathbf{w}^{(j)\top}\mathbf{x}_i = \mathbf{y}_i$ for all $j \in [m]$ and $i \in G_j$;
- (P2) $\|\mathbf{w}^{(j)}\|_2 \le 2\sqrt{n_j/d}$ for all $j \in [m]$.

To establish this claim, we lower-bound the n_j-th largest singular value $\sigma_{n_j}(\mathbf{A}_j)$. Note that $\sigma_{n_j}(\mathbf{A}_j)$ is at least the corresponding singular value of the $n_j \times (d-1)$ submatrix $\mathbf{B}_j$ obtained from $\mathbf{A}_j$ by removing the t-th column of $\mathbf{A}_j$ for some $t \in S$. (If S is empty, we may remove any column.) Since the rows of $\mathbf{A}_j$ are independent, and since the entries of $\mathbf{x}_i$ after removing the t-th one are iid $\mathrm{Unif}(\{\pm 1\})$ random variables, it follows that the $n_j \times (d-1)$ entries of $\mathbf{B}_j$ are iid $\mathrm{Unif}(\{\pm 1\})$ random variables. Hence, the rows of $\mathbf{B}_j^\top$ are independent, mean-zero, isotropic, and $O(1)$-subgaussian. By Lemma 22 and a union bound, with probability at least $1 - 2m\exp(-\min_{j\in[m]}\{n_j\})$,

$$\sigma_{n_j}(\mathbf{A}_j) \ge \sigma_{n_j}(\mathbf{B}_j^\top) \ge \sqrt{d-1} - C_2\sqrt{n_j} \ge \left(\sqrt{1-\frac{1}{d}} - C_2\sqrt{\frac{c_1}{\ln d}}\right)\sqrt{d} \quad \text{for all } j \in [m],$$

where $C_2 > 0$ is twice the absolute constant from Lemma 22, and the final inequality uses the upper-bound on n_j. The fact $d \ge C$ and the upper-bounds on m and n altogether imply that the probability of the above event is at least $5/6$, and also that $\sqrt{1-1/d} - C_2\sqrt{c_1/\ln d} \ge 1/2$. So in this event, for each $j \in [m]$, the column space of $\mathbf{A}_j$ has rank n_j, so the system of linear equations defining $\mathbf{w}^{(j)}$ is feasible, and

$$\|\mathbf{w}^{(j)}\|_2 = \|\mathbf{A}_j^\dagger(z^{(j)}\mathbf{1})\|_2 \le \sigma_1(\mathbf{A}_j^\dagger)\|\mathbf{1}\|_2 = \frac{\sqrt{n_j}}{\sigma_{n_j}(\mathbf{A}_j)} \le 2\sqrt{\frac{n_j}{d}}.$$

This establishes P1 and P2 in the event as claimed.

We further claim that with probability at least $5/6$ (conditional on the "good event"),

- (P3) $|\mathbf{w}^{(j)\top}\mathbf{x}_i| \le 4\|\mathbf{w}^{(j)}\|_2\sqrt{\ln d}$ for all $j \in [m]$ and $i \in [n] \setminus G_j$.

To establish this claim, first observe that $\mathbf{x}_i$ and $\mathbf{w}^{(j)}$ are independent for $i \notin G_j$. Moreover, by Lemma 36, conditional on $\mathbf{w}^{(j)}$ (with $G_j \not\ni i$), $\mathbf{x}_i^\top\mathbf{w}^{(j)}$ is a mean-zero random variable satisfying

$$\mathbb{E}\left[\exp(\mathbf{w}^{(j)\top}\mathbf{x}) \mid \chi_S(\mathbf{x}), \mathbf{w}^{(j)}\right] \le \exp(\|\mathbf{w}^{(j)}\|_2^2).$$

So, by Markov's inequality and a union bound, we have with probability at least $5/6$,

$$|\mathbf{w}^{(j)\top}\mathbf{x}_i| \le (\sqrt{2}\|\mathbf{w}^{(j)}\|_2)\sqrt{2\ln(12mn)} \quad \text{for all } j \in [m] \text{ and } i \in [n] \setminus G_j.$$

Using $d \geq C$ and the upper-bounds on m and n, we obtain $\sqrt{\ln(12mn)} \leq 2\sqrt{\ln d}$, and hence we deduce P3 from the above inequality.

