Learning Deep ReLU Networks Is Fixed-Parameter Tractable

Sitan Chen*
sitanc@mit.edu
MIT

Adam R. Klivans†
klivans@cs.utexas.edu
UT-Austin and IAS

Raghu Meka‡
raghum@cs.ucla.edu
UCLA

September 29, 2020

Abstract

We consider the problem of learning an unknown ReLU network with respect to Gaussian inputs and obtain the first nontrivial results for networks of depth more than two. We give an algorithm whose running time is a fixed polynomial in the ambient dimension and some (exponentially large) function of only the network's parameters.

Our bounds depend on the number of hidden units, depth, spectral norm of the weight matrices, and Lipschitz constant of the overall network (we show that some dependence on the Lipschitz constant is necessary). We also give a bound that is doubly exponential in the size of the network but is independent of spectral norm. These results provably cannot be obtained using gradient-based methods and give the first example of a class of efficiently learnable neural networks that gradient descent will fail to learn.

In contrast, prior work for learning networks of depth three or higher requires *exponential* time in the ambient dimension, even when the above parameters are bounded by a constant. Additionally, all prior work for the depth-two case requires well-conditioned weights and/or positive coefficients to obtain efficient run-times. Our algorithm does not require these assumptions.

Our main technical tool is a type of filtered PCA that can be used to iteratively recover an approximate basis for the subspace spanned by the hidden units in the first layer. Our analysis leverages new structural results on lattice polynomials from tropical geometry.

1 Introduction

We study the problem of learning the following class of concepts:

Definition 1.1 (ReLU Networks)**.** Let $\mathcal{C}_S$ denote the concept class of (feedforward) ReLU networks over $\mathbb{R}^d$ of size S. Specifically, $F \in \mathcal{C}_S$ if there exist weight matrices $\mathbf{W}_0 \in \mathbb{R}^{k_0 \times d}, \mathbf{W}_1 \in \mathbb{R}^{k_1 \times k_0}, \ldots, \mathbf{W}_L \in \mathbb{R}^{k_L \times k_{L-1}}, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L}$ for which

$$F(x) \triangleq \mathbf{W}_{L+1}\phi\left(\mathbf{W}_L\phi\left(\cdots\phi(\mathbf{W}_0 x)\cdots\right)\right),$$

where $\phi(z) \triangleq \max(z, 0)$ is the ReLU activation applied entrywise, and $k_0 + \cdots + k_L = S$. In this case we say that F is computed by a ReLU network with depth $L+2$. We will refer to the rank of $\mathbf{W}_0$ as k, to emphasize that the value of F only depends on a k-dimensional subspace of $\mathbb{R}^d$. We will also let $k_{L+1} = 1$.

*This work was supported in part by a Paul and Daisy Soros Fellowship, NSF CAREER Award CCF-1453261, and NSF Large CCF-1565235.

†Supported by NSF awards AF-1909204, AF-1717896, and the NSF AI Institute for Foundations of Machine Learning (IFML). Work done while visiting the Institute for Advanced Study, Princeton, NJ.

‡Supported by NSF CAREER Award CCF-1553605.

When the weight matrices of two ReLU networks $F, F' \in \mathcal{C}_S$ have the same dimensions (at all layers), then we say that F and F' *have the same architecture*.

For example, a depth two ReLU network of size S in d-dimensions is a function $F : \mathbb{R}^d \to \mathbb{R}$ of the form

$$F(x) = \sum_{i=1}^{S} \lambda_i \phi(\langle w_i, x \rangle),$$

where $\lambda_i \in \mathbb{R}$ are scalars and $w_i \in \mathbb{R}^d$ are arbitrary vectors.

Note that any Boolean function $F : \{\pm 1\}^n \to \{\pm 1\}$ can be computed by an n-layer ReLU network (see Lemma A.2 in Appendix A.2). In particular, if F is a junta depending only on k variables, then it can be computed by a k-layer ReLU network with size that depends only on k.

Learning ReLU Networks The problem of PAC learning an unknown ReLU network from labeled examples is a central challenge in the theory of machine learning. Given samples from a distribution of the form $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ where $y = F(x)$ with F an unknown size-S ReLU network, and x is drawn according to a distribution $\mathcal{D}$, the goal is to output a function $f : \mathbb{R}^d \to \mathbb{R}$ with small *test* error, i.e., $\mathbb{E}_{x,y}[(y - f(x))^2] \leq \varepsilon \, \mathbb{E}[y^2]$. In this work, we focus on the widely studied case where the input distribution on x is Gaussian.

Ideally, we would like an algorithm with sample complexity and running time that is polynomial in all the relevant parameters. As a first step, the algorithm should depend polynomially on the *dimension* (it is often easy to obtain brute-force search algorithms that run in time exponential in the dimension[1]). Even this goal, however, has been elusive: it is not known how to achieve subexponential-time algorithms for general depth two ReLU networks (without making additional assumptions on the network).

In this work, we give the first algorithm for learning ReLU networks whose running time is a fixed polynomial in the dimension, regardless of the depth of the network. Our algorithm is *fixed-parameter tractable*: we show that we can *properly learn* (i.e., the output hypothesis is also a ReLU network) ReLU networks with sample complexity and running time that is a fixed polynomial in the dimension and an exponential function of the network's *parameters*.

More precisely, our main result is as follows. We will also make the (as it turns out necessary) assumption that the ReLU network has a bounded Lipschitz constant: a function $f : \mathbb{R}^d \to \mathbb{R}$ is Λ-Lipschitz if $|f(x) - f(x')| \leq \Lambda \|x - x'\|_2$ for all x, x'.

Theorem 1.2 (Main, see Theorem 5.2 for formal statement)**.** *Let $\mathcal{D}$ be the distribution over pairs $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ where $x \sim \mathcal{N}(0, \mathrm{Id})$ and $y = F(x)$ for a size-S ReLU network F with depth $L + 2$, Lipschitz constant at most Λ, rank of* bottom *weight matrix $\mathbf{W}_0$ being k, and whose weight matrices all have spectral norm at most B.*

There is an algorithm that draws $d \log(1/\delta) \exp\left(\mathrm{poly}(k, S, \Lambda/\varepsilon)\right) B^{O(Lk)}$ samples, runs in time $\widetilde{O}(d^2 \log(1/\delta)) \exp\left(\mathrm{poly}(k, S, \Lambda/\varepsilon)\right) B^{O(LkS^2)}$, and outputs a ReLU network $\widetilde{F}$ such that $\mathbb{E}[(y - \widetilde{F}(x))^2] \leq \varepsilon$ with probability at least $1 - \delta$.[2]

Note that the sample complexity is linear while the run-time is quadratic in the ambient dimension. In particular, in the well-studied special case where the product of the spectral norms of the weight matrices is a constant (see e.g. [GRS18]), in which case the Lipschitz constant of the network is also constant, we can obtain the following corollary:

[1] Although in our specific case even this type of search turns out to be nontrivial.

[2] See Remark 5.3 for a discussion of why this guarantee is scale-invariant.

Corollary 1.3. *Let $\mathcal{D}$ be the distribution over pairs $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ where $x \sim \mathcal{N}(0, \mathrm{Id})$ and $y = F(x)$ for a size-S ReLU network F for which the product of the spectral norms of its weight matrices is a constant.*

Then there is an algorithm that draws $N = d\log(1/\delta)\exp(O(k^3/\varepsilon^2 + kS))$ samples, runs in time $\widetilde{O}(d^2\log(1/\delta))\exp(O(k^3S^2/\varepsilon^2 + kS^3))$, and outputs a ReLU network $\widetilde{F}$ such that $\mathbb{E}[(y - \widetilde{F}(x))^2] \le \varepsilon$ with probability at least $1 - \delta$.

As mentioned earlier, no algorithms that were sub-exponential in d were known even for S, B, ε being constants.

Before going further, we note that a dependence on the Lipschitz constant of the network is necessary even for learning depth two ReLU networks with respect to Gaussians:

Example 1.4. *Let $\Lambda > 0$. Consider the size-3, depth two ReLU network $F : \mathbb{R}^2 \to \mathbb{R}$ given by*

$$F(x_1, x_2) = \phi(x_1 + \Lambda x_2) + \phi(3x_1 + \Lambda x_2) - 2\phi(-x_1 + \Lambda x_2).$$

The Lipschitz constant of F is $\Theta(\Lambda)$: $F(0, 1/\Lambda) = 1$ and $F(1, 1/\Lambda) = 2$. Furthermore, note that for $(x_1, x_2) \in \mathbb{S}^1$, $F(x_1, x_2) = 0$ unless $x_2 \in [-3/\Lambda, 3/\Lambda]$. By rotational symmetry, for $(x_1, x_2) \sim \mathcal{N}(0, \mathrm{Id})$, $F(x_1, x_2) \neq 0$ with probability at most $O(1/\Lambda)$.

Note that for depth two ReLU networks with positive weights, no such dependence on the Lipschitz constant is necessary intuitively because without cancellations between the hidden units, one cannot devise "spiky" functions F which simultaneously have small variance but attain a large value at some bounded-norm x.

Interestingly, our techniques are also general enough to handle general continuous piecewise-linear functions (see Definition 4.2 for a formal definition):

Theorem 1.5 (See Theorem 5.1 for formal statement)**.** *Let $\mathcal{D}$ be the distribution over pairs $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ where $x \sim \mathcal{N}(0, \mathrm{Id})$ and $y = F(x)$ for a continuous piecewise-linear function F which only depends on the projection of x to a k-dimensional subspace V, has at most M linear pieces, and is Λ-Lipschitz.*

There is an algorithm that draws $d\log(1/\delta) \cdot \mathrm{poly}\left(\exp\left(k^3\Lambda^2/\varepsilon^2\right), M^k\right)$ samples, runs in time $\widetilde{O}(d^2\log(1/\delta)) \cdot M^{M^2} \cdot \mathrm{poly}\left(\exp\left(k^4\Lambda^2/\varepsilon^2\right), M^{k^2}\right)$, and outputs a piecewise-linear function $\widetilde{F}$ such that $\mathbb{E}[(y - \widetilde{F}(x))^2] \le \varepsilon$ with probability at least $1 - \delta$.

Note that a size-S ReLU network is a continuous piecewise-linear function with at most 2^S linear pieces. Specializing Theorem 1.5 to ReLU networks gives a guarantee which is incomparable to Theorem 1.2: we obtain an algorithm that depends doubly exponentially on S but has no dependence on the norms of the weight matrices.

1.1 Prior Work on Provably Learning Neural Networks

Algorithmic Results Algorithms for learning neural networks (obtaining small *test error*) have been intensely studied in the literature. In the last few years alone there have been many papers giving provable results for learning restricted classes of neural networks under various settings [JSA15, ZLJ16, ZSJ+17, BG17, GKKT17, LY17, ZPS17, Tia17, GKM18, DLT18, GLM18, GKLW18, MR18, BJW19, GK19, AZLL19, VW19, ZYWG19, DGK+20, GMOV18, LMZ20].

The predominant techniques are spectral or tensor-based dimension reduction [JSA15, ZSJ+17, BJW19, DKKZ20], kernel methods [ZLJ16, GKKT17, Dan17, MR18, GK19], and gradient-based methods [GLM18, GKLW18, VW19]. All prior work takes distributional and/or architectural

assumptions, the most common one being that the inputs come from a standard Gaussian. We will also work in this setting.[3]

As pointed out in [GGJ+20, DGK+20], all existing algorithmic results for Gaussian inputs hold *only for depth two networks* and make at least one of two assumptions on the unknown network F in question:

Assumption (1) Weight matrix $\mathbf{W}_0$ is well-conditioned and, in particular, full rank.

Assumption (2) The vector at the output layer ($\mathbf{W}_1$ when $L = 0$) has all positive entries.

Assumption (1) allows one to use tensor decomposition to recover the parameters of the network and hence PAC learn, an idea that has inspired a long line of works [JSA15, ZSJ+17, GLM18, GKLW18, BJW19]. However, the assumption is not necessary for PAC learning or achieving low-prediction error. For instance, consider a pathological case where $\mathbf{W}_0$ has repeated rows. Here, while parameter recovery is not possible it is still possible to PAC learn. To our knowledge, the only work that can PAC learn depth two networks over Gaussian inputs without a condition number bound on $\mathbf{W}_0$ is [DKKZ20]. However, their work still requires assumption (2) (and only holds for depth two networks). Our work shows that assumption (2) is neither information-theoretically nor computationally necessary.

Limitations of Gradient-Based Methods Two recent works [GGJ+20, DKKZ20] showed that a broad family of algorithms, namely *correlational statistical query (CSQ) algorithms*, fail to PAC learn even depth two ReLU networks; that is, functions of the form $F(x) = \sum_{i=1}^{k} \lambda_i \phi(\langle v_i, x \rangle)$ with respect to Gaussian inputs in time polynomial in d where d is the ambient dimension (in fact, [DKKZ20] rules out running time $d^{o(k)}$). Informally, a CSQ algorithm is limited to using noisy estimates of statistics of the form $\mathbb{E}[y \cdot \sigma(x)]$ for arbitrary bounded σ, where the expectation is over examples (x, y) and $y = F(x)$ is computed by the network. The point is that this already rules out a wide range of algorithmic approaches in theory and practice, including gradient descent on overparameterized networks (i.e., using neural tangent kernels [JGH18] or the mean-field approximation for gradient dynamics [MMN18]). Note that the algorithms of [DKKZ20] for learning depth two ReLU networks with positive coefficients are CSQ algorithms as well.

Note that as a consequence of Theorem 1.2, for any ε a function of k, our algorithm can learn the lower bound instances in [GGJ+20, DKKZ20] to error ε in time $g(k) \cdot \mathrm{poly}(d)$ for some g (note that the norm bounds and Lipschitz constants for these instances are upper bounded by functions of k), which is impossible for any CSQ algorithm. We explain why our algorithm is not a CSQ algorithm in Section 2.

For the classification version of this problem (i.e., taking a softmax) where we observe $Y \in \{0, 1\}$ such that $\mathbb{E}[Y|X] = \sigma(f(X))$ where σ is say sigmoid and $f(X)$ is a depth two ReLU network, Goel et al. [GGJ+20] show that even general SQ algorithms cannot achieve a runtime with polynomial dependence on the dimension. We also remark there is an extensive literature of previous work showing various hardness results for learning certain classes of neural networks [BR89, Vu06, KS09, LSSS14, GKKT17, SVWX17, SSSS17, Sha18, VW19, GKK19, DV20]. We refer the reader to [GGJ+20] for a discussion of how these prior works relate to the above CSQ lower bounds.

1.2 Other Related Work

Multi-Index Models Functions computed by ReLU networks where $\mathbf{W}_0$ has fewer rows than columns are a special case of a *multi-index model*, that is, a function $F : \mathbb{R}^d \to \mathbb{R}$ given by

[3] Other works such as [AZLL19] or kernel-based methods [ZLJ16, GKKT17] require strong norm-based assumptions on the inputs and weights.

$F(x) = f(\mathbf{W}^\top x)$ for some matrix $\mathbf{W} \in \mathbb{R}^{k\times d}$ and some function $f : \mathbb{R}^k \to \mathbb{R}$. In the theoretical computer science literature, these are sometimes referred to as *subspace juntas* [VX11, DMN19].

One of the strongest results in this line of work, and the closest in spirit to the setting we consider, is that of [DH18], which gives various conditions on f under which one can recover $\mathbf{W}$ (under Gaussian inputs) in the special case where $k = 1$, as well as a vector in the row span of $\mathbf{W}$ in the case of general k (although these results do not hold for ReLU). In general, the literature on multi-index models is vast, and we refer to [DH18] for a comprehensive overview of this body of work. Many works were inspired by a simple but powerful connection to Stein's lemma [Li92, Bri12, PV16], which was also a key ingredient in the above algorithms for learning neural networks using tensor decomposition. One technique in this literature which is somewhat similar in spirit to the techniques we employ in this work is that of *sliced inverse regression* [BB+18, Li91], and we elaborate in Remark 2.1 on this connection.

Piecewise-Linear Regression Lastly, we mention that previous works on *segmented regression* (see e.g. [ADLS16] on the references therein) study regression for piecewise-linear functions but work with a different notion of piecewise-linearity that is unrelated to our setting.

2 Proof Overview

Suppose we are given samples (x, y) where $y = F(x)$ is computed by a size S ReLU network as in Definition 1.1. Let $V \subseteq \mathbb{R}^d$ denote the span of the rows of $\mathbf{W}_0$ and let k be its dimension. We will call V the *relevant subspace*, because the value of F only depends on the projection of x to V. In particular, we can write $y = F'(\Pi_V(x))$ for some function $F' : V \to \mathbb{R}$ that is itself a size S ReLU network and Π_V denotes the projection operator onto V. The main focus of our algorithm will be in figuring out the relevant subspace V given samples (x, y). This is the hardest part of the algorithm, because once we learn the relevant subspace to high enough accuracy, we can grid-search over ReLU networks in this subspace. Even this grid search turns out to be non-trivial to analyze and entails proving new *stability* results for piecewise-linear functions.

Filtered PCA Our algorithm builds upon the *filtered PCA* approach, originally introduced in [CM20] for the purposes of learning low-degree polynomials over Gaussian space.[4] For any $\psi : \mathbb{R} \to \mathbb{R}$, let $\mathbf{M}_\psi \triangleq \mathbb{E}[\psi(Y)(XX^T - \mathrm{Id})]$. A basic but important observation is that for any choice of ψ, all vectors orthogonal to the true subspace V are in the kernel of $\mathbf{M}_\psi$. A natural idea for identifying the true subspace then is to look at the nonzero singular vectors of $\mathbf{M}_\psi$ for a suitable ψ. If we could show that $\mathbf{M}_\psi$ has k nonzero singular values all bounded away from 0 by some dimension-independent margin $c(\psi)$, then we could hope to approximately recover V by empirically estimating $\mathbf{M}_\psi$ using $O(d/c(\psi)^2)$, invoking standard matrix concentration, and computing its top-k singular subspace. So the main hurdle is to identify an appropriate ψ for which this is the case.

What should the ψ be? For instance if ψ is the identity function, then the matrix $\mathbf{M}_\psi$ could be identically zero. This is an essential difference between our setting and the setting studied in previous works [DKKZ20, GLM18] (in the $L = 0$ case) where the output layer's coefficients are all positive, for which this choice of ψ would suffice to recover the relevant subspace.

Note that this is consistent with the CSQ lower bounds of [GGJ+20, DKKZ20], as any algorithm that just tries to use the spectrum of $\mathbf{M}_\psi$ for ψ being the identity function would be a CSQ

[4] For readers familiar with the approach there, we explain in Remark 5.15 why a straightforward application of the algorithm there cannot work.

algorithm. Indeed, for any of the 'hard' functions F from those works which are ReLU networks with $L = 0$ we would have $\mathbf{M}_\psi = 0$ if ψ is the identity function.

We will choose ψ not equal to the identity, and in this way our algorithm will be non-CSQ and evade the aforementioned CSQ lower bounds.

Threshold Filter. Motivated by [CM20], our starting point in the present work is to consider ψ given by a univariate threshold, that is, $\psi(z) = \mathbb{1}[|z| > \tau]$ for suitable τ. For brevity, for $\tau \in \mathbb{R}$ define $\mathbf{M}_\tau = \mathbb{E}_{x,y}[\mathbb{1}[|y| > \tau](xx^T - \mathrm{Id})]$. Then we have that

$$\langle \Pi_V, \mathbf{M}_\tau \rangle = \mathop{\mathbb{E}}_{x,y}\Big[\mathbb{1}[|y| > \tau] \cdot (\|\Pi_V x\|^2 - k)\Big].$$

In particular, if one could choose τ for which $|F(x)| > \tau$ only if $\|\Pi_V x\|^2 \geq 2k$ [5], then we would conclude that $\langle \Pi_V, \mathbf{M}_\tau \rangle \geq k \cdot \mathbb{P}[|y| > \tau]$, so some singular value of $\mathbf{M}_\tau$ is at least $\mathbb{P}[|y| > \tau]$. If F is Λ-Lipschitz, we can simply choose τ to be $\sqrt{2k} \cdot \Lambda$, and provided $\mathbb{P}[|y| > \tau]$ is reasonably large, then we conclude that $\mathbf{M}_\tau$ has some reasonably large singular value. Finally, to lower bound $\mathbb{P}[|y| > \tau]$, we prove an *anti-concentration* result for piecewise linear functions over Gaussian space (Lemma 5.4).

In other words, if one conditions on the samples (x, y) whose responses y are sufficiently large in magnitude, then we show that the resulting distribution is noticeably non-Gaussian in some direction, and by taking the top singular vector of the conditional covariance, we can approximately recover some direction inside the relevant subspace V.[6]

Unfortunately, all that the above analysis tells us is that the trace of $\mathbf{M}_\tau$ is non-negligible which in turn helps us guarantee that we identify at least one direction in V. It is not at all clear whether the above threshold approach is enough to identify more than just one vector in the relevant subspace. Indeed, recovering the full relevant subspace turns out to be significantly more challenging, and the core technical contribution of this work is to show how to do this.

Remark 2.1 (Relation to Sliced Inverse Regression). The trick of conditioning only on (x, y) for which $|y|$ is sufficiently large is reminiscent of the technique of *slicing* originally introduced by [Li91] in the context of learning multi-index models. The high-level idea of slicing is that for any fixed value of y, the conditional law of $x|F(x) = y$ is likely to be non-Gaussian in most directions $v \in V$, so in particular, $\mathbb{E}[xx^\top - \mathrm{Id} \mid F(x) = y]$ should be nonzero, and its singular vectors will lie in V. This can be thought of as filtered PCA with the choice of function $\psi(z) = \mathbb{1}[z = y]$. The first issue with using such an approach to get an actual learning algorithm is that $\mathbb{P}_x[F(x) = y] = 0$ for any y, and the workaround in non-asymptotic analyses of sliced inverse regression [BB+18] is to estimate something like $\mathbb{E}_y[\mathbb{E}[xx^\top - \mathrm{Id} \mid F(x) = y]]$ instead. While finite sample estimators for such objects are known, the conditions under which this approach can provably recover the relevant subspace are quite strong and not applicable to our setting.

Learning the Full Subspace: What Doesn't Work One might hope that a more refined analysis shows that for a suitable τ, the spectrum of $\mathbf{M}_\tau$ can identify the entire subspace V. Given

[5] The choice of $2k$ here is for exposition; any bound noticeably more than k, e.g., $k + 1$ will do.

[6] Note that while the goal is to reweight the distribution over x to look non-Gaussian in some relevant direction, the main challenge once we've fixed a reweighting is not to identify that non-Gaussian subspace, which in our setting is trivial and does not require any of the more sophisticated techniques in the non-Gaussian component analysis literature (e.g. [Ver10, GS19]), but to argue that the new distribution is indeed non-Gaussian in some direction in V. In a similar vein, while the work [VX11] gives some moment-based conditions under which it is possible to learn multi-index models over Gaussian inputs, it seems highly nontrivial to verify whether such conditions actually hold for ReLU networks, and in addition their results seem tailored to $\{0, 1\}$-valued functions.

that we can already learn some $w \in V$ with the threshold approach above, a first step would be to try to find a direction in V orthogonal to w, by lower bounding the contribution to the Frobenius norm of $\mathbf{M}_\tau$ from vectors orthogonal to w. Concretely, letting $\Pi_{V\setminus\{w\}}$ denote the projector to the orthogonal complement of w in V, we have that

$$\langle \Pi_{V\setminus\{w\}}, \mathbf{M}_\tau \rangle = \mathop{\mathbb{E}}_{x,y}\Big[\mathbb{1}[|y| > \tau] \cdot (\|\Pi_{V\setminus\{w\}}x\|^2 - (k-1))\Big].$$

As before, if one could choose τ for which $|F(x)| > \tau$ only if $\|\Pi_{V\setminus\{w\}}x\|^2 \geq k$, and if we could lower bound $\mathbb{P}[|y| > \tau]$, then we would conclude that $\langle \Pi_{V\setminus\{w\}}, \mathbf{M}_\tau \rangle \geq \mathbb{P}[|y| > \tau]$, so $\mathbf{M}_\tau$ has some other singular vector, orthogonal to w, with non-negligible singular value. The issue is that such a τ typically does not exist! For x satisfying $\|\Pi_{V\setminus\{w\}}x\|^2 \leq k$, $F(x)$ can be arbitrarily large, because $\|\Pi_w x\|$ can be arbitrarily large.

It may be possible to lower bound the quantity in (3.17) using a more refined argument, but for general deep ReLU networks or piecewise linear functions, this seems very challenging. At the very least, one must be careful not to prove something too strong, like showing that $v^\top \mathbf{M}_\tau v$ is non-negligible for *any* unit vector $v \in V$. For instance, even when $L = 0$, it could be that all but one of the rows of $\mathbf{W}_0$ lie in a proper subspace $W \subsetneq V$, and for the remaining row u of $\mathbf{W}_0$, $\|\Pi_{V\setminus W}u\|/\|u\|$ is arbitrarily small. In this case, for v in the direction of $\Pi_{V\setminus W}u$, the quadratic form $v^\top \mathbf{M}_\tau v$ is arbitrarily small, and it would be impossible to recover all of V from a reasonable number of samples.

More generally, any proposed algorithm for learning all of V had better be consistent with the fact that it is impossible to recover the full subspace V within a reasonable number of samples if almost all of the variance of F is explained by some proper subspace $W \subsetneq V$, or equivalently, if the "leftover variance" $\mathbb{E}_x[(F(x) - F(\Pi_W x))^2]$ is negligible. We emphasize that this is a key subtlety that does not manifest in previous works that consider full-rank, well-conditioned weight matrices.

Learning the Full Subspace: Our Approach We now explain our approach. At a high level, we try to learn orthogonal directions inside the relevant subspace in an iterative fashion. The threshold filter approach above already gives us a single direction in V. Suppose inductively that we've learned some orthogonal vectors $w_1, ..., w_\ell \in V$ spanning a subspace $W \subseteq V$ and want to learn another (note that technically we can only guarantee $w_1, ..., w_\ell$ are approximately within V, but let us temporarily ignore this for the sake of exposition). Motivated by the above consideration regarding "leftover variance," we proceed by a win-win argument: either the leftover variance already satisfies $\mathbb{E}_x[(F(x) - F(\Pi_W x))^2] \leq \varepsilon$ in which case we are already done, or we can learn a new direction via the following crucial modification of the threshold filter.

First, as a thought experiment, consider the following matrix

$$\mathbf{M}_\tau^W \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y}\Big[\mathbb{1}[|y - F(\Pi_W x)| > \tau] \cdot (xx^\top - \mathrm{Id})\Big]\Pi_{W^\perp}.$$

Note the critical fact that we threshold on $y - F(\Pi_W x)$ as opposed to just on y. As before, it is not hard to show that if this matrix is nonzero, then its singular vectors with nonzero singular value must lie in $\mathbf{W}_0$ and be orthogonal to W; thus giving us a new direction in $\mathbf{W}_0$. We claim that if the leftover variance is non-negligible, then the above matrix will give us a new direction in W.

The intuition behind the above matrix is as follows. Let $V\setminus W$ denote the subspace of V orthogonal to W. We can write $F(x) = F(\Pi_V x) = F(\Pi_W x + \Pi_{V\setminus W}x)$. Now, as F is Lipschitz, we can bound $G(x) = y - F(\Pi_W x) = F(\Pi_W x + \Pi_{V\setminus W}x) - F(\Pi_W x)$ as $|G(x)| \leq \Lambda\|\Pi_{V\setminus W}x\|^2$, where Λ is the Lipschitz constant of F. In other words, $G(x)$ is bounded over x for which $\|\Pi_{V\setminus W}x\|$ is

bounded. Recall that the fact that $F(x)$ is not bounded over such x was the key obstacle to using the original threshold filter approach to learn the full subspace.

The upshot is that for a suitably large τ, the only contribution to the matrix $\mathbf{M}_\tau^W$ should be from inputs x that have large projection in $V \setminus W$. We are now in a position to adapt the analysis lower bounding $\langle \Pi_V, \mathbf{M}_\tau \rangle$ to lower bounding $\langle \Pi_{V \setminus W}, \mathbf{M}_\tau^W \rangle$. In particular, we can apply the aforementioned anti-concentration for piecewise linear functions *to the function* G and argue that, provided the leftover variance $\mathbb{E}_x[(F(x) - F(\Pi_W x))^2] = \mathbb{E}_x[G(x)^2]$ is non-negligible, the top singular vector of $\mathbf{M}_\tau^W$ will give us a new vector in $V \setminus W$.

That being said, an obvious obstacle in implementing the above is that along with not knowing the true subspace $\mathbf{W}_0$, we also don't know the true function F. This precludes us from forming the matrix $\mathbf{M}_\tau^W$ as defined above.

To get around this, we will enumerate over a sufficiently fine net of ReLU networks $\widetilde{F}$ *with relevant subspace* W, one of which will be close to the ReLU network $F(\Pi_W x)$. For each $\widetilde{F}$, we will form the matrix

$$\widetilde{\mathbf{M}}_\tau^W \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y} \Big[\mathbb{1}[|y - \widetilde{F}(\Pi_W x)| > \tau] \cdot (xx^\top - \mathrm{Id}) \Big] \Pi_{W^\perp}. \tag{1}$$

and output the top singular vector as our new direction only if it has non-negligible singular value.

Arguing soundness, i.e. that this procedure doesn't yield a "false positive" in the form of an erroneous direction lying far from V, is not too hard. However, analyzing completeness, i.e. that this procedure will find *some* new direction, is surprisingly subtle (see Lemma 5.13). Formally, we need to argue that if we have an approximation $\widetilde{F}$ to the true F (under some suitable metric), then the corresponding matrix $\widetilde{\mathbf{M}}_\tau^W$ is close to the matrix $\mathbf{M}_\tau^W$. This is further complicated by the fact that ultimately, we will only have access to a subspace W which is *approximately* in V, as every direction we find in our iterative procedure is only guaranteed to *mostly* lie within V.

Our key step in proving this is showing a new stability property of affine thresholds of piecewise linear functions and makes an intriguing connection to *lattice polynomials* in tropical geometry.

Stability of Piecewise Linear Functions Following the above discussions, to complete our analysis we need to show *stability* of affine thresholds of ReLU networks in the following sense: if $F, \tilde{F} : \mathbb{R}^d \to \mathbb{R}$ are two RELU networks that are close in some structural sense (i.e., under some parametrization), then $\mathbb{E}[\mathbb{1}[|F(x)| > \tau](xx^T - Id)] \approx \mathbb{E}[\mathbb{1}[|\tilde{F}(x)| > \tau](xx^T - Id)]$. A natural way to approach the above is to upper bound $\mathbb{P}[|F(x)| > \tau \wedge |\widetilde{F}(x)| \leq \tau]$. That is, affine thresholds of ReLU networks that are structurally close disagree with low probability.

A natural way to parametrize closeness is to require the weight matrices of the two networks $F, \tilde{F}$ to be close to each other. While such a statement is not too difficult to show for depth two networks (by a union bound over pairs of ReLUs), proving such a statement for general ReLU networks using a direct approach seems quite challenging. We instead look at proving such a statement for a more general class of functions - continuous piecewise-linear functions which allows us to do a certain kind of hybrid argument more naturally.

Concretely, we show that affine thresholds of piecewise-linear functions that are close in some appropriate structural sense disagree with low probability over Gaussian space. We will elaborate upon the notion of structural closeness we consider momentarily, but for now it is helpful to keep in mind that it specializes to L_2 distance for linear functions.

Lemma 2.2 (Informal, see Lemma 5.6)**.** *Let $F, \widetilde{F} : \mathbb{R}^d \to \mathbb{R}$ be piecewise-linear functions, both consisting of at most m linear pieces, which are "(m, η)-structurally-close" (see Definition 4.10). For any $\tau > 0$,*

$$\mathop{\mathbb{P}}_{x \sim \mathcal{N}(0, \mathrm{Id})} \Big[|F(x)| > \tau \wedge |\widetilde{F}(x)| \leq \tau \Big] \leq O(\eta m^2 / \tau). \tag{2}$$

To get a sense for this, suppose $F, \widetilde{F}$ were even close in the sense that the polyhedral regions over which F is linear are *identical* to those over which $\widetilde{F}$ is linear, and furthermore $\mathbb{E}_x[(F(x) - \widetilde{F}(x))^2]^{1/2} \leq \eta$. Then if we take for granted that Lemma 2.2 holds when $m = 1$, i.e. when $F, \widetilde{F}$ are linear (see Lemma 5.7), it is not hard to show an $O((\eta m/\tau)^c)$ upper bound in (2) under this very strong notion of closeness for some $c < 1$. Because F and $\widetilde{F}$ are L_2-close as functions, for any $t > 0$ we have that with probability $1 - O(\eta^2/t^2)$ the input $x \sim \mathcal{N}(0, \mathrm{Id})$ lies in a polyhedral region for which the corresponding linear functions for F and $\widetilde{F}$ are t-close. By the $m = 1$ case of Lemma 2.2, over any one of these at most m regions, the affine thresholds $\mathbb{1}[|F(x)| > \tau]$ and $\mathbb{1}[|F(x)| > \tau]$ disagree with probability $O(t/\tau)$. Union bounding over these regions as well as the event of probability η^2/t^2 that x does not fall in such a polyhedral region, we can upper-bound the left-hand side of (2) by $O(\eta^2/t^2 + mt/\tau)$, and by taking $t = (\eta^2\tau/m)^{1/3}$, we get a bound of $(\eta m^2/\tau)^{2/3}$.

The issues with this are twofold. First, recall the function $\widetilde{F}$ that we want to apply Lemma 5.6 to is obtained from some enumeration over a fine net of ReLU networks. As such there is no way to guarantee that the polyhedral regions defining F and $\widetilde{F}$ are exactly the same, making adapting the above argument far more difficult, especially for general ReLU networks.

Second, we stress that the *linear* scaling in $O(\eta)$ in (2.2) is essential. If one suffered any polynomial loss in this bound as in the above argument, then upon applying Lemma 2.2 k times over the course of our iterative algorithm for recovering V, we would incur time and sample complexity *doubly exponential* in k. The reason is as follows.

Recall that in the final argument we can only ensure that the directions $w_1, \ldots, w_\ell$ we have found so far are *approximately* within V, and the parameter η will end up scaling with an appropriate notion of subspace distance between W and the true space V. On the other hand, the bound we can show on how far $\widetilde{M}_\tau^W$ deviates from M_τ^W in spectral norm will essentially scale with the right-hand side of (2.2). So if we could only ensure $\widetilde{M}_\tau^W$ and M_τ^W are $O(\eta^c)$-close in spectral norm for $c < 1$, then if we append the top eigenvector of $\widetilde{M}_\tau^W$ to the list of directions $w_1, ..., w_\ell$ we have found so far, the resulting span will only be $O(\eta^c)$-close in subspace distance. Iterating, we would conclude that for the final output of the algorithm to be sufficiently accurate, we would need the error incurred by the very first direction w_1 found to be doubly exponentially small in k!

Lattice Polynomials It turns out that there is a clean workaround to both issues: passing to the *lattice polynomial* representation for piecewise-linear functions. Specifically, we exploit the following powerful tool:

Theorem 2.3 ([Ovc02], Theorem 4.1; see Theorem 4.9 below)**.** *If F is continuous piecewise-linear, there exist linear functions $\{g_i\}_{i \in [M]}$ and subsets $\mathcal{I}_1, ..., \mathcal{I}_m \subseteq [M]$ for which*

$$F(x) = \max_{j \in [m]} \min_{i \in \mathcal{I}_j} g_i(x). \tag{3}$$

In fact, our notion of "structural closeness" will be built around this structural result. Roughly speaking, we say two piecewise linear functions are structurally close if they have lattice polynomial representations of the form (3) with the same set of clauses and whose corresponding linear functions are pairwise close in L_2 (see Definition 4.10).

At a high level, Theorem 2.3 will then allow us to implement a hybrid argument in the proof of Lemma 2.2 and carefully track how the affine threshold computed by a piecewise-linear function changes as we interpolate between F and $\widetilde{F}$. In this way, we end up with the desired linear dependence on η in (2.2).

With Lemma 2.2 in hand, we can argue that even with only access to a subspace W approximately within V and with only a function $\widetilde{F}$ that approximates $F(\Pi_W x)$, the top singular vector of (1) mostly lies within V, and we can make progress.

Finally, we remark that as an added bonus, Theorem 4.9 also gives us a way to enumerate over general continuous piecewise-linear functions! In this way, we can adapt our algorithm for learning ReLU networks to learning arbitrary piecewise-linear functions, with some additional computational overhead (see Theorem 5.1).

Enumerating Over Piecewise-Linear Functions and ReLU Networks There is in fact one more subtlety to implementing the above approach for ReLU networks and getting singly exponential dependence on k.

First note that whereas one can always enumerate over functions computed by lattice polynomials of the form (3) in time $\exp(\mathrm{poly}(M))$ (see Lemma 4.14), for ReLU networks of size S this can be as large as doubly exponential in S. Instead, we enumerate over ReLU networks in the naive way, that is, enumerating over the $\exp(O(S))$ many possible architectures and netting over weight matrices with respect to spectral norm, giving us only singly exponential dependence on S.

Here is the subtlety. Obviously two ReLU networks with the same architecture and whose weight matrices are pairwise close in spectral norm will be close in L_2. But how do we ensure that the corresponding lattice polynomials guaranteed by Theorem 2.3 are structurally close? In particular, getting anything quantitative would be a nightmare if the clause structure of these lattice polynomials depended in some sophisticated, possibly discontinuous fashion on the precise entries of the weight matrices.

Our workaround is to open up the black box of Theorem 2.3 and give a proof for the special case of ReLU networks from scratch. In doing so, we will find out that there are lattice polynomial representations for ReLU networks which only depend on the architecture and the *signs* of the entries of the weight matrices (see Theorem 4.15). In this way, we can guarantee that a moderately fine net will contain a network which is structurally close to the true network.

3 Technical Preliminaries

In this section we collect notation and technical tools that will be useful in the sequel.

3.1 Miscellaneous Notation and Definitions

We will use $\|\cdot\|_p$ to denote the L_p norm of a vector or of a random variable. When the random variable is given by a function over Gaussian space, e.g. $F(x)$ for $x \sim \mathcal{N}(0, \mathrm{Id})$ and $F : \mathbb{R}^d \to \mathbb{R}$, we use the short-hand $\|F\|_p$ to denote $\mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{Id})}[F(x)^p]^{1/p}$. When $p = 2$, we will omit the subscript. We use $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ to denote operator and Frobenius norms respectively. When we refer to a function as Λ-Lipschitz, unless stated otherwise we mean with respect to L_2.