So, by a union bound, with probability at least $2/3$ (conditional on the "good event"), the properties P1, P2, and P3 all hold simultaneously. We can now establish the desired properties of $\mathbf{g}$. Using $d \geq C$, P2, and the upper-bounds on m and n, we obtain

$$\|\mathbf{g}\|_{\mathcal{R}} \leq 2\sum_{j=1}^{m} \|\mathbf{w}^{(j)}\|_2 \leq 4\sum_{j=1}^{m} \sqrt{\frac{n_j}{d}} \leq 4\sqrt{m\sum_{j=1}^{m}\frac{n_j}{d}} = 4\sqrt{\frac{mn}{d}} \leq \frac{4n\sqrt{\ln d}}{d}.$$

Furthermore, by P1, we have for any $j \in [m]$ and $i \in G_j$,

$$2z^{(j)}\mathbf{w}^{(j)\mathsf{T}}\mathbf{x}_i - 1 = 2z^{(j)}\mathbf{y}_i - 1 = 1.$$

And by P2, P3, and the upper-bound on n_j, we have for any $j \in [m]$ and $i \in [n] \setminus G_j$,

$$2z^{(j)}\mathbf{w}^{(j)\mathsf{T}}\mathbf{x}_i - 1 \leq 2|\mathbf{w}^{(j)\mathsf{T}}\mathbf{x}_i| - 1 \leq 16\sqrt{\frac{n_j \ln d}{d}} - 1 \leq 16\sqrt{c_1} - 1 = 0,$$

and hence $\varphi_{\mathrm{r}}(2z^{(j)}\mathbf{w}^{(j)\mathsf{T}}\mathbf{x}_i - 1) = 0$. Therefore, for any $i \in [n]$, if $i \in G_j$,

$$\mathbf{g}(\mathbf{x}_i) = z^{(j)}\varphi_{\mathrm{r}}(2z^{(j)}\mathbf{w}^{(j)\mathsf{T}}\mathbf{x}_i - 1) = z^{(j)} = \mathbf{y}_i. \qquad \blacksquare$$

E.2. Proofs for Section 4.2

Theorem 11 *There is an absolute constant $C > 0$ such that the following holds. For any $\epsilon \in (0,1)$ and $\delta \in (0,1)$, if $n \geq C(\log(1/\delta) + d^3/\epsilon^2)$, then with probability at least $1-\delta$, every solution $\mathbf{g}\colon \Omega \to \mathbb{R}$ to (VP) for the sampled parity dataset satisfies $\|\chi - \mathrm{clip} \circ \mathbf{g}\|_{L^2(\nu)}^2 \leq \epsilon$, where* $\mathrm{clip}(t) := \min\{\max\{t, -1\}, 1\}$.

Proof. Let $\mathbf{G}$ denote all solutions to (VP) on the sampled parity dataset, so $\|\mathbf{g} - \chi\|_{L^2(\boldsymbol{\nu}_n)} = 0$ for all $\mathbf{g} \in \mathbf{G}$. By Proposition 25, we can write each $\mathbf{g}$ as $\mathbf{g}(x) = g_{\boldsymbol{\mu}}(x) + \mathbf{v}^{\mathsf{T}}x + \mathbf{c}$, where $\boldsymbol{\mu} \in \mathcal{M}$, $\mathbf{v} \in \mathbb{R}^d$, and $\mathbf{c} \in \mathbb{R}$. Furthermore, we can assume that $g_{\boldsymbol{\mu}}(0) = 0$ by absorbing the value of $g_{\boldsymbol{\mu}}(0)$ into $\mathbf{c}$ (at the cost of losing the evenness of $\boldsymbol{\mu}$, but evenness is not needed in the sequel). Lemma 1, Theorem 26, and Theorem 5 together imply that every $\mathbf{g} \in \mathbf{G}$ satisfies $\|\mathbf{g}\|_{\mathcal{R}} \leq Cd$ for some absolute constant $C > 0$. Let $\mathcal{E}$ be the event that $\max\{|\mathbf{c}|, \|\mathbf{v}\|_2\} \leq 8Cd$ (for all $\mathbf{g} \in \mathbf{G}$), and let $\mathcal{E}^{\mathrm{c}}$ be its complement; event $\mathcal{E}$ occurs with probability at least $1 - \delta/2$ by Lemma 34, for another absolute constant $C' > 0$.