Given a subspace $V \subset \mathbb{R}^d$, let Π_V denote the orthogonal projector to that subspace. Let $\mathbb{S}_V \subset \mathbb{R}^d$ denote the set of vectors in V of unit norm. When the ambient space $\mathbb{R}^d$ is clear from context, we let $V^\perp$ denote the orthogonal complement of V. For a subspace $W \subseteq V$, we will denote the orthogonal complement of W inside V by $V \backslash W$.

Given $x \in \mathbb{R}$, let $\mathcal{N}(0, 1, x)$ denote the standard Gaussian density's value at x. Let $\mathrm{erfc}(z) \triangleq \mathbb{P}_{g\sim\mathcal{N}(0,1)}[|g| > z]$ (note that we eschew the usual normalization). Let χ^2_m denote the chi-squared distribution with m degrees of freedom.

Recall that we denote the ReLU activation function by $\phi(z) \triangleq \max(z, 0)$. Additionally, for $\eta > 0$, let $\mathsf{clip}_\eta : \mathbb{R} \to \mathbb{R}$ denote the function given by

$$\mathsf{clip}_\eta(z) = \begin{cases} z & \text{if } |z| \le \eta \\ 0 & \text{otherwise} \end{cases}$$

Overloading notation, given a vector $v \in \mathbb{R}^m$, we will use $\mathsf{clip}_\eta(v)$ to refer to the vector in $\mathbb{R}^m$ obtained by applying clip_η entrywise.

We will use the following basic property of the clipping operation:

Fact 3.1. *Suppose $v, v' \in \mathbb{R}^m$ satisfy $\|v - v'\|_\infty \le \eta$, and define $v'' \triangleq \mathsf{clip}_\eta(v')$. Then for any $i \in [m]$, $v_i v_i'' \ge 0$.*

Proof. If $v_i'' > 0$, then $v_i'' = v_i' > \eta$ and by triangle inequality, $v_i > 0$. Similarly, if $v_i'' < 0$, then $v_i'' = v_i' < -\eta$ and by triangle inequality, $v_i < 0$. □

Lastly, we will use $\vee$ and $\wedge$ to denote max and min respectively. The following class of functions will be useful for us.

Definition 3.2. The set of *lattice polynomials over the reals* is the set of real-valued functions defined inductively as follows: for any $d \ge 1$, any constant real-valued function $\mathbb{R}^d \to \mathbb{R}$ is a lattice polynomial, and any function $h : \mathbb{R}^d \to \mathbb{R}$ which can be written as $h(x) = f(x) \vee g(x)$ or $h(x) = f(x) \wedge g(x)$ for two lattice polynomials $f, g : \mathbb{R}^d \to \mathbb{R}$ is also a lattice polynomial.

3.2 Concentration and Anti-Concentration

Fact 3.3 (Elementary anticoncentration)**.** *If Z is a random variable for which $|Z| \le M$ almost surely, and $\mathbb{E}[Z^2] \ge \sigma^2$, then $\mathbb{P}[|Z| \ge t] \ge \frac{1}{M^2}(\sigma^2 - t^2)$.*

Proof. We have

$$\begin{aligned} \sigma^2 \le \mathbb{E}\big[Z^2\big] &= \mathbb{E}\big[Z^2 \mid |Z| \ge t\big] \cdot \mathbb{P}[|Z| \ge t] + \mathbb{E}\big[Z^2 \mid |Z| < t\big] \cdot \mathbb{P}[|Z| < t] \\ &\le M^2 \cdot \mathbb{P}[|Z| \ge t] + t^2, \end{aligned}$$

from which the claimed bound follows upon rearranging. □

Fact 3.4. *For any integer $m \ge 1$ and $t \ge 0$, $\operatorname{erfc}(z) \ge \sqrt{2/\pi} \cdot \frac{t \cdot e^{-t^2/2}}{t^2+1}$.*

Fact 3.5. *The function $f : \mathbb{R}_{\ge 0} \to \mathbb{R}$ given by $f(z) = \operatorname{erfc}(1/\sqrt{z}) \cdot z$ is convex over $\mathbb{R}_{\ge 0}$.*

Proof. We can explicitly compute

$$f''(z) = \frac{e^{-1/2z}(1+z)}{2z^{5/2}\sqrt{2\pi}},$$

which is clearly nonnegative for any $z \ge 0$. □

Lemma 3.6 ([Ver10])**.** *Let $f : \mathbb{R} \to [0, 1]$ be any function. Let $\mathbf{M} = \mathbb{E}_{x \sim \mathcal{N}(0, \mathrm{Id}_d)}[f(x) \cdot (xx^\top - \mathrm{Id})]$. For any $\varepsilon, \delta > 0$, if $x_1, ..., x_N \sim \mathcal{N}(0, \mathrm{Id}_d)$ for $N = \Omega\left(\frac{1}{\varepsilon^2}(d + \log 1/\delta)\right)$, then*

$$\mathbb{P}\left[\left\|\mathbf{M} - \frac{1}{N}\sum_i f(x_i) \cdot (x_i x_i^\top - \mathrm{Id})\right\|_{op} \ge \varepsilon\right] \le \delta.$$

Proof. This follows from standard sub-Gaussian concentration; see e.g. Remark 5.40 in [Ver10]. ☐

Fact 3.7 (Sub-exponential tail bounds, see e.g. [Ver10], Proposition 5.16)**.** *If $X_1, ..., X_N$ are i.i.d. random variables with mean zero and sub-exponential norm*[7] *K, then*

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} X_i\right| \geq t\right] \leq 2\exp\left(-\Omega\left(\frac{Nt^2}{K^2} \wedge \frac{Nt}{K}\right)\right). \tag{4}$$

In particular, for any $\delta > 0$, if we take $N = \Theta\left(\frac{K^2}{t^2} \vee \frac{K}{t}\right) \cdot \log 1/\delta$, then (4) is at most δ.

Fact 3.8 (e.g. [Ver18], Corollary 4.2.13)**.** *For any $\varepsilon > 0$, there is an ε-net (in L_2 norm) of size $(1 + 2/\varepsilon)^m$ for the unit L_2 ball in m dimensions.*

Corollary 3.9. *For any $\varepsilon, \beta > 0$, there is an ε-net (in operator norm) for the set of $m_1 \times m_2$ matrices of operator norm at most β of size at most $(1 + 2\beta/\varepsilon)^{m_1 m_2}$.*

Proof. As operator norm is upper bounded by Frobenius norm, an ε-net in Frobenius norm for the set of $m_1 \times m_2$ matrices of Frobenius norm at most β would contain the claimed ε-net. The former can be obtained from scaling an ε/β-net in Frobenius norm for the set of $m_1 \times m_2$ matrices of unit Frobenius norm, and such a net with size $(1 + 2\beta/\varepsilon)^{m_1 m_2}$ exists by Fact 3.8. ☐

3.3 Power Method, Subspace Distances, and Perturbation Bounds

Fact 3.10 (Power method, see [RST09])**.** *Let $\mathbf{M} \in \mathbb{R}^{d\times d}$, let $k \leq d$ be a non-negative integer, and let $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_d$ be the nonzero singular values of $\mathbf{M}$. For any $k = 1, \ldots, d-1$, let $\mathsf{gap}_k = \sigma_k/\sigma_{k+1}$. Suppose there is a* matrix-vector oracle *which runs in time R, and which, given $v \in \mathbb{R}^d$, outputs $\mathbf{M}v$. Then, for any $\eta, \delta > 0$, there is an algorithm* ApproxBlockSVD$(\mathbf{M}, \eta, \delta)$ *which runs in time $\widetilde{O}(kR \log \frac{1}{\eta\cdot\delta\cdot\mathsf{gap}_k})$, and with probability at least $1-\delta$ outputs a matrix $\mathbf{U} \in R^{d\times k}$ with orthonormal columns so that $\|\mathbf{U} - \mathbf{U}_k\|_2 < \eta$, where $\mathbf{U}_k$ is the matrix whose columns are the top k right singular vectors of $\mathbf{M}$.*

Lemma 3.11 (Gap-free Wedin, see [AZL16] Lemma B.3)**.** *Let $\varepsilon, \xi, \mu > 0$. For symmetric matrices $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{d\times d}$ for which $\|\mathbf{A} - \widehat{\mathbf{A}}\|_{op} \leq \varepsilon$, if $\widehat{\mathbf{U}}$ is the matrix whose columns consist of the singular vectors of $\widehat{\mathbf{A}}$ with singular value at least μ, and $\mathbf{U}$ is the matrix whose columns consist of the singular vectors of $\mathbf{A}$ with singular value at most $\mu - \xi$, then $\|\widehat{\mathbf{U}}^\top \mathbf{U}\|_{op} \leq \varepsilon/\xi$.*

Corollary 3.12. *Let $\lambda \geq 2\varepsilon > 0$. For symmetric matrices $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{d\times d}$ for which $\|\mathbf{A} - \widehat{\mathbf{A}}\|_{op} \leq \varepsilon$ and $\|\widehat{\mathbf{A}}\|_{op} \geq \lambda - \varepsilon$, if $w \in \mathbb{S}^{d-1}$ is the top singular vector of $\widehat{\mathbf{A}}$, and $V \subset \mathbb{R}^d$ is the orthogonal complement of the kernel of $\mathbf{A}$, then $\|\Pi_V w\|_{op} \geq 1 - 4\varepsilon^2/\lambda^2$.*

Proof. If we take $\xi = \mu = \|\widehat{\mathbf{A}}\|$ in Lemma 3.11, then the columns of $\mathbf{U}$ (resp. $\widehat{\mathbf{U}}$) in Lemma 3.11 consist of an orthonormal basis $B \in \mathbb{R}^{d\times k}$ for the kernel of $\mathbf{A}$ (resp. w and other singular vectors of $\mathbf{A}$, if any, with the same singular value), where k is the dimension of $\ker(\mathbf{A})$. We have that

$$\|\Pi_{V^\perp} w\| \leq \|\widehat{\mathbf{U}}^\top \mathbf{U}\|_{\mathrm{op}} \leq \varepsilon/\|\widehat{\mathbf{A}}\|_{\mathrm{op}} \leq \frac{\varepsilon}{\lambda - \varepsilon},$$

[7] Here we define the sub-exponential norm of a random variable X to be $\sup_{p\geq 1} \frac{1}{p}\mathbb{E}[|X|^p]^{1/p}$

from which we conclude that

$$\|\Pi_V w\| \geq \left(1 - \left(\frac{\varepsilon}{\lambda - \varepsilon}\right)^2\right)^{1/2} \geq 1 - 4\varepsilon^2/\lambda^2$$

as claimed. □

Definition 3.13 (Frames)**.** A set of orthonormal vectors $\widetilde{w}_1, ..., \widetilde{w}_\ell$ is a *frame*. Given subspace $V \subset \mathbb{R}^d$, we say that this frame is *ν-nearly within V* if $\|\Pi_V \widetilde{w}_i\| \geq 1 - \nu$ for all i. We will sometimes refer to their span $\widetilde{W}$ as a frame ν-nearly within to V, when the choice of orthonormal basis for $\widetilde{W}$ is clear from context.

Definition 3.14 (Subspace distances)**.** Given ℓ-dimensional subspaces $U_1, U_2 \subset \mathbb{R}^d$, let $M_1, M_2 \in \mathbb{R}^{d\times\ell}$ denote any two matrices whose columns consists of basis vectors for U_1, U_2 respectively. The *chordal distance* $d_C(U_1, U_2)$ between U_1 and U_2 is defined by

$$d_C(U_1, U_2) = \left(\ell - \|M_1^\top M_2\|_F^2\right)^{1/2}.$$

The *Procrustes distance* $d_P(U_1, U_2)$ between U_1 and U_2 is defined by

$$\inf_{\mathbf{O}\in O(r)} \|U_2 - U_1 \cdot \mathbf{O}\|_F,$$

where $O(r)$ denotes the group of $r \times r$ orthogonal matrices.

Fact 3.15 (See e.g. [CM20], Lemma 3.26)**.** *Given ℓ-dimensional subspaces $U_1, U_2 \subset \mathbb{R}^d$,*

$$d_P(U_1, U_2) \leq \sqrt{2} d_C(U_1, U_2).$$

Lemma 3.16. *Let $\nu \leq O(1/\ell^2)$. If Π is an orthogonal projector to a subspace $V \subset \mathbb{R}^d$, and $\widetilde{w}_1, ..., \widetilde{w}_\ell$ are a frame ν-nearly within V, then there exists an orthonormal set of vectors $w_1, ..., w_\ell$ spanning $W \subset V$ for which $d_C(\widetilde{W}, W) \leq \sqrt{2\nu \cdot \ell}$ and $\|w_i - \widetilde{w_i}\| \leq 2\sqrt{\nu \cdot \ell}$ for all $i \in [\ell]$.*

Proof. Let $\widetilde{W}$ be the subspace spanned by $\widetilde{w}_1, ..., \widetilde{w}_\ell$, and let W be the subspace spanned by $\Pi_V \widetilde{w}_1, ..., \Pi_V \widetilde{w}_\ell$. First note that because $\nu \leq \frac{1}{2\ell^2}$, $\widetilde{W}$ and W have the same dimension, that is, $\Pi_V \widetilde{w}_1 ..., \Pi_V \widetilde{w}_\ell$ are linearly independent. Indeed, we have that $\langle \widetilde{w}_i, \Pi_V \widetilde{w}_i \rangle \geq (1 - \nu)^2 \geq 1 - 2\nu$, while $\langle \widetilde{w}_i, \Pi_V \widetilde{w}_j \rangle = \langle \widetilde{w}_i, \Pi_{V^\perp} \widetilde{w}_j \rangle \leq (1 - (1 - \nu)^2)^{1/2} = \sqrt{2\nu}$ for $i \neq j$, so the Gram matrix of these vectors is diagonally dominant provided $\nu \leq O(1/\ell^2)$.

Overloading notation, let W (resp. $\widetilde{W}$) also denote the $d \times \ell$ matrices whose columns consist of some orthonormal basis vectors for W (resp. the vectors $\widetilde{w}_1, ..., \widetilde{w}_\ell$). The chordal distance $d_C(W, \widetilde{W})$ satisfies

$$d_C(W, \widetilde{W})^2 = \ell - \|W^\top \widetilde{W}\|_F^2 = \ell - \sum \|\Pi_W \widetilde{w}_i\|^2 \leq \ell - \ell \cdot (1 - \nu)^2 \leq 2\nu\ell$$

Letting $O^* \triangleq \arg\inf_{O\in O(r)} \|W - O\widetilde{W}\|_F$ in the definition of $d_P(W, \widetilde{W})$, we can take $w_1, ..., w_\ell$ in the lemma statement to be the columns of OW. Then we have that $d_P(W, \widetilde{W})^2 = \sum \|w_i - \widetilde{w}_i\|^2 \leq 4\nu \cdot \ell$ by Fact 3.15, from which the lemma follows. □

Lemma 3.17. *For any $\mathbf{M} \in \mathbb{R}^{d\times d}$ and a frame $\widetilde{W} \in \mathbb{R}^{d\times\ell}$ which is ν-nearly within an ℓ-dimensional subspace W, we have that*

$$\|\Pi_{\widetilde{W}^\perp} \mathbf{M} \Pi_{\widetilde{W}^\perp} - \Pi_{W^\perp} \mathbf{M} \Pi_{W^\perp}\|_{op} \leq \sqrt{2} \cdot \|\mathbf{M}\|_{op} \cdot d_C(\widetilde{W}, W).$$

Proof. We bound $\|(\Pi_{\widetilde{W}^\perp} - \Pi_{W^\perp})\mathbf{M}\Pi_{\widetilde{W}^\perp}\|_{\text{op}}$ and $\|\Pi_{W^\perp}\mathbf{M}(\Pi_{\widetilde{W}^\perp} - \Pi_{W^\perp})\|_{\text{op}}$ and apply triangle inequality. By sub-multiplicativity of the operator norm and the fact that projections have operator norm 1, $\|(\Pi_{\widetilde{W}^\perp} - \Pi_{W^\perp})\mathbf{M}\Pi_{\widetilde{W}^\perp}\|_{\text{op}} \leq \|\Pi_{\widetilde{W}^\perp} - \Pi_{W^\perp}\|_{\text{op}} \cdot \|\mathbf{M}\|_{\text{op}}$. Finally, note that

$$\|\Pi_{\widetilde{W}^\perp} - \Pi_{W^\perp}\|_2^2 \leq \|\Pi_{\widetilde{W}^\perp} - \Pi_{W^\perp}\|_F^2 = \|\Pi_{\widetilde{W}} - \Pi_W\|_F^2 = 2(\ell - \langle \Pi_{\widetilde{W}}, \Pi_W \rangle) = 2d_C(\widetilde{W}, W)^2,$$

from which the claim follows. □

4 Continuous Piecewise-Linear Functions and Lattice Polynomials

In this section, we introduce tools for reasoning about continuous piecewise-linear functions, culminating in a structural result (Theorem 4.15) giving an explicit representation of arbitrary ReLU networks as lattice polynomials (see Definition 3.2).

4.1 Basic Notions

We will work with functions which only depend on some low-dimensional projection of the input.

Definition 4.1 (Subspace juntas)**.** A function $F : \mathbb{R}^d \to \mathbb{R}$ is a *subspace junta* if there exist $v_1, ..., v_k \in \mathbb{S}^{d-1}$ and a function $h : \mathbb{R}^k \to \mathbb{R}$ for which $F(x) = h(\langle v_1, x\rangle, ..., \langle v_k, x\rangle)$ for all $x \in \mathbb{R}^d$. We will refer to $V \triangleq \text{span}(v_1, ..., v_k)$ as the *relevant subspace* of F, to $v_1, ..., v_k$ as the *relevant directions* of F, and to h as the *link function* of F.

Definition 4.2 (Piecewise Linear Functions)**.** Given vector space W, a function $h : W \to \mathbb{R}$ is said to be *piecewise-linear* (resp. piecewise-affine-linear) if there exist finitely many linear (resp. affine linear) functions $\{g_i : W \to \mathbb{R}\}_{i\in[M]}$ and a partition of W into finitely many polyhedral cones $\{S_i\}_{i\in\mathcal{I}}$ such that $G(x) = \sum_i \mathbb{1}[x \in S_i]g_i(x)$. We will say that h is *realized by M pieces* $\{(g_i, S_i)\}$ (note that h can have infinitely many realizations). If each g_i is given by $g_i(x) = \langle u_i, x\rangle + b_i$ for some $u_i \in W, b_i \in \mathbb{R}$, then we will also refer to the pieces of h by $\{(\langle u_i, \cdot\rangle + b_i, S_i)\}$.

We are now ready to define the concept class we will work with in this paper.

Definition 4.3 ("Kickers")**.** We call a subspace junta F with link function h a *kicker* if h is continuous piecewise-linear. Note that a kicker is itself a continuous piecewise-linear function, and for any realization of its link function by M pieces, there is a realization of F by M pieces.

Henceforth, fix a subspace junta $F : \mathbb{R}^d \to \mathbb{R}$ with link function h and relevant directions $v_1, ..., v_k$ spanning relevant subspace $V \subset \mathbb{R}^d$.