Since, for each $\mathbf{g} \in \mathbf{G}$, we have $\mathbf{g}(\mathbf{x}_i) = \chi(\mathbf{x}_i)$ for every example $(\mathbf{x}_i, \chi(\mathbf{x}_i))$ in the sampled parity dataset, it follows that $\|\mathrm{clip} \circ \mathbf{g} - \chi\|_{L^2(\boldsymbol{\nu}_n)} = \|\mathbf{g} - \chi\|_{L^2(\boldsymbol{\nu}_n)} = 0$ for all such $\mathbf{g} \in \mathbf{G}$. For any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\mathbf{g}\in\mathbf{G}} \|\mathrm{clip}\circ\mathbf{g} - \chi\|^2_{L^2(\nu)} \geq t\right] \\
&\leq \mathbb{P}\left[\sup_{\mathbf{g}\in\mathbf{G}} \|\mathrm{clip}\circ\mathbf{g} - \chi\|^2_{L^2(\nu)} \geq t \mid \mathcal{E}\right] + \mathbb{P}\left[\mathcal{E}^{\mathrm{c}}\right] \\
&= \mathbb{P}\left[\sup_{\mathbf{g}\in\mathbf{G}} \|\mathrm{clip}\circ\mathbf{g} - \chi\|^2_{L^2(\nu)} - \|\mathrm{clip}\circ\mathbf{g} - \chi\|^2_{L^2(\boldsymbol{\nu}_n)} \geq t \mid \mathcal{E}\right] + \mathbb{P}\left[\mathcal{E}^{\mathrm{c}}\right] \\
&\leq \mathbb{P}\left[\sup_{g\in\mathcal{G}_0} \|\mathrm{clip}\circ g - \chi\|^2_{L^2(\nu)} - \|\mathrm{clip}\circ g - \chi\|^2_{L^2(\boldsymbol{\nu}_n)} \geq t\right] + \delta/2,
\end{aligned}$$

where

$$\mathcal{G}_0 := \left\{x \mapsto g(x) + v^\mathsf{T} x + c : \|g\|_{\mathcal{R}} \leq Cd,\ \max\{\|v\|_2, |c|\} \leq 8Cd\right\}.$$

Define

$$t_0 := 4\,\underbrace{\mathbb{E}\left[\sup_{g\in\mathcal{G}_0} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i g(\mathbf{x}_i)\right]}_{\mathrm{Rad}_n(\mathcal{G}_0)} + 4\sqrt{\frac{\log(2/\delta)}{n}}.$$

Above, $\mathrm{Rad}_n(\mathcal{G}_0)$ denotes the Rademacher complexity of $\mathcal{G}_0$, where $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n \sim_{\mathrm{iid}} \mathrm{Unif}(\{\pm 1\})$, independent of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Since, for any $y \in \{\pm 1\}$, the mapping $z \mapsto (y - \mathrm{clip}(z))^2 = (1 - y\,\mathrm{clip}(z))^2$ is 4-Lipschitz and has range $[-4, 4]$, it follows by standard Rademacher complexity arguments (see, e.g., Meir and Zhang, 2003, Theorem 8) that

$$\mathbb{P}\left[\sup_{g\in\mathcal{G}_0} \|\mathrm{clip}\circ g - \chi\|^2_{L^2(\nu)} - \|\mathrm{clip}\circ g - \chi\|^2_{L^2(\boldsymbol{\nu}_n)} \geq t_0\right] \leq \delta/2.$$