Example 4.4 (ReLU Networks)**.** *Feedforward ReLU networks as defined in Definition 1.1 are kickers with relevant subspace of dimension at most k, where k is the row span of the weight matrix $\mathbf{W}_0$, the link function is defined by*

$$h(z) = \mathbf{W}_{L+1}\phi(\mathbf{W}_L\phi(\cdots \mathbf{W}_1\phi(z)\cdots)),$$

and the pieces in one possible realization of h correspond to the different possible sign patterns that the activations could take on, that is the different possible values of the vector

$$\{\mathbf{W}_a\phi(\mathbf{W}_{a-1}\phi(\cdots \mathbf{W}_1\phi(z)\cdots))\}_{0\leq a\leq L} \in \prod_{a=0}^{L}\{\pm 1\}^{k_a}$$

as z ranges over $\mathbb{R}^k$.

Lemma 4.5. *If F is a Λ-Lipschitz kicker, then for any realization of its link function h by pieces $\{(\langle w_i, \cdot\rangle, S_i)\}$, there is a realization by pieces $\{(\langle w_i', \cdot\rangle, S_i)\}$ for which* $\max_i \|g_i\| \leq L$..

Proof. Consider any piece $(\langle w_i, \cdot\rangle, S_i)$. If there is some $x \in S_i$ for which there exists a ball of nonzero radius r around x contained in S_i, then clearly $L \geq \|w_i\|$: take x and $x + r \cdot w_i$ and note that

$$L \geq \frac{F(x + r \cdot w_i) - F(x)}{\|(x + r \cdot w_i) - x\|} = \frac{r\|w_i\|^2}{r\|w_i\|} = \|w_i\|.$$

If no such x and ball exist, then S_i is not full-dimensional and therefore contained in a hyperplane $W \subset V$. Then if we replace $(\langle w_i, \cdot\rangle, S_i)$ in the realization of h with $(\langle \Pi_W w_i, \cdot\rangle, S_i)$, this is still a realization of h. Again, it would suffice for there to exist a ball, now in the subspace W, of nonzero radius around some point in S_i. If this is not the case, then S_i is not a full-dimensional subset of W and thus lies in a codimension 1 subspace of W. Continuing thus, we eventually obtain some (possibly zero) vector w_i' for which replacing $(\langle w_i, \cdot\rangle, S_i)$ in the realization of h with $(\langle w_i', \cdot\rangle, S_i)$ still gives a realization of h, and furthermore $\|w_i'\| \leq L$. □

Definition 4.6 (Restrictions)**.** Given any nonzero linear subspace $W \subseteq V$, let $F|_W : W \to \mathbb{R}$ denote the restriction of F to the subspace W. By abuse of notation, we will sometimes also regard $F|_W$ as a function over $\mathbb{R}^d$ given by $F|_W(x) = F(\Pi_W x)$.

One of the main properties of kickers that we exploit is *positive homogeneity*:

Fact 4.7 (Positive homogeneity)**.** *For any $\lambda \geq 0$ and $x \in \mathbb{R}^k$, $F(\lambda \cdot x) = \lambda F(x)$.*

The following property of restrictions of Lipschitz functions will be important.

Lemma 4.8. *For any nonzero linear subspace $W \subseteq V$, and Λ-Lipschitz function $F : \mathbb{R}^d \to \mathbb{R}$,*

$$\sup_{x : \|\Pi_{V \setminus W} x\| \leq 1} |F(x) - F(\Pi_W x)| \leq \Lambda.$$

Proof. Because $F(x) = F(\Pi_V x)$ and $F(\Pi_W x) = F(\Pi_W \Pi_V x)$, we may assume without loss of generality that $x \in V$. For any $x \in V$ for which $\|\Pi_{V \setminus W} x\| \leq 1$, we have that

$$|F(x) - F(\Pi_W x)| \leq \Lambda \|x - \Pi_W x\| = \Lambda \|\Pi_{V \setminus W} x\| \leq \Lambda,$$

as claimed. □

4.2 A Generic Lattice Polynomial Representation

Essential to our analysis is the following structural result from [Ovc02] which says that, perhaps surprisingly, *any* piecewise linear function can be expressed as a relatively simple lattice polynomial.

Theorem 4.9 ([Ovc02], Theorem 4.1)**.** *If $h : \mathbb{R}^n \to \mathbb{R}$ is a continuous piecewise-linear function which has a realization by pieces $\{(g_i, S_i)\}_{i \in [M]}$, there exists a collection of* clauses *$\mathcal{I}_1, ..., \mathcal{I}_m \subseteq [M]$ for which*

$$h(x) = \max_{j \in [m]} \min_{i \in \mathcal{I}_j} g_i(x) \tag{5}$$

We will work with the following notion of approximation for such lattice polynomials:

Definition 4.10. Two continuous piecewise-linear functions $G, \widetilde{G} : \mathbb{R}^d \to \mathbb{R}$ are (M, η)-*structurally-close* if there exist linear functions $g_1, ..., g_M$ and $\widetilde{g}_1, ..., \widetilde{g}_M$ and subsets $\mathcal{I}_1, ..., \mathcal{I}_m \subseteq [M]$ for which

$$G(x) = \max_{j \in [m]} \min_{i \in \mathcal{I}_i} g_i(x) \qquad \widetilde{G}(x) = \max_{j \in [m]} \min_{i \in \mathcal{I}_i} \widetilde{g}_i(x)$$

and $\|g_i - \widetilde{g}_i\| \leq \eta$ for all i.

Structural closeness of continuous piecewise-linear functions in the above sense is stronger than L_2-closeness.

Lemma 4.11. *Take continuous piecewise-linear functions $G, \widetilde{G} : \mathbb{R}^m \to \mathbb{R}$ which are (M, η)-structurally-close. Then $\|G - \widetilde{G}\| \leq \eta\sqrt{m}$. In particular, if G is a piecewise-linear function which is realized by pieces $\{(\langle u_i, \cdot \rangle, S_i)\}$ satisfying $\|u_i\| \leq \eta$, then $\|G\| \leq \eta\sqrt{m}$.*

To show this, we need the following helper lemma:

Lemma 4.12. *If $\{g_i\}_{i \in [M]}$ and $\{\widetilde{g}_i\}_{i \in [M]}$ are two collections of linear functions, then for any x,*

$$|\max_{j \in [m]} \min_{i \in \mathcal{I}_j} g_i(x) - \max_{j \in [m]} \min_{i \in \mathcal{I}_j} \widetilde{g}_i(x)| \leq \max_i |g_i(x) - \widetilde{g}_i(x)|$$

Proof. This simply follows by induction using the fact that if $f_1, f_2 : \mathbb{R}^a \to \mathbb{R}$ are both 1-Lipschitz with respect to L_∞, then $f_1 \vee f_2$ and $f_1 \wedge f_2$ are as well. □

Proof of Lemma 4.11. Let $\{(\langle u_i, \cdot \rangle, S_i)\}_{i \in [M]}$ and $\{(\langle \widetilde{u}_i, \cdot \rangle, S_i)\}_{i \in [M]}$ be the realizations of $G, \widetilde{G}$ for which $\|u_i - \widetilde{u}_i\| \leq \eta$. By Lemma 4.12 applied to these pieces, together with Cauchy-Schwarz, for any x we have that $|G(x) - \widetilde{G}(x)| \leq \eta\|x\|$. So $\|G - \widetilde{G}\| \leq \eta \cdot \mathbb{E}[\|x\|^2]^{1/2} = \eta\sqrt{m}$. □

As discussed in Section 2, for our application to learning general kickers, we will leverage the lattice polynomial representation in Theorem 4.9 to grid over piecewise-linear functions. Note that *a priori*, even if we knew exactly the set of linear functions $\{g_i\}_{i \in [M]}$ in a realization of a piecewise-linear function, enumerating over all lattice polynomials of the form (5) would require time doubly exponential in M, as there are 2^M possible clauses $\mathcal{I}_j$ and 2^{2^M} possible sets of clauses $\{\mathcal{I}_j\}$.

By being slightly more careful, we can enumerate over piecewise linear functions in time $\exp(\text{poly}(M))$.

Definition 4.13. An *order type on n elements* is specified by a function $\omega : [n] \to [n]$ for which every element from 1 to $\max_i \omega(i)$ is present. We say that a set of n real numbers $z_1, ..., z_n$ has order type ω (denoted $\{z_1, ..., z_n\} \vdash \omega$ if $z_i = z_j$ (resp. $z_i > z_j$, $z_i < z_j$) if and only if $\omega(i) = \omega(j)$ (resp. $\omega(i) > \omega(j)$, $\omega(i) < \omega(j)$). Denote the set of order types on n elements by Ω_n. Note that any set of real numbers has exactly one order type.

Lemma 4.14. *If F has a realization by pieces $\{(g_i, S_i)\}_{i \in [M]}$, then there is a function $A : \Omega_M \to [M]$ such that for any x,*

$$F(x) = \sum_{\omega \in \Omega_M} \mathbb{1}\Big[\{g_i(x)\}_{i \in [M]} \vdash \omega\Big] \cdot g_{A(\omega)}(x).$$

Proof. Let $F(x) = \max_{j \in [m]} \min_{i \in \mathcal{I}_j} g_i(x)$ be the max-min representation guaranteed by Theorem 4.9. This representation implies that for a fixed order type ω, there is some index $i \in [M]$ for which $F(x) = g_i(x)$ for all x satisfying $\{g_i(x)\}_{i \in [M]} \vdash \omega$. This gives the desired mapping A. □

Note that the set of functions $A : \Omega_M \to [M]$ is only of size $(M!)^M \leq M^{M^2}$, so by Lemma 4.14, to enumerate over piecewise-linear functions with M pieces we can simply enumerate over linear functions $\{g_i\}$ together with all possible functions A (see Algorithm 1 below).

4.3 Lattice Polynomials for ReLU Networks

Here we give an explicit proof of Theorem 4.9 in the special case of ReLU networks. We emphasize that the specific nature of the construction exhibited in this theorem will be important in the proof of our main result for learning ReLU networks, and that simply applying Theorem 4.9 in a black-box fashion will not suffice for our purposes.

Theorem 4.15. *If $F \in \mathcal{C}_S$ is a ReLU network with weight matrices $\mathbf{W}_0 \in \mathbb{R}^{k_0 \times d}, \mathbf{W}_1 \in \mathbb{R}^{k_1 \times k_0}, \ldots, \mathbf{W}_L \in \mathbb{R}^{k_L \times k_{L-1}}, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L}$, and if F' is a ReLU network with the same architecture as F, with weight matrices $\mathbf{W}'_0, ..., \mathbf{W}'_{L+1}$, such that*

$$(\mathbf{W}_a)_{i,j} \cdot (\mathbf{W}'_a)_{i,j} \geq 0 \qquad \forall\ 0 \leq a \leq L+1, (i,j) \in [k_a] \times [k_{a-1}],$$

then there exist vectors $v_1, ..., v_M, v'_1, ..., v'_M$ and clauses $\mathcal{I}_1, ..., \mathcal{I}_m \subseteq [M]$, where $M = 2^S$, for which

$$F(x) = \max_{j \in [m]} \min_{i \in \mathcal{I}_j} \langle v_i, x \rangle$$
$$F'(x) = \max_{j \in [m]} \min_{i \in \mathcal{I}_j} \langle v'_i, x \rangle.$$

Specifically, $v_1, ..., v_M$ consist of all vectors of the form $\mathbf{W}_{L+1}\mathbf{\Sigma}_L\mathbf{W}_L\mathbf{\Sigma}_{L-1} \cdot \cdots \cdot \mathbf{\Sigma}_0\mathbf{W}_0$ for diagonal matrices $\mathbf{\Sigma}_i \in \{0,1\}^{k_i \times k_i}$, and $v'_1, ..., v'_M$ are defined analogously.

We prove Theorem 4.15 by induction by exhibiting max-min representations for ReLUs, scalings, and sums of max-min formulas. Let $G : \mathbb{R}^d \to \mathbb{R}$ be a piecewise-linear function given by $G(x) \triangleq \max_{j \in [m]} \min_{i \in \mathcal{I}_j} \langle u_i, x \rangle$ for some subsets $\{\mathcal{I}_1, ..., \mathcal{I}_m\}$ of $[M]$ and vectors $\{u_1, ..., u_M\}$ in $\mathbb{R}^d$.

Lemma 4.16. *Let $u_{M+1} = 0$ and let $\mathcal{I}_{m+1} = \{M+1\}$. Then for all $x \in \mathbb{R}^d$,*

$$\phi(G(x)) = \max_{j \in [m+1]} \min_{i \in \mathcal{I}_j} \langle u_i, x \rangle.$$

Proof. This is immediate from the definition of ϕ. □

Lemma 4.17. *For any $\lambda \in \mathbb{R}$, there exist subsets $\{\mathcal{J}_1, ..., \mathcal{J}_{m'}\}$ of $[M]$ such that for all $x \in \mathbb{R}^d$,*

$$\lambda G(x) = \max_{j \in [m']} \min_{i \in \mathcal{J}_j} \langle \lambda u_i, x \rangle.$$

Furthermore, these subsets only depend on $\mathcal{I}_1, ..., \mathcal{I}_m$ and the sign of λ.

Proof. For $\lambda > 0$, we have $\mathcal{J}_j = \mathcal{I}_j$ for all j. So it remains to show the claim for $\lambda = -1$. We can write $-G(x)$ as $\min_{j \in [m]} \max_{i \in \mathcal{I}_j} \langle u_i, x \rangle$. This is a lattice polynomial over the reals, and any lattice polynomial over a distributive lattice can be written in disjunctive normal form as $\max_{j \in [m']} \min_{i \in \mathcal{J}_j} \langle u_i, x \rangle$ for some subsets $\{\mathcal{J}_j\}$ (see e.g. [Bir40, Section II.5, Lemma 3]), from which the claim follows. □

Lemma 4.18. *For any $k' \in \mathcal{N}$ and $b \in [k']$, let $G_b(x) = \max_{j \in [m_b]} \min_{i \in \mathcal{I}_j^b} \langle u_i^b, x \rangle$ for some subsets $\{\mathcal{I}_j^b\}$ of $[M_b]$ and vectors $\{u_i^b\}$ in $\mathbb{R}^d$. For all $x \in \mathbb{R}^d$,*

$$\sum_{b=1}^{k'} G_b(x) = \max_{(j_1,...,j_{k'}) \in [m_1] \times \cdots \times [m_{k'}]} \ \min_{(i_1,...,i_{k'}) \in \mathcal{I}_{j_1} \times \cdots \times \mathcal{I}_{j_{k'}}} \langle u_{i_1}^1 + \cdots + u_{i_{k'}}^{k'}, x \rangle. \tag{6}$$

Proof. Take any $x \in \mathbb{R}^d$, and for $b \in [k']$ suppose that $G_b(x) = \langle u^b_{i^*_b}, x\rangle$ for some index $i^*_b \in [M]$. Note that for any $\mathcal{I}^1_{j_1}, \dots, \mathcal{I}^{k'}_{j_{k'}}$ containing $i^*_1, \dots, i^*_{k'}$ respectively,

$$\min_{(i_1,\dots,i_{k'})\in \mathcal{I}^1_{j_1}\times\cdots\times\mathcal{I}^{k'}_{j_{k'}}} \langle u^1_{i_1} + \cdots + u^{k'}_{i_{k'}}, x\rangle = \langle u^{k'}_{i^*_{k'}} + \cdots + u^{k'}_{i^*_{k'}}, x\rangle.$$

This shows that the right-hand side of (6) is lower bounded by the left-hand side.

We now show the other direction. For any $i'_1, \dots, i'_{k'}$ for which $\langle u^1_{i'_1} + \ldots + u^{k'}_{i'_{k'}}, x\rangle > G_1(x) + \cdots + G_{k'}(x)$, we must have $\langle u^b_{i'_b}, x\rangle > G_b(x)$ for some $b \in [k']$. In this case, we know that for every clause $\mathcal{I}^b_{j_b}$ in G_b which contains i'_b, there is some $i \in \mathcal{I}^b_{j_b}$ for which $\langle u^b_i, x\rangle < \langle u^b_{i'_b}, x\rangle$. So for any $\mathcal{I}^1_{j_1}, \dots, \mathcal{I}^{k'}_{j_{k'}}$ containing $i'_1, \dots, i'_{k'}$ respectively, the corresponding clause on the right-hand side of (6) satisfies $\min_{(i_1,\dots,i_{L'})\in \mathcal{I}^1_{j_1}\times\cdots\times\mathcal{I}^{L'}_{j_{L'}}} \langle u^1_{i_1} + \cdots + u^{L'}_{i_{L'}}, x\rangle < \langle u^1_{i'_1} + \cdots + u^{L'}_{i'_{L'}}, x\rangle$. This concludes the proof that the left-hand side of (6) is upper bounded by the left-hand side. ☐

We can now prove Theorem 4.15:

Proof. The claim is trivially true for $L = -1$. Suppose inductively that for some layer $0 \le a \le L$, we have that for all $b \in [k_a]$, if we denote

$$F_{a,b} \triangleq \mathbf{W}^b_a \phi\left(\mathbf{W}_{a-1}\phi\left(\cdots\phi\left(\mathbf{W}_0 x\right)\right)\right)$$
$$F'_{a,b} \triangleq \mathbf{W}'^b_a \phi\left(\mathbf{W}'_{a-1}\phi\left(\cdots\phi\left(\mathbf{W}'_0 x\right)\right)\right),$$

where $\mathbf{W}^b_a$ denotes the b-th row of $\mathbf{W}_a$, then $F_{a,b}$ and $F'_{a,b}$ can be expressed as max-min formulas $\max_{j\in[m_{a,b}]}\min_{i\in\mathcal{I}^{a,b}_j}\langle v^{a,b}_i, \cdot\rangle$ and $\max_{j\in[m_{a,b}]}\min_{i\in\mathcal{I}^{a,b}_j}\langle v'^{a,b}_i, \cdot\rangle$ for some clauses $\{\mathcal{I}^{a,b}_j\}$ and vectors $v^{a,b}_i, v'^{a,b}_i$ comprised respectively of vectors of the form $\mathbf{W}^b_a\mathbf{\Sigma}_{a-1}\cdots\mathbf{\Sigma}_0\mathbf{W}_0$ and $\mathbf{W}'^b_a\mathbf{\Sigma}_{a-1}\cdots\mathbf{\Sigma}_0\mathbf{W}'_0$ for all possible diagonal matrices $\mathbf{\Sigma}_i \in \{0,1\}^{k_i\times k_i}$. Then for any $b \in [k_{a+1}]$, note that $F_{a+1,b} = \mathbf{W}^b_{a+1}\phi(F_{a,1}, ..., F_{a,k_a})$ and $F'_{a+1,b} = \mathbf{W}'^b_{a+1}\phi(F'_{a,1}, ..., F'_{a,k_a})$. By Lemma 4.16 and Lemma 4.17, if the entries of $\mathbf{W}^b_a$ and $\mathbf{W}'^b_a$ are $w_1, ..., w_{k_{a+1}}$ and $w'_1, ..., w'_{k_{a+1}}$ respectively, then for every $b' \in [k_a]$, if $w_{b'}\cdot w'_{b'} \ge 0$, then there exist max-min representations for $w_{b'}\phi(F_{a,b'})$ and $w'_{b'}\phi(F_{a,b'})$ with the same set of clauses.