So it remains to show that $t_0 \leq \epsilon$ under the assumption $n \geq C_0((d^3 + \log(1/\delta))/\epsilon^2)$ for suitably large absolute constant $C_0 > 0$. The second term in the definition of t_0 is at most $\epsilon/2$ provided that C_0 is chosen large enough. To bound the first term ($\mathrm{Rad}_n(\mathcal{G}_0)$), we use the fact that

$$\mathrm{Rad}_n(\mathcal{G}_0) = \mathrm{Rad}_n(\mathcal{G}_1) + \mathrm{Rad}_n(\mathcal{G}_2)$$

where $\mathcal{G}_1 := \{g : \|g\|_{\mathcal{R}} \leq Cd\}$ and $\mathcal{G}_2 := \{x \mapsto v^\mathsf{T} x + c : \max\{\|v\|_2, |c|\} \leq 8Cd\}$. Theorem 10 of Parhi and Nowak (2021a) implies

$$\mathrm{Rad}_n(\mathcal{G}_1) \leq \frac{2\cdot(Cd)\cdot\sqrt{d}}{\sqrt{n}} = O\left(\sqrt{\frac{d^3}{n}}\right),$$

while Theorem 3 of Kakade et al. (2008) implies

$$\mathrm{Rad}_n(\mathcal{G}_2) \leq \sqrt{d+1}\cdot\sqrt{(8Cd)^2}\cdot\sqrt{\frac{2}{n}} = O\left(\sqrt{\frac{d^3}{n}}\right).$$

By choosing C_0 large enough, it follows that $\mathrm{Rad}_n(\mathcal{G}_0) \leq \epsilon/8$. Hence, we have shown that $t_0 \leq \epsilon$ as required. ■

Appendix F. Proofs for Section 5

F.1. Proof of Theorem 12

Theorem 37 (Detailed version of Theorem 12) *Suppose $f\colon \Omega \to [-1,1]$ is given by $f(x) = \phi(v^\intercal x)$ for some unit vector $v \in \mathbb{S}^{d-1}$ and some $\phi\colon [-\sqrt{d}, \sqrt{d}] \to [-1,1]$ that is L-Lipschitz and ρ-periodic for $\rho \in [\|v\|_\infty, 1]$. Let $\sigma_{\rho,v} := \sqrt{2\rho\|v\|_1 - 1}$, and fix any $\epsilon \in (0,1)$. There exists a function $g\colon \Omega \to \mathbb{R}$ such that the following properties hold:*

1. *$|f(x) - g(x)| \le \epsilon$ for all $x \in \{\pm 1\}^d$;*
2. *g is represented by a neural network of width at most*
$$O\left(\frac{dL(\sigma_{\rho,v}\sqrt{\log(1/\epsilon)} + \rho\log(1/\epsilon))}{\epsilon^2}\right);$$
3. *g satisfies*
$$\|g\|_{\mathcal{R}} = O\left(\frac{L^2(\sigma_{\rho,v}\sqrt{\log(1/\epsilon)} + \rho\log(1/\epsilon))(\sigma_{\rho,v} + \sqrt{\log(d/\epsilon)})}{\epsilon}\right).$$

Proof. We first describe the (randomized) construction of our approximating neural network $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}$. For $w \in \mathbb{Z}^d$, define $h_w\colon \mathbb{R}^d \to \mathbb{R}$ by $h_w(x) := \phi(v^\intercal x + \rho w^\intercal x)$. Let $\mathbf{w} \in \mathbb{Z}^d \setminus \{-(1/\rho)v\}$ be a random vector with distribution to be specified later in the proof. Let $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(k)}$ be i.i.d. copies of $\mathbf{w}$ for a positive integer $k > (9(d+1)\ln(2))/\epsilon$, and let $\mathbf{h}^{(j)} := h_{\mathbf{w}^{(j)}}$ for each j. Observe that each $\mathbf{h}^{(j)}$ can be written as $\mathbf{h}^{(j)}(x) = \phi^{(j)}(x^\intercal \mathbf{u}^{(j)})$ for

$$\mathbf{u}^{(j)} := \frac{1}{\|v + \rho\mathbf{w}^{(j)}\|_2}(v + \rho\mathbf{w}^{(j)}) \in \mathbb{S}^{d-1} \quad \text{and} \quad \phi^{(j)}(z) := \phi(\|v + \rho\mathbf{w}^{(j)}\|_2 z),$$

where $\phi^{(j)}\colon \mathbb{R} \to [-1,1]$ is L_j-Lipschitz for $L_j := L\|v + \rho\mathbf{w}^{(j)}\|_2$ (using the L-Lipschitzness of ϕ). Let $\tau > 0$ be a value (depending on ρ, v, and ϵ) also to be specified later. By Lemma 38 (with $t := \tau/\|v + \rho\mathbf{w}^{(j)}\|_2$ and $\delta := \epsilon/3$), there exist $\tilde{\mathbf{h}}^{(1)}, \dots, \tilde{\mathbf{h}}^{(k)}$ such that:

- (H1) $\tilde{\mathbf{h}}^{(j)}\colon \mathbb{R}^d \to \mathbb{R}$ is represented by a neural network of width at most $O(\tau L/\epsilon)$;
- (H2) $\|\tilde{\mathbf{h}}^{(j)}\|_{\mathcal{R}} = O(\tau L^2\|v + \rho\mathbf{w}^{(j)}\|_2/\epsilon)$;
- (H3) $|\tilde{\mathbf{h}}^{(j)}(x) - \mathbf{h}^{(j)}(x)| \le \epsilon/3$ for all $x \in \{\pm 1\}^d$ such that $|x^\intercal \mathbf{u}^{(j)}| \le \tau/\|v + \rho\mathbf{w}^{(j)}\|_2$;
- (H4) $|\tilde{\mathbf{h}}^{(j)}(x) - \mathbf{h}^{(j)}(x)| \le 1$ for all $x \in \mathbb{R}^d$.

Our approximating neural network $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\mathbf{g}(x) := \frac{1}{k}\sum_{j=1}^{k} \tilde{\mathbf{h}}^{(j)}(x).$$

By construction and using properties H1 and H2 (above), the following properties of $\mathbf{g}$ are immediate:

- (G1) $\mathbf{g}$ is represented by a neural network of width at most $O(k\tau L/\epsilon)$;
- (G2) $\max\{\|\mathbf{g}\|_{\mathcal{R}}, \|\mathbf{g}\|_{\mathscr{V}_2}\} = O(\tau L^2 \max_{j\in[k]} \|v + \rho\mathbf{w}^{(j)}\|_2/\epsilon)$.

Note that these properties are given in terms of τ, which has yet to be specified, as well as $\max_{j\in[k]} \|v+\rho\mathbf{w}^{(j)}\|_2$, which is a random variable. So, in the remainder of the proof, we choose a particular distribution for $\mathbf{w}$ (and hence also for $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(k)}$) and a value of τ that, together, will ultimately allow us to establish the existence of an approximating neural network with the desired properties via the probabilistic method.

We first specify the probability distribution of $\mathbf{w}$ and establish some of its properties. We let $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_d)$ be a vector of independent random variables $\mathbf{w}_1, \dots, \mathbf{w}_d$ with $p_i := \mathbb{P}[\mathbf{w}_i = -2\,\mathrm{sign}\,(v_i)] = |v_i|/(2\rho)$ and $\mathbb{P}[\mathbf{w}_i = 0] = 1 - p_i$. Note that $p_i \in [0, 1/2]$ for all i since we have assumed $\rho \geq \|v\|_\infty$, so the distribution of $\mathbf{w}$ is well-defined. Furthermore, observe that $\mathbf{w} \neq -(1/\rho)v$ almost surely (since $v \neq 0$ and $\rho \geq \|v\|_\infty$ by assumption), $\mathbb{E}[v + \rho\mathbf{w}] = 0$, and

$$\mathbb{E}\left[\|v + \rho\mathbf{w}\|_2^2\right] = \sum_{i=1}^d \mathrm{Var}(\rho\mathbf{w}_i) = \sum_{i=1}^d 4\rho^2 \cdot \frac{|v_i|}{2\rho} \cdot \left(1 - \frac{|v_i|}{2\rho}\right) = 2\rho\,\|v\|_1 - \|v\|_2^2 = \sigma_{\rho,v}^2.$$

Moreover, $\|v + \rho\mathbf{w}\|_2$ is a function of independent random variables $\mathbf{w}_1, \dots, \mathbf{w}_d$ that satisfies the $(2|v_1|, \dots, 2|v_d|)$-bounded differences property. By McDiarmid's inequality (Lemma 20) and Jensen's inequality,

$$\mathbb{P}\left[\|v + \rho\mathbf{w}\|_2 \geq \sigma_{\rho,v} + \sqrt{2\ln(1/\delta)}\right] \leq \delta \quad \text{for all } \delta \in (0,1). \tag{2}$$

Finally, for any fixed $x \in \{\pm 1\}^d$, $x^\mathsf{T}(v + \rho\mathbf{w}) = \sum_{i=1}^d x_i(v_i + \rho\mathbf{w}_i)$ is a sum of d independent, mean-zero random variables, with variance $\mathrm{Var}\,(x^\mathsf{T}(v + \rho\mathbf{w})) = \sigma_{\rho,v}^2$ and $|x_i(v_i + \rho\mathbf{w}_i)| \leq 2\rho$ almost surely for each i. By Bernstein's inequality (Lemma 19),

$$\mathbb{P}\left[|x^\mathsf{T}(v + \rho\mathbf{w})| \geq \sigma_{\rho,v}\sqrt{2\ln(2/\delta)} + 2\rho\ln(2/\delta)/3\right] \leq \delta \quad \text{for all } x \in \{\pm 1\}^d,\, \delta \in (0,1). \tag{3}$$

We now show that $\mathbf{g}$ has the desired properties with positive probability. Since $w^\mathsf{T}x$ is an integer for any $w \in \mathbb{Z}^d$ and $x \in \{\pm 1\}^d$, and since $v + \rho\mathbf{w}^{(j)} \neq 0$ almost surely, the ρ-periodicity of ϕ implies that $g_{\mathbf{w}^{(j)}}(x) = f(x)$ for all $x \in \{\pm 1\}^d$ and all $j \in [k]$. Therefore, the intermediate (random) function $\mathbf{g}_1 \colon \mathbb{R}^d \to [-1, 1]$ defined by $\mathbf{g}_1(x) := \frac{1}{k}\sum_{j=1}^k \mathbf{h}^{(j)}(x)$ satisfies $\mathbf{g}_1(x) = f(x)$ for all $x \in \{\pm 1\}^d$. For each $x \in \{\pm 1\}^d$, let

$$\mathbf{r}(x) := |\{j \in [k] : |x^\mathsf{T}(v + \rho\mathbf{w}^{(j)})| \geq \tau\}| = |\{j \in [k] : |x^\mathsf{T}\mathbf{u}^{(j)}| \geq \tau/\|v + \rho\mathbf{w}^{(j)}\|_2\}|.$$

Using the approximation properties of $\tilde{\mathbf{h}}^{(j)}$ (i.e., H3 and H4 from above), we have for each $x \in \{\pm 1\}^d$,

$$|\mathbf{g}(x) - \mathbf{g}_1(x)| = \frac{1}{k}\left|\sum_{j=1}^k (\tilde{\mathbf{h}}^{(j)}(x) - \mathbf{h}^{(j)}(x))\right| \leq \left(1 - \frac{\mathbf{r}(x)}{k}\right) \cdot \frac{\epsilon}{3} + \frac{\mathbf{r}(x)}{k} \cdot 1.$$

This final expression is at most ϵ if $\mathbf{r}(x) \leq 2k\epsilon/3$. We choose τ such that for any $x \in \{\pm 1\}^d$, we have $\mathbb{P}[|x^\mathsf{T}(v + \rho\mathbf{w})| > \tau] \leq \epsilon/3$. By (3), it suffices to choose

$$\tau := \sigma_{\rho,v}\sqrt{2\ln(6/\epsilon)} + 2\rho\ln(6/\epsilon)/3.$$

By a multiplicative Chernoff bound (Lemma 18) and a union bound over all $x \in \{\pm 1\}^d$,

$$\mathbb{P}\left[\exists x \in \{\pm 1\}^d \text{ s.t. } \mathbf{r}(x) > 2k\epsilon/3\right] \leq 2^d \cdot e^{-k\epsilon/9} < \frac{1}{2},$$

where the final inequality uses the choice of $k > (9(d+1)\ln 2)/\epsilon$. Therefore, with probability more than $1/2$, we have $\mathbf{r}(x) \leq 2k\epsilon/3$ for all $x \in \{\pm 1\}^d$, and hence

$$|\mathbf{g}(x) - f(x)| = |\mathbf{g}(x) - \mathbf{g}_1(x)| \leq \epsilon \quad \text{for all } x \in \{\pm 1\}^d. \tag{4}$$