Finally, by Lemma 4.18, there exist max-min representations for the scalar-valued functions $F_{a+1,b} = \sum_{b'=1}^{k_a} w_{b'}\phi(F_{a,b'})$ and $F'_{a+1,b} = \sum_{b'=1}^{k_a} w'_{b'}\phi(F'_{a,b'})$ with the same set of clauses. And the vectors in this max-min representation consist of all vectors of the form $\mathbf{W}^b_{a+1}\mathbf{\Sigma}_a\cdot\cdots\cdot\mathbf{\Sigma}_0\mathbf{W}_0$ and $\mathbf{W}'^b_{a+1}\mathbf{\Sigma}_a\cdot\cdots\cdot\mathbf{\Sigma}_0\mathbf{W}'_0$ respectively for $\mathbf{\Sigma}_i \in \{0,1\}^{k_i\times k_i}$. This completes the inductive step. ☐

5 Filtered PCA

In this section we prove our main results on learning kickers and ReLU networks. Throughout, we will make the followingbase assumption about the function F.

Assumption 1. *F is a kicker which is Λ-Lipschitz for some $\Lambda \ge 1$ and has at most M pieces.*

While our techniques are general enough to work under just this assumption, for our main application to learning ReLU networks (Definition 1.1), we can obtain improved runtime guarantees by making the following additional assumption on F.

Assumption 2. *F is computed by a size-S ReLU network*[8] *with depth $L+2$ and weight matrices $\mathbf{W}_0 \in \mathbb{R}^{k_0 \times d}, \dots \mathbf{W}_L \in \mathbb{R}^{k_L \times k_{L-1}}, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L}$ satisfying $\|\mathbf{W}_i\|_{op} \le B$ for all $0 \le i \le L+1$, for some $B \ge 1$.*[9]

In this section, unless stated otherwise, we will only assume F satisfies Assumption 1, but in certain parts of the proof (e.g. Section 5.5), we will get better bounds by additionally making Assumption 2. Formally, our main results are the following:

Theorem 5.1. *Given access to samples from the distribution $\mathcal{D}$ corresponding to kicker F satisfying Assumption 1,* FilteredPCA*$(\mathcal{D}, \varepsilon, \delta)$ outputs a kicker $\widetilde{F}$ for which $\mathbb{E}[(y - \widetilde{F}(x))^2] \le \varepsilon^2$ with probability at least $1 - \delta$. Furthermore,* FilteredPCA *has sample complexity*

$$d \log(1/\delta) \cdot \mathrm{poly}\left(\exp\left(k^3 \Lambda^2 / \varepsilon^2\right), M^k\right)$$

and runtime

$$\widetilde{O}(d^2 \log(1/\delta)) \cdot M^{M^2} \cdot \mathrm{poly}\left(\exp\left(k^4 \Lambda^2 / \varepsilon^2\right), M^{k^2}\right).$$

Theorem 5.2. *Given access to samples from the distribution $\mathcal{D}$ corresponding to feedforward ReLU network F satisfying Assumption 2,* FilteredPCA*$(\mathcal{D}, \varepsilon, \delta)$ outputs a ReLU network $\widetilde{F}$ for which $\mathbb{E}[(y - \widetilde{F}(x))^2] \le \varepsilon^2$ with probability at least $1 - \delta$. Furthermore,* FilteredPCA *has sample complexity*

$$d \log(1/\delta) \mathrm{poly}\left(\exp\left(k^3 \Lambda^2 / \varepsilon^2\right), 2^{kS}, \left(B^{(L+2)} / \Lambda\right)^k\right)$$

and runtime

$$\widetilde{O}(d^2 \log(1/\delta)) \cdot \mathrm{poly}\left(\exp\left(k^3 S^2 \Lambda^2 / \varepsilon^2\right), 2^{kS^3}, \left(B^{L+2} / \Lambda\right)^{kS^2}\right).$$

Remark 5.3 (Scale Invariance). Often, guarantees for PAC learning ReLU networks are stated scale-invariantly in terms of the relative error $\mathbb{E}[(y - \widetilde{F}(x))^2] / \mathbb{E}[y^2]$, or equivalently the absolute error $\mathbb{E}[(y - \widetilde{F}(x))^2]$ for the true F satisfying $\mathbb{E}[y^2] = 1$.

In our general setting, recall from Example 1.4 that some dependence on the Lipschitz constant of F is needed. One standard way to achieve this is to normalize the weight matrices of the true underlying network F to have operator norm at most B, in which case the Lipschitz constant of F is at most B^{L+2} and, with our techniques, we can obtain guarantees depending just on B by using Theorem 5.1. To obtain improved guarantees, we can additionally assume a better bound of Λ on the Lipschitz constant, and this gives rise to Theorem 5.2 above.

Under this normalization in terms of Λ and B, note that the sample complexity and runtime in Theorem 5.2 are scale invariant as the quantities Λ/ε and B^{L+2}/Λ are invariant under arbitrary rescalings of the $L+2$ weight matrices of F. Also note that Λ can be any *upper bound* on the actual Lipschitz constant of F, that is, the runtime guarantee in Theorem 5.2 does not degrade with the actual Lipschitz constant of F.

In Section 5.1, we prove an anti-concentration result for piecewise-linear functions. We use this in Section 5.2 to prove that in an idealized scenario where we had exact access to some ℓ-dimensional $W \subset V$ as well as exact query access to $F|_W$, we would be able to approximately recover a vector in $V \backslash W$ by running one iteration of the main loop of FilteredPCA. In the remaining sections, we show how to pass from this idealized scenario to the setting we actually care about, in which we

[8] Note that this implies $M \le 2^S$.

[9] Recall from Definition 1.1 that we will refer to the rank of $\mathbf{W}_0$ as k to emphasize that F is a kicker with relevant subspace V of dimension k.

only samples $(x, F(x))$. In Section 5.3 we show that affine thresholds of piecewise-linear functions are stable under small perturbations of the function. Then in Section 5.4, we show how to grid over the set of kickers, and in Section 5.5 we show how to grid over ReLU networks more efficiently and formally state our algorithm. In Section 5.6 we combine these ingredients to argue that as long as we have sufficiently good approximate access to W and $F|_W$, a single iteration of the main loop of FilteredPCA will approximately recover a vector from $V \backslash W$. Lastly, in Section 5.7 we conclude the proofs of Theorem 5.1 and 5.2. At the very end, we discuss briefly why merely adapting the approach of [CM20] does not work.

5.1 Anti-Concentration of Piecewise Linear Functions

In this section, we show that for any continuous piecewise-linear function with some variance, the probability that it exceeds any given threshold is non-negligible.

Lemma 5.4. *If $G : \mathbb{R}^m \to \mathbb{R}$ is continuous piecewise-linear and Λ-Lipschitz and $\mathbb{E}[G^2] \geq \sigma^2$, then for any $s \geq 0$,*

$$\mathbb{P}[|G| > s] \geq \Omega(\exp(-3ms^2/\sigma^2)) \cdot \frac{s\sigma}{\sqrt{m}\Lambda^2}.$$

Proof. Let $\{(g_i, S_i)\}$ be the pieces of some realization G, and for every i let $u_i \in \mathbb{R}^m$ be the vector for which $g_i(\cdot) = \langle u_i, \cdot \rangle$. By Lemma 4.5, we can assume $\|u_i\| \leq \Lambda$ for all i.

Take any i and define

$$\sigma_i^2 \triangleq \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,\mathrm{Id})} [\langle u_i, x \rangle^2 \mid x \in S_i]$$

Note that if i is chosen with probability $\mathbb{P}[x \in S_i]$, then $\mathbb{E}_i[\sigma_i^2] \geq \sigma^2$. Because each S_i is a polyhedral cone, sampling $x \sim \mathcal{N}(0, \mathrm{Id})$ conditioned on $x \in S_i$ is equivalent to sampling $r \sim \chi^2_m$, independently sampling $\widehat{x} \sim \mathbb{S}^{m-1}$ conditioned on $\widehat{x} \in S_i$, and outputting $r^{1/2} \cdot \widehat{x}$. It follows that

$$\sigma_i^2 = \mathop{\mathbb{E}}_{r \sim \chi^2_m, \widehat{x} \sim \mathbb{S}^{m-1}} [r \cdot \langle u_i, \widehat{x} \rangle^2 \mid \widehat{x} \in S_i] = \mathop{\mathbb{E}}_{r \sim \chi^2_m} [r] \cdot \mathop{\mathbb{E}}_{\widehat{x} \sim \mathbb{S}^{m-1}} [\langle u_i, \widehat{x} \rangle^2 \mid \widehat{x} \in S_i] = m \cdot \mathop{\mathbb{E}}_{\widehat{x} \sim \mathbb{S}^{m-1}} [\langle u_i, \widehat{x} \rangle^2 \mid \widehat{x} \in S_i].$$

By Fact 3.3, $\mathbb{P}[|\langle u_i, \widehat{x} \rangle| \geq \sigma_i / \sqrt{2m} \mid \widehat{x} \in S_i] \geq \frac{\sigma_i^2}{2m\|u_i\|^2}$. We conclude that for any $s > 0$,

$$\begin{aligned} \mathbb{P}[|\langle u_i, x \rangle| \geq s \mid x \in S_i] &\geq \mathop{\mathbb{P}}_{r \sim \chi^2_m} \left[r > 2ms^2/\sigma_i^2 \right] \cdot \frac{\sigma_i^2}{2m\|u_i\|^2} \\ &\geq \operatorname{erfc}(s\sqrt{2m}/\sigma_i) \cdot \frac{\sigma_i^2}{2m\Lambda^2} \end{aligned} \tag{7}$$

By Fact 3.5, the right-hand side of (7) is convex as a function of σ_i^2, so

$$\begin{aligned} \mathbb{P}[|G(x)| > s] &\geq \mathop{\mathbb{E}}_i \left[\operatorname{erfc}(s\sqrt{2m}/\sigma_i) \cdot \frac{\sigma_i^2}{2m\Lambda^2} \right] \\ &\geq \operatorname{erfc}(s\sqrt{2m}/\mathop{\mathbb{E}}_i[\sigma_i^2]^{1/2}) \cdot \frac{\mathbb{E}_i[\sigma_i^2]}{2m\Lambda^2} \\ &\geq \operatorname{erfc}(s\sqrt{2m}/\sigma) \cdot \frac{\sigma^2}{2m\Lambda^2} \\ &\geq \sqrt{2/\pi} \cdot \frac{s\sqrt{2m} \cdot \exp(-ms^2/\sigma^2)}{\sigma \cdot (2ms^2/\sigma^2 + 1)} \cdot \frac{\sigma^2}{2m\Lambda^2} \\ &\geq \Omega(\exp(-3ms^2/\sigma^2)) \cdot \frac{s\sigma}{\sqrt{m}\Lambda^2}, \end{aligned}$$

where the second step follows by Jensen's and the fourth step follows by Fact 3.4. ∎

5.2 An Idealized Calculation

Suppose we had access to an orthonormal collection of vectors $w_1, \dots, w_\ell$ that are *exactly* in V. Let W denote their span. Suppose further that we had access to the matrix

$$\mathbf{M}_\tau^W \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y}\Big[\mathbb{1}[|y - F(\Pi_W x)| > \tau] \cdot (xx^\top - \mathrm{Id})\Big]\Pi_{W^\perp}.$$

When the threshold τ is clear from context, we will just refer to this matrix as $\mathbf{M}^W$.

As we will see, if this matrix is nonzero, then its singular vectors with nonzero singular value must lie in V and be orthogonal to $w_1, \dots, w_\ell$. The main challenge will be to show that this matrix is nonzero. The following proof also applies to the case of $\ell = 0$, in which case $F(\Pi_W x)$ specializes to the zero function and (8) specializes to

$$\mathbf{M}_\tau^\emptyset \triangleq \mathop{\mathbb{E}}_{x,y}\Big[\mathbb{1}[|y| > \tau] \cdot (xx^\top - \mathrm{Id})\Big]. \tag{8}$$

In particular, (8) is a matrix we actually have access to at the beginning of the algorithm, and one consequence of the warmup argument below is an algorithm for finding a single vector in V.

We first show that for appropriately chosen τ, either the top singular value of $\mathbf{M}_\tau^W$ is non-negligible, or $\mathbb{E}[(F(x) - F(\Pi_W x)^2]$ is small, that is, F is already sufficiently well-approximated by the function $F|_W$.

Lemma 5.5. *Suppose $\mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{Id})}[(F(x) - F(\Pi_W x))^2] \geq \rho^2$ for some $\rho > 0$. For any $\tau > 0$, if a vector is not in the kernel of $\mathbf{M}_\tau^W$, then it must lie in $V\backslash W$. For $\tau \geq \sqrt{2(k-\ell)} \cdot \Lambda$,*

$$\left\langle \mathbf{M}_\tau^W, \Pi_{V\backslash W} \right\rangle \geq \Omega\left(e^{-3k\tau^2/\rho^2}\right) \cdot \frac{(k-\ell)\tau\rho}{\sqrt{k}\Lambda^2}. \tag{9}$$

In particular, for this choice of τ, the top singular vector of $\mathbf{M}_\tau^W$ lies in $V\backslash W$ and has singular value at least $\lambda_\tau^{(\ell)} \triangleq \Omega\left(e^{-3k\tau^2/\rho^2}\right) \cdot \frac{\tau\rho}{\sqrt{k}\Lambda^2}$.

Proof. The first part just follows from the fact that any $u \in \Pi_W$ is clearly in the kernel, and for any $u \in \mathbb{S}^{d-1}$ orthogonal to V, $\langle u, x\rangle$ and $F(x)$ are independent, so

$$u^\top \mathbf{M}_\tau^W u = \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,1)}[g^2 - 1] \cdot \mathop{\mathbb{E}}_x[\mathbb{1}[|F(x) - F(\Pi_W v)| > \tau]] = 0.$$

For (9), we would like to apply Lemmas 4.8 and 5.4 to the continuous piecewise-linear function $G(x) \triangleq F(x) - F(\Pi_W x)$. Pick an orthonormal basis $w_{\ell+1}, \dots, w_k$ for $V\backslash W$. For any x for which $\|\Pi_{V\backslash W} x\| \leq 1$, Lemma 4.8 implies $|G(x)| \leq \Lambda$. So by positive homogeneity (see Fact 4.7) of $G(x)$ and the definition of τ, $|G(x)| > \tau$ only if $\|\Pi_{V\backslash W} x\|^2 \geq 2(k-\ell)$, so

$$\begin{aligned}\sum_{i=\ell+1}^{k} w_i^\top \mathbf{M}_\tau^W w_i &= \mathop{\mathbb{E}}_x\Big[\mathbb{1}[|G(x)| > \tau] \cdot \Big(\|\Pi_{V\backslash W} x\|^2 - (k-\ell)\Big)\Big] \\ &\geq (k-\ell) \cdot \mathop{\mathbb{P}}_x[G(x) > \tau].\end{aligned}$$

(9) then follows from Lemma 5.4 applied to G.

The final statement in Lemma 5.5 follows by averaging. ∎

If ε is the target L_2 error to which we want to learn F, we will only ever work with $\rho \geq \Omega(\varepsilon)$. In the sequel, we will take

$$\tau = c\sqrt{k} \cdot \Lambda \tag{10}$$

for sufficiently large absolute constant $c > 0$. As a result, we have that

$$\lambda_\tau^{(\ell)} \geq \Omega\left(e^{-O(k^2\Lambda^2/\varepsilon^2)}\right) \cdot (\varepsilon/\Lambda) \triangleq \underline{\lambda}. \tag{11}$$

5.3 Stability of Piecewise Linear Threshold Functions

To get an iterative algorithm for finding all relevant directions of F, we need to show an analogue of Lemma 9 in the setting when we only have access to directions $\widetilde{w}_1, \ldots, \widetilde{w}_\ell$ which are *close* to the span of V, and when we only have access to an *approximation* of the function $F|_W$.

In this section, we show the following stability result for affine thresholds of piecewise-linear functions:

Lemma 5.6. *Let $f, g, g' : \mathbb{R}^d \to \mathbb{R}$ be piecewise-linear functions. For any $\tau > 0$, if g, g' are (m, η)-structurally-close and f has a realization with at most m pieces, then*

$$\mathop{\mathbb{P}}_{x \sim \mathcal{N}(0,\mathrm{Id})}\left[|g(x) - f(x)| > \tau \wedge |g'(x) - f(x)| \leq \tau\right] \leq 9\eta m^2/\tau \tag{12}$$

An important building block of the proof is the special case where $f = 0$ and g, g' are linear:

Lemma 5.7. *For $\tau > 0$ and vectors $v, v' \in \mathbb{R}^d$,*

$$\mathop{\mathbb{P}}_{x \sim \mathcal{N}(0,\mathrm{Id})}[\langle v, x\rangle > \tau \wedge \langle v', x\rangle \leq \tau] \leq O\left(\frac{\|v - v'\|}{\tau}\right) \tag{13}$$

Proof. First note that without loss of generality, we may assume that $\|v\| \geq \|v'\|$; if not, then the random variable $\mathbb{1}[\langle v, x\rangle > \tau \wedge \langle v', x\rangle \leq \tau]$ is stochastically dominated by $\mathbb{1}[\langle v, x\rangle > \tau \wedge \langle \zeta v', x\rangle \leq \tau]$ for $\zeta = \|v\|/\|v'\|$, and furthermore $\|v - \zeta v'\| \leq \|v - v'\|$ by the Pythagorean theorem.

Also note that we may assume $\|v'\| > \|v - v'\|$. Otherwise, we would have $\|v\| \leq 2\|v - v'\|$. But then we could upper bound the left-hand side of (13) by

$$\mathbb{P}[\langle v, x\rangle > \tau] \leq e^{-\tau^2/2\|v\|^2} \leq e^{-\frac{\tau^2}{8\|v-v'\|^2}} \leq 2\|v - v'\|/\tau.$$

Now define $\widehat{v} = v/\|v\|$ and $\widehat{v'} = v'/\|v'\|$ so that (13) equals $\mathbb{P}[\langle \widehat{v}, x\rangle > \widehat{\tau} \wedge \langle \widehat{v'}, x\rangle \leq \widehat{\tau}']$ for $\widehat{\tau} \triangleq \tau/\|v\|$ and $\widehat{\tau}' \triangleq \tau/\|v'\|$. Write $\widehat{v}' = \alpha\widehat{v} + \sqrt{1-\alpha^2}v^\perp$ for $v^\perp$ orthogonal to $\widehat{v}$, and denote the random variables $\langle \widehat{v}, x\rangle$ and $\langle \widehat{v}', x\rangle$ by γ and γ' respectively (these are α-correlated standard Gaussians).

Note that by the assumption that $\|v\| \geq \|v'\| \geq \|v - v'\|$, the angle between v and v' is at most $\pi/3$, so $\alpha \geq 1/2$.

We are now ready to upper bound (13). We will split into two cases, either $\gamma > \widehat{\tau}'/\alpha$ or $\widehat{\tau} \leq \gamma \leq \widehat{\tau}'$, and upper bound the contribution of either case to the probability in (13) by $O(\|v - v'\|/\tau)$, from which the lemma will follow.

Case 1: $\gamma > \widehat{\tau}'/\alpha$.