Finally, by (2) an a union bound over all $j \in [k]$, we have that with probability more than $1/2$,

$$\max_{j \in [k]} \|v + \rho \mathbf{w}^{(j)}\|_2 \leq \sigma_{\rho,v} + \sqrt{2\ln(2k)}. \tag{5}$$

So, there is a positive probability that both (4) and (5) hold simultaneously, and in this event, it can be checked (via G1 and G2 above) that the function $\mathbf{g}$ satisfies the desired properties in the theorem. ■

Lemma 38 *Suppose $f(x) = \phi(v^\top x)$ is an L-Lipschitz function for $v \in \mathbb{S}^{d-1}$, $\phi : \mathbb{R} \to [-1,1]$, and $L \geq 1$. For any $t \in [1, \sqrt{d} - 1]$ and $\delta \in (0,1)$, there exists a neural network g of width $O(\frac{tL}{\delta})$ such that:*

1. *$\|g\|_{\mathcal{R}} = O(\frac{tL^2}{\delta})$;*
2. *$|f(x) - g(x)| \leq \delta$ for all x with $|v^\top x| \leq t$;*
3. *$|f(x) - g(x)| \leq 1$ for all $x \in \mathbb{R}^d$;*
4. *$g(x) = 0$ for all x with $|v^\top x| \geq t + \frac{1}{L}$; and*
5. *g is a ridge function that in direction v.*

Proof. We first introduce an L-Lipschitz function ϕ_t (visualized in Figure 1) that perfectly fits ϕ on the interval $[-t, t]$ and is zero in $(\infty, -t - \frac{1}{L}] \cup [t + \frac{1}{L}, \infty]$:

$$\phi_t(z) := \begin{cases} \phi(z) & \text{if } z \in [-t, t]; \\ \operatorname{sign}(\phi(t)) \max\{|\phi(t)| - L(z - t)), 0\} & \text{if } z \geq t; \\ \operatorname{sign}(\phi(-t)) \max\{|\phi(-t)| - L(-z + t)), 0\} & \text{if } z \leq -t. \end{cases}$$

Then, there exists a piecewise-linear function ψ_t that

- point-wise approximates ϕ_t to accuracy δ;
- has $\psi_t(z) = \phi_t(z)$ for all $z \notin [-t, t]$;
- has $\frac{2tL}{\delta}$ evenly-spaced knots on the interval $[-t, t]$ where ψ_t exactly fits ϕ_t; and
- is L-Lipschitz.

As a result ψ_t can be written as a neural network with $\psi_t(z) = \sum_{j=1}^m a^{(j)} \varphi_{\mathrm{r}} z - b^{(j)}$ where $m = \frac{2Lt}{\delta}$, $b^{(j)} \in [-t - \frac{1}{L}, t + \frac{1}{L}]$, and $\left|a^{(j)}\right| \leq 2L$.

By taking $g(x) := \psi_t(v^\top x)$, we have a neural network that satisfies conditions 2, 3, 4, and 5. The bound on $\|g\|_{\mathcal{R}}$ is immediate from the fact that g can be expressed as a neural network with $O(\frac{tL}{\delta})$ neurons with unit weights, biases in $[-\sqrt{d}, \sqrt{d}]$, and bounded coefficients $a^{(j)}$. ■

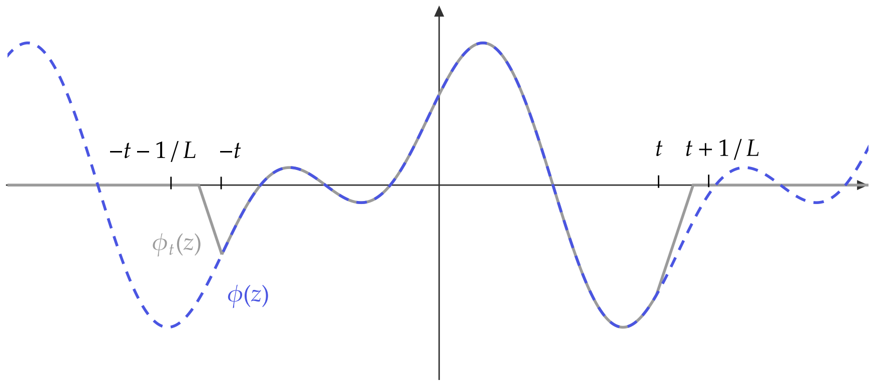

Figure 1: A visualization of how the truncated ϕ_t (gray) is generated from ϕ (blue), t, and L.

F.2. Proof of Theorem 14

Theorem 14 *Assume d is even. Let Ridge_d be the set of functions $g\colon \Omega \to \mathbb{R}$ such that $g(x) = \phi(w^\intercal x)$ for some $w \in \mathbb{S}^{d-1}$ and Lipschitz continuous $\phi\colon [-\sqrt{d}, \sqrt{d}] \to \mathbb{R}$. Let $\rho := 4q/\sqrt{d}$ for $q \in \{1, 2, \dots, \lfloor \sqrt{d}/4 \rfloor\}$ and $f(x) := \cos((2\pi/(\rho\sqrt{d}))\mathbf{1}^\intercal x)$. Then*

$$\inf\{\|g\|_{\mathcal{R}} : g \in \mathrm{Ridge}_d,\ \|g - f\|_{L^\infty(\nu)} \leq 1/2\} = \Omega(\sqrt{d}/\rho^2).$$

Proof. We prove the claim by a reduction to Theorem 4. That is, we show that an interpolant with better $\mathcal{R}$-norm than the bound stipulates can be used to construct a neural network that contradicts Theorem 4.

To do so, we consider a lower dimension $d' = 4\lfloor d/4q \rfloor - 4$ and create a mapping from points $z \in \{\pm 1\}^{d'}$ to $x_z \in \{\pm 1\}^d$. We define $a \in [0, 4q-1]$ such that $2a \equiv d \pmod{4q}$. For any z, we define x_z as follows:

$$x_z = (\underbrace{z_1, \dots, z_1}_{q}, \dots, \underbrace{z_{d'}, \dots, z_{d'}}_{q}, \underbrace{1, \dots, 1}_{a}, \underbrace{-1, \dots, -1}_{d-d'q-a}).$$

Observe that

$$\mathbf{1}^\intercal x_z = q\mathbf{1}^\intercal z + 2a - d + d'q \equiv q\mathbf{1}^\intercal z \pmod{4q}.$$

Due to the periodicity of cosine and the fact that d' is a multiple of 4,

$$\cos(\frac{2\pi}{\rho} v^\intercal x_z) = \cos(\frac{\pi}{2}\mathbf{1}^\intercal z) = \chi(z).$$

Consider some $g(x) = \phi(w^\intercal x)$ with $\|g - \cos(\frac{2\pi}{\rho} v^\intercal \cdot)\|_\infty \leq \frac{1}{2}$. Define $w' \in \mathbb{R}^{d'}$ such that $w'_i := \sum_{j=1}^q w_{(i-1)q+j}$. Observe that $\|w'\|_2 \leq \sqrt{q}$ and that $w^\intercal x_z = w'^\intercal z + c_w$, where c_w depends only on the remaining elements of w. Define $\tilde{g}(z) = \phi(w'^\intercal z + c_w)$. Then,

$$|\tilde{g}(z) - \chi(z)| = |\phi(w^\intercal x_z) - \cos(\frac{2\pi}{\rho} v^\intercal x_z)| \leq \frac{1}{2}$$

for all $z \in \{\pm 1\}^{d'}$. Since translation can only decrease the $\mathcal{R}$-norm (by exhausting some neurons to effectively behave linearly in the domain) namely, $\|\tilde{g}\|_{\mathcal{R}} \leq \|w'\|_2 \|\phi'\|_{\mathrm{TV}} = \|w'\|_2 \|g\|_{\mathcal{R}}$, Theorem 4 implies that $\|g\|_{\mathcal{R}} = \Omega(d'^{3/2}/\sqrt{q})$. The theorem statement follows by plugging in q and d'. ■