The density of γ' relative to γ is given by

$$\int_{-\infty}^{\frac{\widehat{\tau}' - \alpha\gamma}{\sqrt{1-\alpha^2}}} \mathcal{N}(0,1,x)dx = \frac{1}{2}\operatorname{erfc}\left(\frac{\alpha\gamma - \widehat{\tau}'}{\sqrt{1-\alpha^2}}\right) \leq \frac{1}{2}\exp\left(-\frac{(\alpha\gamma - \widehat{\tau}')^2}{2(1-\alpha^2)}\right).$$

We have that

$$\mathop{\mathbb{E}}_{\gamma}\left[\frac{1}{2}\exp\left(-\frac{(\alpha\gamma-\widehat{\tau}')^2}{2(1-\alpha^2)}\right)\cdot\mathbb{1}[\gamma>\widehat{\tau}']\right]=\frac{1}{4}\sqrt{1-\alpha^2}\cdot\exp(-\widehat{\tau}'^2/2)\cdot\operatorname{erfc}(\widehat{\tau}'\sqrt{1-\alpha^2}/\alpha)$$
$$\leq\frac{1}{4}\sqrt{1-\alpha^2}\cdot\exp(-\widehat{\tau}'^2/2\alpha^2)$$
$$\leq\frac{\|v-v'\|}{4\sqrt{2}\|v'\|}\cdot\frac{|\alpha|\sqrt{2}}{\widehat{\tau}'}\leq\frac{\|v-v'\|}{4\tau},$$

where the first step is standard Gaussian integration, the second step uses the inequality $\operatorname{erfc}(z)\leq e^{-z^2/2}$ for all $z\geq 0$, and the third step uses the fact that $\exp(-x)\leq 1/x$ for all $x>0$ and the fact that $\sqrt{1-\alpha^2}=\frac{1}{\sqrt{2}}\|\widehat{v}-\widehat{v}'\|\leq\frac{\|v-v'\|}{\sqrt{2}\|v'\|}$.

Case 2: $\widehat{\tau}<\gamma\leq\widehat{\tau}'/\alpha$.

We can naively upper bound the probability $\widehat{\tau}<\gamma\leq\widehat{\tau}'/\alpha$ and $\gamma'\leq\widehat{\tau}'$ by the probability $\widehat{\tau}<\gamma\leq\widehat{\tau}'/\alpha$, which is at most $e^{-\widehat{\tau}^2/2}\cdot(\widehat{\tau}'/\alpha-\widehat{\tau})$. Note that

$$\widehat{\tau}'/\alpha-\widehat{\tau}\leq\tau\cdot\left(\frac{1/\alpha}{\|v'\|}-\frac{1}{\|v\|'+\|v-v'\|}\right)\leq\frac{\tau}{\alpha}\cdot\frac{(1-\alpha)\|v'\|+\|v-v'\|}{\|v'\|^2}\leq\frac{3\tau\|v-v'\|}{2\alpha\|v'\|^2},\tag{14}$$

where in the last step we have used that $1-\alpha=\frac{1}{2}\|\widehat{v}-\widehat{v}'\|\leq\frac{\|v-v'\|}{2\|v'\|}$.

Suppose to the contrary that $e^{-\widehat{\tau}^2/2}\cdot(\widehat{\tau}'/\alpha-\widehat{\tau})>\frac{9\|v-v'\|}{\tau}$ so that by (14),

$$e^{\widehat{\tau}^2/2}<\frac{\tau^2}{6\alpha\|v'\|^2}.\tag{15}$$

Recall that we may assume that $\|v'\|\geq\|v-v'\|$, so $\widehat{\tau}\geq\frac{\tau}{2\|v'\|}$, and that $\alpha\geq 1/2$. From this, (15) would imply that $e^{\frac{\tau^2}{8\|v'\|^2}}<\frac{\tau^2}{3\|v'\|^2}$, and such an inequality cannot hold. □

We are now ready to prove Lemma 5.6.

Proof of Lemma 5.6. The left-hand side of (12) is at most

$$\mathop{\mathbb{P}}_{x\sim\mathcal{N}(0,\mathrm{Id})}[g(x)-f(x)>\tau\wedge g'(x)-f(x)\leq\tau]+\mathop{\mathbb{P}}_{x\sim\mathcal{N}(0,\mathrm{Id})}[g(x)-f(x)<-\tau\wedge g'(x)-f(x)\geq-\tau],\tag{16}$$

and by symmetry it suffices to upper bound the former probability on the right-hand side of (16) by $O(\eta m^2/\tau)$.

By definition of (m,η)-structural-closeness, we can express g and g' as $\max_j\min_{i\in\mathcal{I}_j}\langle u_i,\cdot\rangle$ and $\max_j\min_{i\in\mathcal{I}_j}\langle u_i',\cdot\rangle$ respectively, for vectors $\{u_i\}_{i\in[m]}$ and $\{u_i'\}_{i\in[m]}$ for which $\|u_i-u_i'\|\leq\eta$ for all i.

We proceed via a hybrid argument. Take any $0\leq i\leq m$. Let $u_1^{(i)},\ldots,u_{i-1}^{(i)}$ be $u_1,\ldots,u_{i-1}$, and let $u_i^{(i)},\ldots,u_m^{(i)}$ be the vectors $u_i',\ldots,u_m'$. Define the function $g^{(i)}=\max_a\min_{b\in\mathcal{I}_a}\langle u_i^{(i)},x\rangle$ so that $g^{(0)}(x)=\max_a\min_{b\in\mathcal{I}_a}\langle u_b',x\rangle$ and $g^{(m)}(x)=\max_a\min_{b\in\mathcal{I}_a}\langle u_b,x\rangle$.

We claim that for any x, $g^{(i-1)}(x)$ and $g^{(i)}(x)$ are sandwiched between $\langle u_i',x\rangle$ and $\langle u_i,x\rangle$, in the sense that

$$\langle u_i',x\rangle\geq g^{(i-1)}(x)\geq g^{(i)}(x)\geq\langle u_i,x\rangle\qquad\text{or}\qquad\langle u_i',x\rangle\leq g^{(i-1)}(x)\leq g^{(i)}(x)\leq\langle u_i,x\rangle.\tag{17}$$

This would imply

$$\mathbb{P}[g^{(i)}(x)-f(x)>\tau\wedge g^{(i-1)}(x)-f(x)\leq\tau]\leq\mathbb{P}[\langle u_i,x\rangle-f(x)>\tau\wedge\langle u_i',x\rangle-f(x)\leq\tau]\tag{18}$$

because either the left-hand side of (18) is zero, or or the event on the left-hand side immediately implies the one on the right-hand side.

Denote by $\{(\langle w_i, \cdot\rangle, S_i)\}_{i\in[m]}$ the pieces of some realization of f. We would then have

$$\begin{aligned}
&\mathbb{P}[g(x) - f(x) > \tau \wedge g'(x) - f(x) \le \tau] \\
&\quad\le \sum_{i=1}^{m} \mathbb{P}[\langle u_i, x\rangle - f(x) > \tau \wedge \langle u_i', x\rangle - f(x) \le \tau] \\
&\quad= \sum_{\ell=1}^{m}\sum_{i=1}^{m} \mathbb{P}[x \in S_\ell \wedge \langle u_i - w_\ell, x\rangle > \tau \wedge \langle u_i' - w_\ell, x\rangle \le \tau] \\
&\quad\le \sum_{\ell=1}^{m}\sum_{i=1}^{m} \mathbb{P}[\langle u_i - w_\ell, x\rangle > \tau \wedge \langle u_i' - w_\ell, x\rangle \le \tau] \le O\left(\eta m^2/\tau\right),
\end{aligned}$$

where the first step follows by triangle inequality and (18), and the last step follows by Lemma 5.7.

To complete the proof, we now turn to proving that the quantities $g^{(i)}(x)$ and $g^{(i-1)}(x)$ are sandwiched between $\langle u_i', x\rangle$ and $\langle u_i, x\rangle$, which will imply (18). Suppose that $g^{(i-1)}(x) = \langle u_j^{(i-1)}, x\rangle$ for some index j.

Case 1: $\langle u_i', x\rangle \ge \langle u_j^{(i-1)}, x\rangle$.

In this case $\min_{b\in\mathcal{I}_a}\langle u_b^{(i-1)}, x\rangle \le \langle u_i', x\rangle$ for all a. If $\langle u_i, x\rangle \ge \langle u_j', x\rangle$, then changing u_i' to u_i will not change the values of any of the clauses. So suppose $\langle u_i, x\rangle < \langle u_j', x\rangle$, in which case the value of the function cannot increase. Then if index i appears in any clause $\mathcal{I}_a$ for which $\min_{b\in\mathcal{I}_a}\langle u_b^{(i-1)}, x\rangle = \langle u_j^{(i-1)}, x\rangle$, then $g^{(i)}(x) \ge \langle u_i, x\rangle$. Otherwise, the value of the function stays the same. We conclude that the first inequality in (17) holds.

Case 2: $\langle u_i', x\rangle < \langle u_j^{(i-1)}, x\rangle$.

In this case there is some $\mathcal{I}_a$ for which $\langle u_j^{(i-1)}, x\rangle = \min_{b\in\mathcal{I}_a}\langle u_b^{(i-1)}, x\rangle$ and in which index i does not appear. If $\langle u_i, x\rangle \le \langle u_i', x\rangle$, then changing u_i' to u_i will not change the value of this $\mathcal{I}_a$ clause, and the values of the other clauses will not increase, so the value of the function will not change. So suppose $\langle u_i, x\rangle > \langle u_i', x\rangle$. Changing u_i' to u_i will not affect any clause $\mathcal{I}_a$ not containing i or for which $\min_{b\in\mathcal{I}_a}\langle u_b^{(i-1)}, x\rangle \le u_i'$. For all other clauses, their value will either stay the same or increase to u_i, in which case $g^{(i)}(x) \le \langle u_i, x\rangle$. We conclude that the second inequality in (17) holds. □

5.4 Netting Over Piecewise Linear Functions

Suppose we have recovered an ℓ-dimensional subspace $\widetilde{W}$ that approximately lies within V. In this section we show how to produce a finite list of candidate kickers with relevant subspace $\widetilde{W}$, one of which is guaranteed to approximate F restricted to some ℓ-dimensional subspace W. Ignoring the finiteness of this list for now, we first show that as long as $\widetilde{W}$ is sufficiently close to lying within V, there exists *some* kicker close to *some* restriction $F|_W$.

Lemma 5.8. *Let $\widetilde{w}_1, \dots, \widetilde{w}_\ell$ be a frame ν-nearly within V, with span $\widetilde{W}$. There exist an ℓ-dimensional subspace $W \subset V$ and a Λ-Lipschitz kicker $\widetilde{F}^*$ with relevant subspace $\widetilde{W}$ which is $(M, 2\sqrt{\nu}\cdot\ell\Lambda)$-structurally-close to $F|_W$.*

Proof of Lemma 5.8. By Lemma 3.16, there exist orthonormal vectors $w_1, \dots, w_\ell$ for which $\|w_i - \widetilde{w}_i\| \le 2\sqrt{\nu\ell}$. Let W be their span.

The function $F|_W$ is a continuous piecewise-linear function with at most M pieces, so by Theorem 4.9 and Lemma 4.5, there exist vectors $u_1, \dots, u_M \in W$ and subsets $\mathcal{I}_1, \dots, \mathcal{I}_m \subseteq [M]$

for which $F(x) = \max_{j\in[m]} \min_{i\in\mathcal{I}_j} \langle u_i, x\rangle$ and $\|u_i\| \le \Lambda$ for all i. For any $i \in [M]$, write $u_i = \sum_{i'\in[\ell]} \alpha_{i,i'} w_{i'}$. Define $\widetilde{u}_i^* \triangleq \sum_{i'\in[\ell]} \alpha_{i,i'} \widetilde{w}_{i'}$ and define the kicker $\widetilde{F}^*$ with relevant subspace $\widetilde{W}$ by $\widetilde{F}^*(x) \triangleq \max_{j\in[m]} \min_{i\in\mathcal{I}_j} \langle \widetilde{u}_i^*, x\rangle$.

Note that for any i,

$$\|\widetilde{u}_i^* - u_i\| = \sum_{i'\in[\ell]} \alpha_{i,i'} \|\widetilde{w}_{i'} - w_{i'}\| \le 2\sqrt{\nu\ell} \cdot \sum_{i'} |\alpha_{i,i'}| \le 2\sqrt{\nu} \cdot \ell \|u_i\| \le 2\sqrt{\nu} \cdot \ell\Lambda,$$

where the penultimate step is by Cauchy-Schwarz, so $\widetilde{F}^*$ is $(M, 2\sqrt{\nu} \cdot \ell\Lambda)$-structurally-close to $F|_W$ as claimed. Lastly, note that $\|\widetilde{u}_i^*\| = \|u_i\| \le \Lambda$, so $\widetilde{F}^*$ is indeed Λ-Lipschitz. □

We now show that the existential guarantee of Lemma 4.14 implies that if we enumerate over a fine enough net of kickers, then we can recover an approximation to $\widetilde{F}^*$ from Lemma 5.8 in time singly exponential in poly(M).

Algorithm 1: EnumerateKickers$(\widetilde{W}, \varepsilon')$

Input: Subspace $\widetilde{W}$ spanned by orthonormal vectors $\widetilde{w}_1, \dots, \widetilde{w}_\ell$, granularity $\varepsilon' > 0$
Output: List of kickers $\widetilde{F}$ with relevant subspace $\widetilde{W}$

1 $\mathcal{L} \leftarrow \emptyset$.
2 Let $\mathcal{N}$ be an $\varepsilon'\Lambda$-net over the set of vectors in $\widetilde{W}$ with norm at most Λ.
3 **for** $\widetilde{u}_1, \dots, \widetilde{u}_M \in \mathcal{N}$ **do**
4 **for** *functions* $A : \Omega_M \to [M]$ **do**
5 Let $\widetilde{F}$ be the kicker given by

$$\widetilde{F}(x) = \sum_{\omega\in\Omega_M} \mathbb{1}\left[\{\langle \widetilde{u}_i, x\rangle\}_{i\in[M]} \vdash \omega\right] \cdot \langle \widetilde{u}_{A(\omega)}, x\rangle.$$

  Append $\widetilde{F}$ to $\mathcal{L}$.
6 **return** $\mathcal{L}$.

Lemma 5.9. *Take any $\varepsilon' > 0$. Given a frame $\widetilde{w}_1, \dots, \widetilde{w}_\ell$ with span $\widetilde{W}$, for any Λ-Lipschitz kicker $\widetilde{F}^*$ with relevant subspace $\widetilde{W}$, there exists a kicker $\widetilde{F}$ with relevant subspace $\widetilde{W}$ in the output $\mathcal{L}$ of* EnumerateKickers$(\widetilde{W}, \varepsilon')$ *which is $(M, \varepsilon'\Lambda)$-structurally-close to $\widetilde{F}$. Furthermore, $|\mathcal{L}| \le M^{M^2} \cdot (1 + 2/\varepsilon')^\ell$.*

In particular, if $\widetilde{w}_1, \dots, \widetilde{w}_\ell$ is a frame ν-nearly within V, then for $\varepsilon' = 2\sqrt{\nu} \cdot \ell$, $\mathcal{L}$ contains a kicker $\widetilde{F}$ which is $(M, C_{\mathsf{piecewise}}\sqrt{\nu})$-structurally-close to $F|_W$ for some ℓ-dimensional subspace $W \subseteq V$, where

$$C_{\mathsf{piecewise}} \triangleq 4k\Lambda.$$

Furthermore, $|\mathcal{L}| \le M^{M^2} O(1/\sqrt{\nu})^\ell$ in this case.

Proof. By Lemma 4.14, the function $\widetilde{F}^*$ in the hypothesis can be written in the form $\widetilde{F}^*(x) = \sum_{\omega\in\Omega_M} \mathbb{1}\left[\{\langle \widetilde{u}_i^*, x\rangle\}_{i\in[M]} \vdash \omega\right] \cdot \langle \widetilde{u}_{A(\omega)}^*, x\rangle$ for some vectors $\{\widetilde{u}_i^*\}_{i\in[M]}$ and function $A : \Omega_M \to [M]$.

Because $\mathcal{N}$ in Step 2 of EnumerateKickers is an $\varepsilon'\Lambda$-net over the set of vectors in $\widetilde{W}$ with norm at most Λ, there exist vectors $\widetilde{u}_1, ..., \widetilde{u}_M \in \mathcal{N}$ for which $\|\widetilde{u}_i - \widetilde{u}_i^*\| \le \varepsilon'\Lambda$. If we define $\widetilde{F}$ by

$\widetilde{F}(x) = \sum_{\omega \in \Omega_M} \mathbb{1}\left[\{\langle \widetilde{u}_i, x\rangle\}_{i\in[M]} \vdash \omega\right] \cdot \langle \widetilde{u}_{A(\omega)}, x\rangle$, then by design, $\widetilde{F}$ is $(M, \varepsilon'\Lambda)$-structurally-close to $\widetilde{F}$.

It remains to bound the size of $\mathcal{L}$. For any $\varepsilon' > 0$ there is an ε'-net $\mathcal{N}'_{\varepsilon'}$ for the L_2 unit ball in $\widetilde{W}$ of size at most $(1+2/\varepsilon')^\ell$. Define $\mathcal{N} \triangleq \Lambda \cdot \mathcal{N}'_{\varepsilon'}$. Furthermore, there are $|\Omega_M|^M \le M^{M^2}$ functions $A : \Omega_M \to [M]$. This yields the desired bound on $|\mathcal{L}|$.

The final part of the lemma follows by invoking Lemma 5.8 and noting that the lattice polynomial representation of $\widetilde{F}^*$ and that of $F|_W$ are identical in the proof of Lemma 5.8, so the structural closeness of $\widetilde{F}$ to $F|_W$ follows by triangle inequality. □

5.5 Netting Over Neural Networks

Enumerating over arbitrary kickers with M pieces requires runtime scaling exponentially in $\mathrm{poly}(M)$. For ReLU networks of size S, M could be as large as $\exp(S)$, so naively using EnumerateKickers in our application to learning ReLU networks would incur doubly exponential dependence on k in the runtime. In this section we show how to enumerate over ReLU networks more efficiently. We first prove the analogue of Lemma 5.8 for ReLU networks.

Lemma 5.10. *Suppose F additionally satisfies Assumption 2. Let $\widetilde{w}_1, \dots, \widetilde{w}_\ell$ be a frame ν-nearly within V, with span $\widetilde{W}$. There exist an ℓ-dimensional subspace $W \subset V$ and weight matrix $\mathbf{W}_0^* \in \mathbb{R}^{k_0 \times d}$ with rows in $\widetilde{W}$ for which*

$$\|\mathbf{W}_0 \Pi_W - \mathbf{W}_0^*\|_{op} \le 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B \tag{19}$$

and for which $\|\mathbf{W}_0^*\|_{op} \le B$.

Proof. As in the proof of Lemma 5.8, Lemma 3.16 yields orthonormal vectors $w_1, \dots, w_\ell$ for which $\|w_i - \widetilde{w}_i\| \le 2\sqrt{\nu\ell}$. Let W be their span.

If F has weight matrices $\mathbf{W}_0 \in \mathbb{R}^{k_0 \times d}, \mathbf{W}_1 \in \mathbb{R}^{k_1 \times k_0}, \dots, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L}$, then $F|_W$ is a ReLU network with weight matrices $\mathbf{W}_0\Pi_W, \mathbf{W}_1, \dots, \mathbf{W}_{L+1}$. Denoting the rows of $\mathbf{W}_0\Pi_W \in \mathbb{R}^{k_0 \times d}$ as $u_1, \dots, u_{k_0}$, we may write them as $u_i = \sum_{i' \in [\ell]} \alpha_{i,i'} w_{i'}$ for $i \in [k_0]$.

Define $\widetilde{u}_i^* \triangleq \sum_{i' \in [\ell]} \alpha_{i,i'} \widetilde{w}_{i'}$. As in the proof of Lemma 5.8, we have that

$$\|\widetilde{u}_i^* - u_i\| \le 2\sqrt{\nu} \cdot \ell \|u_i\| \le 2\sqrt{\nu} \cdot \ell B,$$

where in the last step we have used the fact that the maximum norm of any row of $\mathbf{W}_0\Pi_W$ is at most the maximum norm of any row of $\mathbf{W}_0$, which is upper bounded by $\|\mathbf{W}_0\|_{\mathrm{op}} \le B$.

Let $\widetilde{\mathbf{W}}_0^*$ denote the matrix whose rows consist of $\widetilde{u}_1^*, \dots, \widetilde{u}_{k_0}^*$. We have that

$$\|\mathbf{W}_0\Pi_W - \widetilde{\mathbf{W}}_0^*\|_{\mathrm{op}} \le \|\mathbf{W}_0\Pi_W - \widetilde{\mathbf{W}}_0^*\|_F \le 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B$$

as claimed. Finally, the bound on $\|\mathbf{W}_0^*\|_{\mathrm{op}}$ follows from the fact that $\mathbf{W}_0^* = \mathbf{W}_0 \cdot \mathbf{O} \cdot \Pi_W$ for an orthogonal matrix $\mathbf{O}$ mapping the frame $\{w_1, ..., w_\ell\}$ to $\{\widetilde{w}_1, ..., \widetilde{w}_\ell\}$. □

We can now show the analogue of Lemma 5.9 for ReLU networks.

Lemma 5.11. *Take any $0 < \varepsilon' \le B$ and any frame $\widetilde{w}_1, \dots, \widetilde{w}_\ell$ with span $\widetilde{W}$. For any ReLU network $\widetilde{F}^*$ of size S with relevant subspace $\widetilde{W}$ and depth $L+2$ whose weight matrices have operator norm at most B, there exists a ReLU network $\widetilde{F}$ with relevant subspace $\widetilde{W}$ in the output $\mathcal{L}$ of* EnumerateNetworks*$(\widetilde{W}, \varepsilon')$ which is $(2^S, 2^{O(L)}B^{L+1}\varepsilon')$-structurally-close (as a piecewise-linear function) to $\widetilde{F}$. Furthermore, $|\mathcal{L}| \le 2^{O(S)} \cdot (1 + 4B/\varepsilon')^{O(S^2)}$.*

Algorithm 2: EnumerateNetworks($\widetilde{W}$, ε')

Input: Subspace $\widetilde{W}$ spanned by orthonormal vectors $\widetilde{w}_1, \dots, \widetilde{w}_\ell$, granularity $\varepsilon' > 0$
Output: List of size-S ReLU networks $\widetilde{F}$ with relevant subspace $\widetilde{W}$

$\mathcal{L} \leftarrow \emptyset$.
for *tuples* $(\widetilde{k}_0, ..., \widetilde{k}_{L+1}) \in \mathbb{Z}_{>0}^{L+2}$ *satisfying* $\sum_{i=0}^{L+1} \widetilde{k}_i = S$ **do**
  For every $0 \le i \le L+1$, let $\mathcal{N}_i$ be an ε'-net (in operator norm) over the set of matrices in $\mathbb{R}^{\widetilde{k}_i \times \widetilde{k}_{i-1}}$ with operator norm at most $B + \varepsilon'$.
  for $\widetilde{\mathbf{W}}_0 \in \mathcal{N}_0, ..., \widetilde{\mathbf{W}}_{L+1} \in \mathcal{N}_{L+1}$ **do**
    Define the ReLU network $\widetilde{F}$ with weight matrices $\mathsf{clip}_{\varepsilon'}(\mathbf{W}_0), ..., \mathsf{clip}_{\varepsilon'}(\mathbf{W}_{L+1})$.
    Append $\widetilde{F}$ to $\mathcal{L}$.
return $\mathcal{L}$.

In particular, if $\widetilde{w}_1, ..., \widetilde{w}_\ell$ is a frame ν-nearly within V, then for $\varepsilon' = 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B$, $\mathcal{L}$ contains a ReLU network $\widetilde{F}$ which is $(M, C_{\mathsf{network}}\sqrt{\nu})$-structurally-close to $F|_W$ for some ℓ-dimensional subspace $W \subseteq V$, where

$$C_{\mathsf{network}} \triangleq 2^{O(L)} B^{L+2} k^{3/2}$$

Furthermore, $|\mathcal{L}| \le O(1/\sqrt{\nu})^{O(S^2)}$ in this case.

Proof. Let $\mathbf{W}'_0 \in \mathbb{R}^{k'_0 \times d}, \dots, \mathbf{W}_{L+1} \in \mathbb{R}^{1 \times k_L}$ denote the weight matrices of $\widetilde{F}^*$. Consider the iteration of the outer loop of EnumerateNetworks in which the architecture of $\widetilde{F}^*$ is guessed correctly, that is, for which $\widetilde{k}_i = k'_i$ for all $0 \le i \le L+1$. By the choice of nets, there is some iteration of the inner loop of the algorithm for which the weight matrices $\{\widetilde{\mathbf{W}}_i\}$ satisfy

$$\|\mathbf{W}'_i - \widetilde{\mathbf{W}}_i\|_{\mathrm{op}} \le \varepsilon' \ \forall \ 0 \le i \le L+1. \tag{20}$$

Define the ReLU network $\widetilde{F}$ with relevant subspace $\widetilde{W}$ to have weight matrices $\widetilde{\mathbf{W}}_0, \widetilde{\mathbf{W}}_1, \dots, \widetilde{\mathbf{W}}_{L+1}$. By the fact that operator norm closeness implies entrywise closeness, together with Fact 3.1 and Theorem 4.15, there are lattice polynomial representations for $\widetilde{F}^*$ and $\widetilde{F}$ with identical clauses, and for which the vectors at the leaves consist of $\mathbf{W}'_{L+1}\mathbf{\Sigma}_L\mathbf{W}'_L \cdots \mathbf{\Sigma}_0\mathbf{W}'_0\Pi_W$ and $\widetilde{\mathbf{W}}_{L+1}\mathbf{\Sigma}_L\widetilde{\mathbf{W}}_L \cdots \mathbf{\Sigma}_0\widetilde{\mathbf{W}}_0$ respectively for all possible diagonal matrices $\mathbf{\Sigma}_i \in \{0,1\}^{k'_i \times k'_i}$. For any such choice of matrices $\{\mathbf{\Sigma}_i\}$, note that

$$\begin{aligned}
&\|\mathbf{W}'_{L+1}\mathbf{\Sigma}_L\mathbf{W}'_L \cdots \mathbf{W}'_0 - \widetilde{\mathbf{W}}_{L+1}\mathbf{\Sigma}_L\widetilde{\mathbf{W}}_L \cdots \widetilde{\mathbf{W}}_0\| \\
&\quad \le \|(\mathbf{W}'_{L+1} - \widetilde{\mathbf{W}}_{L+1})\mathbf{\Sigma}_L\mathbf{W}'_L \cdots \mathbf{W}'_0\| + \cdots + \|\widetilde{\mathbf{W}}_{L+1}\mathbf{\Sigma}_L\widetilde{\mathbf{W}}_L \cdots (\mathbf{W}'_0 - \widetilde{\mathbf{W}}_0)\| \\
&\quad \le \|\mathbf{W}'_{L+1} - \widetilde{\mathbf{W}}_{L+1}\| \prod_{i=0}^{L} \|\mathbf{W}'_i\|_{\mathrm{op}} + \cdots + \prod_{i=1}^{L+1} \|\widetilde{\mathbf{W}}_i\|_{\mathrm{op}} \|\mathbf{W}'_0 - \widetilde{\mathbf{W}}_0\|_{\mathrm{op}} \\
&\quad \le (L+2) \cdot (B + \varepsilon')^{L+1} \cdot \varepsilon' \\
&\quad \le 2^{O(L)} B^{L+1} \cdot \varepsilon',
\end{aligned} \tag{21}$$

where in the last step we used the assumption that $\varepsilon' \le B$. This implies the claim about structural closeness.

We next bound the size of $|\mathcal{L}|$. For any choice of $\widetilde{k}_0, ..., \widetilde{k}_{L+1}$, note that by Corollary 3.9,

$$\left|\mathcal{N}_{\widetilde{k}_0} \times \cdots \times \mathcal{N}_{\widetilde{k}_{L+1}}\right| \le (1 + 4B/\varepsilon')^{L\widetilde{k}_0 + \widetilde{k}_0\widetilde{k}_1 + \cdots + \widetilde{k}_L\widetilde{k}_{L+1} + \widetilde{k}_{L+1}} \le (1 + 4B/\varepsilon')^{O(S^2)}$$

where in the penultimate step we used that

$$L\widetilde{k}_0 + \widetilde{k}_0\widetilde{k}_1 + \cdots + \widetilde{k}_L\widetilde{k}_{L+1} + \widetilde{k}_{L+1} \le (L + \widetilde{k}_0 + \cdots + \widetilde{k}_{L+1})(\widetilde{k}_0 + \cdots + \widetilde{k}_{L+1} + 1) = (L+S)(S+1) \le O(S^2).$$

There are $\binom{S+L+1}{L+1} = 2^{O(S)}$ choices of $(\widetilde{k}_0, \ldots, \widetilde{k}_{L+1})$ in the outer loop of EnumerateNetworks, so $|\mathcal{L}| \le 2^{O(S)} \cdot (1 + 4B/\varepsilon')^{O(S^2)}$ as claimed.

Finally, to obtain the last part of the lemma, we can take $\widetilde{F}^*$ above to have the same weight matrices as F except for the input layer, which we will take to be $\mathbf{W}_0' \triangleq \widetilde{\mathbf{W}}_0^*$ for the weight matrix guaranteed by Lemma 5.10. By (19), this choice of $\mathbf{W}_0'$ is close to $\mathbf{W}_0\Pi_W$ for some subspace $W \subseteq V$. Take $\varepsilon' = 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B$. For $\{\widetilde{\mathbf{W}}_i\}$ satisfying (20), by triangle inequality (19) we get that

$$\|\mathbf{W}_0\Pi_W - \widetilde{\mathbf{W}}_0\|_{\mathrm{op}} \le \|\mathbf{W}_0\Pi_W - \mathbf{W}_0'\|_{\mathrm{op}} + \|\mathbf{W}_0' - \widetilde{\mathbf{W}}_0\|_{\mathrm{op}} \le 2\varepsilon'.$$

Using this, by a calculation analogous to the one leading to (21), we find that $\widetilde{F}$ is $(2^S, 2^{O(L)}B^{L+1}\varepsilon')$-structurally-close to $F|_W$, from which the claim follows by our choice of $\varepsilon' = 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B$. In this case, we get that $|\mathcal{L}| \le 2^{O(S)}(1 + 2/\sqrt{\nu})^{O(S^2)} \le O(1/\sqrt{\nu})^{O(S^2)}$ as claimed. □

With subroutines for enumerating over ReLU networks and kickers in hand, we can now formally state our algorithm, FilteredPCA (see Algorithm 3 below). The algorithm as stated applies to the case where F is a neural network satisfying Assumptions 1 and 2, but we can easily modify the algorithm to work in the case where F is only a kicker satisfying Assumption 1 by replacing the call to EnumerateNetworks$(\widetilde{W}, 2\sqrt{\nu_0} \cdot \ell\sqrt{k} \cdot B)$ in Line 9 with a call to EnumerateKickers$(\widetilde{W}, 2\sqrt{\nu_0} \cdot \ell)$, the call to EnumerateNetworks$(\widetilde{W}, B^{-L-1}2^{-\Omega(L)} \cdot \varepsilon/\sqrt{k})$ in Line 18 with a call to EnumerateKickers$(\widetilde{W}, \varepsilon/(2\sqrt{k}\Lambda))$, and the assignment $N' \leftarrow \mathrm{poly}(B^{L+2}, k, 1/\varepsilon) \cdot \log(1/\delta)$ in Line 19 with the assignment $N' \leftarrow \mathrm{poly}(\Lambda, k, 1/\varepsilon) \cdot \log(1/\delta)$).

5.6 Perturbation Bounds

We now show how to leverage Lemma 5.6 to show that even with access to a subspace $\widetilde{W}$ which is only approximately within V as well as the restriction of F to that subspace, we can recover another vector orthogonal to $\widetilde{W}$ which mostly lies within V.

The first step is to show that in this approximate setting, the analogue of $\mathbf{M}^W$ from Section 5.2 is spectrally close to $\mathbf{M}^W$. It is in showing this perturbation bound that we invoke the stability result of Section 5.3.

Lemma 5.12. *Suppose F only satisfies Assumption 1 (resp. both Assumptions 1 and 2). Let $\widetilde{w}_1, \ldots, \widetilde{w}_\ell \in \mathbb{S}^{d-1}$ be a frame ν-nearly within V, with span $\widetilde{W}$. For $* \in \{\mathsf{piecewise}, \mathsf{network}\}$, define*

$$\xi_*(\nu) \triangleq O\left(k\left(\frac{C_*\sqrt{\nu}M^2}{c\sqrt{k}\Lambda}\right)^{1-1/k} \vee \sqrt{\nu k}\right) \tag{23}$$

and suppose $N \ge \Omega(\{d \vee \log(1/\delta)\}/\xi_*^2)$.

Algorithm 3: FilteredPCA$(\mathcal{D}, \varepsilon, \delta)$

Input: Sample access to $\mathcal{D}$, target error ε, failure probability δ

Output: Size-S ReLU network $\widetilde{F} : \mathbb{R}^d \to \mathbb{R}$ for which $\|\widetilde{F} - F\| \le O(\varepsilon)$ with probability at least $1 - \delta$

$\mathcal{W} \leftarrow \emptyset$.

$\tau \leftarrow c\sqrt{k} \cdot \Lambda$ as in (10).

$\nu_0 \leftarrow \mathrm{poly}(k^k, 1/\underline{\lambda}^k, M^k, \Lambda)^{-1}$, where $\underline{\lambda}$ is defined in (11).

$\xi \leftarrow O\left(k\left(\sqrt{\nu_0 k} \cdot M^2/c\right)^{1-1/k}\right)$ as in (23).

$N \leftarrow \Omega(\{d \vee \log(2k/\delta)\}/\xi^2)$.

for $0 \le \ell \le k-1$ **do**

  Draw samples $(x_1, y_1), \ldots, (x_N, y_N) \sim \mathcal{D}$.

  If $\mathcal{W} = \{\widetilde{w}_1, \ldots, \widetilde{w}_\ell\}$, let $\widetilde{W}$ denote the span of these vectors.

  $\mathcal{L} \leftarrow$ EnumerateNetworks$(\widetilde{W}, 2\sqrt{\nu_0} \cdot \ell\sqrt{k} \cdot B)$.

  for $\widetilde{F} \in \mathcal{L}$ **do**

    Form the matrix

$$\widetilde{\mathbf{M}}_{\mathsf{emp}}^{\widetilde{W}} \triangleq \Pi_{\widetilde{W}^\perp}\left(\sum_{i=1}^{N} \mathbb{1}\left[|y_i - \widetilde{F}(\Pi_{\widetilde{W}} x)| > \tau\right] \cdot (x_i x_i^\top - \mathrm{Id})\right)\Pi_{\widetilde{W}^\perp}. \tag{22}$$

    Run ApproxBlockSVD$(\widetilde{\mathbf{M}}_{\mathsf{emp}}^{\widetilde{W}}, \underline{\lambda}/1000, \delta/(2|\mathcal{L}|k))$ to obtain approximate top singular vector $\widetilde{w}^{\ell+1}$.

    $\lambda \leftarrow (\widetilde{w}^{\ell+1})^\top \widetilde{\mathbf{M}}_{\mathsf{emp}}^{\widetilde{W}} \widetilde{w}^{\ell+1}$.

    if $\lambda \ge 9\underline{\lambda}/16$ **then**

      Append $\widetilde{w}^{\ell+1}$ to $\mathcal{W}$ and exit out of this inner loop and increment ℓ.

  if *no* $\widetilde{w}^{\ell+1}$ *was appended to* $\mathcal{W}$ **then**

    return $\mathcal{W}$.

Let $\widetilde{W}$ denote the span of the vectors in $\mathcal{W}$.

$\mathcal{L} \leftarrow$ EnumerateNetworks$(\widetilde{W}, B^{-L-1}2^{-\Omega(L)} \cdot \varepsilon/\sqrt{k})$.

$N' \leftarrow \mathrm{poly}(B^{L+2}, k, 1/\varepsilon) \cdot \log(1/\delta)$.

for $\widetilde{F} \in \mathcal{L}$ **do**

  Form an empirical estimate $\hat{\varepsilon}$ for $\|\widetilde{F} - F\|$ by drawing N' samples.

  if $\hat{\varepsilon} \le 3\varepsilon$ **then**

    return $\widetilde{F}$.

Given subspace $W \subseteq V$ and $\widetilde{F}$ for which $F|_W$ and $\widetilde{F}$ are $(M, C_{\mathsf{piecewise}}\sqrt{\nu})$-structurally-close (resp. $(M, C_{\mathsf{network}}\sqrt{\nu})$-structurally close), then we have that

$$\|\widetilde{\mathbf{M}}_{\mathsf{emp}}^{\widetilde{W}} - \mathbf{M}^{W}\|_{op} \le 3\xi(\nu)$$

with probability at least $1-\delta$.

Proof. For convenience denote $\widetilde{\mathbf{M}}_{\mathsf{emp}}^{\widetilde{W}}$ and $\mathbf{M}^W$ by $\widetilde{\mathbf{M}}_{\mathsf{emp}}$ and $\mathbf{M}$ respectively. Also, depending on whether F only satisfies Assumption 1 or both Assumptions 1 and 2, define $C_* \triangleq C_{\mathsf{piecewise}}$ or $C_* \triangleq C_{\mathsf{network}}$ respectively. It will also be convenient to define

$$\mathbf{M}' \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y}\Big[\mathbb{1}[|y - F|_W(x)| > \tau] \cdot (xx^\top - \mathrm{Id})\Big]\Pi_{W^\perp}$$

as well as the population version of $\widetilde{\mathbf{M}}_{\mathsf{emp}}$, that is, $\widetilde{\mathbf{M}} \triangleq \mathbb{E}_{(x_1,y_1),\ldots,(x_N,y_N)}[\widetilde{\mathbf{M}}_{\mathsf{emp}}]$.

We will upper bound

$$\|\widetilde{\mathbf{M}}_{\mathsf{emp}} - \mathbf{M}\|_{\mathrm{op}} \le \|\widetilde{\mathbf{M}}_{\mathsf{emp}} - \widetilde{\mathbf{M}}\|_{\mathrm{op}} + \|\widetilde{\mathbf{M}} - \mathbf{M}'\|_{\mathrm{op}} + \|\mathbf{M}' - \mathbf{M}\|_{\mathrm{op}}.$$

by upper bounding each of the summands on the right-hand side by ξ_*.

By Lemma 3.6 and our choice of N, $\|\widetilde{\mathbf{M}}_{\mathsf{emp}} - \widetilde{\mathbf{M}}\|_{\mathrm{op}} \le \xi_*$ with probability at least $1-\delta$.

To upper bound $\|\widetilde{\mathbf{M}} - \mathbf{M}'\|_{\mathrm{op}}$, we can naively upper bound

$$\left\| \mathop{\mathbb{E}}_{x,y}\Big[\mathbb{1}[|y - F|_W(x)| > \tau] \cdot (xx^\top - \mathrm{Id})\Big] \right\| \le 2,$$

so by Lemma 3.17 and Lemma 3.16 we have

$$\|\widetilde{\mathbf{M}} - \mathbf{M}'\|_{\mathrm{op}} \le 2\sqrt{2} \cdot d_C(\widetilde{W}, W) \le 4\sqrt{\nu \cdot k} \le \xi_*$$

Finally, we upper bound $\|\mathbf{M}' - \mathbf{M}\|_{\mathrm{op}}$. For any test vector $v \in \mathbb{S}^{d-1}$ orthogonal to W,

$$\begin{aligned}
v^\top(\mathbf{M} - \mathbf{M}')v &= \mathop{\mathbb{E}}_{x}\Big[\Big(\mathbb{1}[|y - F|_W(x)| > \tau] - \mathbb{1}[|y - \widetilde{F}(\Pi_{\widetilde{W}}x)| > \tau]\Big) \cdot (\langle v, x\rangle^2 - 1)\Big] \\
&\le \mathop{\mathbb{P}}_{x}\Big[\mathrm{sgn}(|y - F|_W(x)| - \tau) \neq \mathrm{sgn}(|y - \widetilde{F}(\Pi_{\widetilde{W}}x)| - \tau)\Big]^{1-1/k} \cdot O(k) \\
&\le O\left(k\left(\frac{C_*\sqrt{\nu}M^2}{\tau}\right)^{1-1/k}\right) = O\left(k\left(\frac{C_*\sqrt{\nu}M^2}{c\sqrt{k}\Lambda}\right)^{1-1/k}\right) \le \xi_*
\end{aligned}$$

where the second step follows by Holder's and the fact that $\mathbb{E}_{g\sim\mathcal{N}(0,1)}[(g^2-1)^k]^{1/k} \le O(k)$, and the third step follows by Lemma 5.6, which we may apply because $\widetilde{F}$ and $F|_W$ are $(M, 4\sqrt{\nu}\cdot\ell\Lambda)$-structurally-close. ∎

Finally, we use the above perturbation bound to show that in a single iteration of the main outer loop of FilteredPCA, if there is some variance unexplained by the subspace $\widetilde{W}$ found so far (see (24)), then we will find another "good" direction orthogonal to $\widetilde{W}$ which is also approximately within the span of V. Note that this claim has two components: *completeness*, i.e. in the list of candidate functions we have enumerated, there is *some* function for which the top singular vector of (22) is a good direction, and *soundness*, i.e. whatever direction is ultimately chosen in Step 14 of FilteredPCA is a good direction.

Lemma 5.13. *Suppose F only satisfies Assumption 1 (resp. both Assumptions 1 and 2). Suppose $\nu \le \varepsilon^2/(4kC_{\mathsf{piecewise}}^2)$ (resp. $\nu \le \varepsilon^2/(4kC_{\mathsf{network}}^2)$). For $0 \le \ell < k$, let $\widetilde{w}_1, \dots, \widetilde{w}_\ell$ be a frame ν-nearly within V, with span $\widetilde{W}$. Define $\xi = \xi_{\mathsf{piecewise}}(\nu)$ (resp. $\xi = \xi_{\mathsf{network}}(\nu)$) according to (23), and suppose $N \ge \Omega(\{d \vee \log(1/\delta)\}/\xi^2)$ and $\tau = c\sqrt{k} \cdot \Lambda$.*

Suppose $\xi \le \underline{\lambda}/6$, and suppose

$$\mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,\mathrm{Id})}[(F(x) - F(\Pi_{\widetilde{W}} x))^2] \ge \varepsilon^2. \tag{24}$$

Let $\mathcal{L}$ be the output of EnumerateKickers$(\widetilde{W}, 2\sqrt{\nu} \cdot \ell)$ *(resp.* EnumerateNetworks$(\widetilde{W}, 2\sqrt{\nu} \cdot \ell\sqrt{k} \cdot B)$*). With probability at least $1 - |\mathcal{L}| \cdot \delta$ over the randomness of the N samples, the following hold:*

1. ***Completeness**: There exists some $\widetilde{F} \in \mathcal{L}$ such that, if $\widetilde{\mathbf{M}}_{\mathsf{emp}}^{\widetilde{W}}$ is defined according to (22), its top singular value is at least $\underline{\lambda} - 3\xi$.*
2. ***Soundness**: For any $\widetilde{F} \in \mathcal{L}$ for which $\|\widetilde{\mathbf{M}}_{\mathsf{emp}}^{\widetilde{W}}\|_{op} \ge \underline{\lambda} - 3\xi$, the top singular vector w satisfies $\|\Pi_V w\| \ge 1 - c'\xi^2/\underline{\lambda}^2$ for some absolute constant $c' > 0$ and is orthogonal to $\widetilde{W}$.*

Proof. When the choice of $\widetilde{F}$ is clear from context, for convenience we will denote $\mathbf{M}^W$ and $\widetilde{\mathbf{M}}_{\mathsf{emp}}^{\widetilde{W}}$ by $\mathbf{M}$ and $\widetilde{\mathbf{M}}_{\mathsf{emp}}$ respectively.

By Lemma 5.9 (resp. Lemma 5.11) and our assumed bound on ν, there exists $\widetilde{F}$ in the output of EnumerateKickers (resp. EnumerateNetworks) which is $(M, \varepsilon/2k)$-structurally-close to $F|_W$ for some ℓ-dimensional subspace $W \subsetneq V$.

By triangle inequality, Lemma 4.11, and (24), and our assumed bounds on ν, we have that $\|F - F|_W\| \ge \varepsilon/2$. So by Lemma 5.5 and (11), we know $\|\mathbf{M}\| \ge \underline{\lambda}$.

Because this $\widetilde{F}$ is $(M, C_{\mathsf{piecewise}}\sqrt{\nu})$-structurally-close (resp. $(M, C_{\mathsf{network}}\sqrt{\nu})$-structurally close) to $F|_W$, Lemma 5.12 implies that with probability $1 - \delta$, $\|\mathbf{M} - \widetilde{\mathbf{M}}_{\mathsf{emp}}\|_{\mathrm{op}} \le 3\xi$, so $\widetilde{\mathbf{M}}_{\mathsf{emp}}$ has top singular value at least $\underline{\lambda} - 3\xi$. This proves completeness.

Now take any $\widetilde{F}$ for which $\|\widetilde{\mathbf{M}}_{\mathsf{emp}}\|_{\mathrm{op}} \ge \underline{\lambda} - 3\xi$. The fact that the top singular vector w is orthogonal to $\widetilde{W}$ is immediate. And by Lemma 5.12, with probability $1 - \delta$ over the samples, $\|\mathbf{M} - \widetilde{\mathbf{M}}_{\mathsf{emp}}\|_{\mathrm{op}} \le 3\xi$. So if we take $\lambda, \varepsilon, \mathbf{A}, \widehat{\mathbf{A}}$ in Corollary 3.12 to be $\underline{\lambda}$, 3ξ, $\mathbf{M}$, and $\widetilde{\mathbf{M}}_{\mathsf{emp}}$ respectively, then because $\xi \le \underline{\lambda}/6$, we get that the top singular vector w of $\widetilde{\mathbf{M}}_{\mathsf{emp}}$ satisfies $\|\Pi_V w\| \ge 1 - O(\xi^2/\underline{\lambda}^2)$. This proves soundness, upon union bounding over all $\widetilde{F} \in \mathcal{L}$. ∎

5.7 Putting Everything Together

To conclude the proof of Theorems 5.1 and 5.2, we first show that for the subspace $\widetilde{W}$ formed in Step 17, if $\widetilde{W}$ is sufficiently close to the true relevant subspace V or if (24) is violated, then one can run EnumerateKickers (resp. EnumerateNetworks) one more time to produce a function with small squared error relative to F.

Lemma 5.14. *Suppose F only satisfies Assumption 1 (resp. both Assumptions 1 and 2). Define*

$$\varepsilon^* \triangleq \varepsilon/(2\sqrt{k}\Lambda) \qquad \textit{(resp. } \varepsilon^* \triangleq B^{-L-1}2^{-\Omega(L)} \cdot \varepsilon/\sqrt{k}\textit{)}$$

Let $\widetilde{w}_1, ..., \widetilde{w}_\ell$ be a frame with span $\widetilde{W}$. If either 1) $\ell = k$ and this frame is $\varepsilon^2/4kC_{\mathsf{piecewise}}^2$-nearly (resp. $\varepsilon^2/4kC_{\mathsf{network}}^2$-nearly) within V, or 2) inequality (24) is violated. Then the output $\mathcal{L}$

of EnumerateKickers$(\widetilde{W}, \varepsilon^*)$ *(resp.* EnumerateNetworks$(\widetilde{W}, \varepsilon^*)$*) contains a function* $\widetilde{F}$ *for which* $\|F - \widetilde{F}\| \le O(\varepsilon)$*. Furthermore,* $|\mathcal{L}| \le M^{M^2} \cdot O(\Lambda/\varepsilon)^k$ *(resp.* $|\mathcal{L}| \le O(B^{L+2}2^{O(L)}/\varepsilon)^{O(S^2)}$*).*

In particular, if 1) or 2) holds for the subspace $\widetilde{W}$ *at the end of running* FilteredPCA*, then the output* $\widetilde{F}$ *of* FilteredPCA *satisfies* $\|F - \widetilde{F}\| \le O(\varepsilon)$*.*

Proof. We first show that if either 1) or 2) holds, then there exists $\widetilde{F}$ in $\mathcal{L}$ for which $\|\widetilde{F} - F\| \le O(\varepsilon)$.

Suppose 1) holds. If F only satisfies Assumption 1 (resp. Assumptions 1 and 2), then by the final part of Lemma 5.9 (resp. Lemma 5.11), there is a function $\widetilde{F}$ in $\mathcal{L}$ which is $(M, \varepsilon/2k)$-structurally-close (resp. $(2^S, \varepsilon/2k)$-structurally-close) to $F|_W$ for ℓ-dimensional subspace $W \subseteq V$. Because $\ell = k$ when 1) holds, this subspace must be V, so in fact $F|_W = F$ and therefore $\widetilde{F}$ is structurally-close to F. By Lemma 4.11, we conclude that $\|\widetilde{F} - F\| \le \varepsilon$.

Suppose 2) holds. If F only satisfies Assumption 1 (resp. Assumptions 1 and 2), then we can take $\widetilde{F}^*$ in the first part of Lemma 5.9 (resp. Lemma 5.11) to be the function $x \mapsto F(\Pi_{\widetilde{W}} x)$, which is clearly also a Λ-Lipschitz kicker (resp. ReLU network of size S whose weight matrices have operator norm at most B) with relevant subspace $\widetilde{W}$. It follows that $\mathcal{L}$ contains some function $\widetilde{F}$ which is $(M, \varepsilon/(2\sqrt{k}))$- (resp. $(2^S, \varepsilon/(2\sqrt{k}))$-structurally-close to $\widetilde{F}^*$. By Lemma 4.11, we conclude that $\|\widetilde{F} - F\| \le 3\varepsilon/2$.

For the last part of the lemma, note that by Lemma A.1 in Appendix A.1 that for any function $\widetilde{F}$ for which $\|\widetilde{F} - F\|^2 \le \mu$, we can estimate $\|\widetilde{F} - F\|^2$ to error $O(\varepsilon^2)$ from $O((\mu + \Lambda^2 k)\log(1/\delta)/\varepsilon^4)$ samples (resp. $O((\mu + B^{2L+4}k)\log(1/\delta)/\varepsilon^2)$). Note that for any $\widetilde{F} \in \mathcal{L}$, by the second part of Lemma 4.11 we have that $\|\widetilde{F} - F\| \le O(\Lambda\sqrt{k})$ (resp. $\|\widetilde{F} - F\| \le O(B^{L+2}\sqrt{k})$). ∎

We can now conclude the proof of correctness for FilteredPCA.

Proof of Theorem 5.1. First note that the only randomness in FilteredPCA comes from calling ApproxBlockSVD and drawing samples, so henceforth we will condition on the event that the former always succeeds and on the success of Lemma 3.6 for every batch of samples drawn in Step 7 of FilteredPCA. By our choice of parameters in FilteredPCA and a union bound, this event happens with probability at least $1 - \delta$.

If F satisfies Assumption 1 only (resp. both Assumptions 1 and 2), let $\xi(\nu) = \xi_{\mathsf{piecewise}}(\nu)$ and $C_* = C_{\mathsf{piecewise}}$ (resp. $\xi(\nu) = \xi_{\mathsf{network}}(\nu)$ and $C_* = C_{\mathsf{network}}$), recalling the definition from (23).

Call $\nu \ge 0$ *admissible* if $\nu \le \varepsilon^2/4kC_*^2$ and $\xi(\nu) \le \underline{\lambda}/6$. Let $\iota : \mathbb{R} \to \mathbb{R}$ be the function given by $\iota(\nu) = c'\xi(\nu)^2/\underline{\lambda}^2$, where c' is the absolute constant in Lemma 5.13. Note that if we define

$$\beta \triangleq (c'/\underline{\lambda}^2) \cdot O\left(k^2 \cdot \left(\frac{C_*^2 M^4}{c^2 k \Lambda^2}\right)^{1-1/k}\right),$$

then $\iota(\nu) = (\beta \cdot \nu^{1-1/k}) \vee (k\nu)$.

Because we are conditioning on every invocation of ApproxBlockSVD succeeding, the quantity λ computed in Step 12 is certainly $\xi(\nu)/2$-close to the true top singular value of $\widetilde{\mathbf{M}}^{\widetilde{W}}$. So Lemma 5.13 tells us that in any iteration ℓ of the main loop in FilteredPCA, if $\{\widetilde{w}_1, \dots, \widetilde{w}_\ell\}$ is a frame ν-nearly within V for admissible ν, then either 1) we reach Line 14 in the inner loop and append some $\widetilde{w}_{\ell+1}$ for which $\{\widetilde{w}_1, \dots, \widetilde{w}_{\ell+1}\}$ is a frame $\iota(\nu)$-nearly within V, or 2) (24) is violated, in which case condition 2) of Lemma 5.14 implies that FilteredPCA would output a function $\widetilde{F}$ for which $\|F - \widetilde{F}\| \le O(\varepsilon)$.

So all we need to verify is that there is a choice of ν_0 for which the k numbers

$$\nu_0, \iota(\nu_0), \dots, \underbrace{\iota(\iota(\cdots \iota}_{k-1}(\nu_0)\cdots)) \tag{25}$$

are all admissible, after which we can invoke condition 1) of Lemma 5.14 to conclude that FilteredPCA outputs a function $\widetilde{F}$ for which $\|F - \widetilde{F}\| \leq O(\varepsilon)$. It is clear that for ν sufficiently small, ι is increasing in ν. So it suffices to choose ν_0 sufficiently small that the last number in the sequence (25) is admissible.

Then the last number in (25) is at most

$$\left(\beta^{\sum_{j=0}^{k-1}(1-1/k)^j} \cdot \nu_0^{(1-1/k)^k}\right) \vee (k^k \nu_0) \leq \left(\beta^k \cdot \nu_0^{1/e}\right) \vee (k^k \nu_0).$$

If F satisfies Assumption 1 only and we take $C_* = C_{\mathsf{piecewise}}$, then

$$\beta^k = (c'/\underline{\lambda}^2)^k \cdot O\left(k^{2k} \cdot (kM^4/c^2)^{k-1}\right),$$

so for

$$\nu_0 \triangleq \mathrm{poly}(k^k, 1/\underline{\lambda}^k, M^k, \Lambda/\varepsilon)^{-1} = \mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, M^k)^{-1}$$

sufficiently small, we have that $(\beta^k \cdot \nu_0^{1/e}) \vee (k^k \nu_0)$ is admissible.

And because in each of the at most k iterations of the main loop of FilteredPCA,

$$N = O(\{d \vee \log(Mk/\delta)\}/\xi(\nu_0)^2) \leq d\log(1/\delta)\mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, M^k)$$

samples are drawn, the final sample complexity is $d\log(1/\delta)\mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, M^k)$ as claimed. The runtime is dominated by the at most $M^{M^2}O(1/\sqrt{\nu_0})^\ell = M^{M^2} \cdot \mathrm{poly}(e^{k^4\Lambda^2/\varepsilon^2}, M^{k^2})$ calls to ApproxBlockSVD, one for each element of $\mathcal{L}$ output by EnumerateKickers, (note that the runtime and sample complexity cost of running EnumerateKickers at the very end is of much lower order). As there is a matrix-vector oracle for the matrices on which we run ApproxBlockSVD which takes time $O(d^2)$, by Fact 3.10 each of these calls takes, up to lower order factors that will be absorbed elsewhere, $\widetilde{O}(d^2\log(1/\delta))$ time, so we conclude that FilteredPCA runs in time

$$\widetilde{O}(d^2\log(1/\delta)) \cdot M^{M^2} \cdot \mathrm{poly}(e^{k^4\Lambda^2/\varepsilon^2}, M^{k^2})$$

as claimed.

If F satisfies Assumptions 1 and 2 and we take $C_* = C_{\mathsf{network}}$, then

$$\beta^k = (c'/\underline{\lambda}^2)^k \cdot O\left(k^{2k}\left(\frac{2^{O(L)}B^{2L+4}k2^{4S}}{c^2\Lambda^2}\right)^{k-1}\right),$$

where we have used that $M \leq 2^S$ for size-S ReLU networks. So for

$$\nu_0 \triangleq \mathrm{poly}(k^k, 1/\underline{\lambda}^k, 2^{kS}, (B^{L+2}/\Lambda)^k, \Lambda/\varepsilon)^{-1} = \mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, 2^{kS}, (B^{L+2}/\Lambda)^k)$$

sufficiently small, we have that $(\beta^k \cdot \nu_0^{1/e}) \vee (k^k \nu_0)$ is admissible.

And because in each of the at most k iteration of the main loop of FilteredPCA,

$$N = O(\{d \vee \log(2^S k/\delta)\}/\xi(\nu_0)^2) \leq d\log(1/\delta)\mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, 2^{kS}, B^{(L+2)k}/\Lambda^k)$$

samples are drawn, the final sample complexity is $d\log(1/\delta)\mathrm{poly}(e^{k^3\Lambda^2/\varepsilon^2}, 2^{kS}, B^{(L+2)k}/\Lambda^k)$ as claimed. The runtime is dominated by the at most $O(1/\sqrt{\nu_0})^{O(S^2)} = \mathrm{poly}(e^{k^3S^2\Lambda^2/\varepsilon^2}, 2^{kS^3}, B^{(L+2)kS^2}/\Lambda^{kS^2})$ calls to ApproxBlockSVD, one for each element of $\mathcal{L}$ output by EnumerateNetworks (note that the runtime and sample complexity cost of running EnumerateNetworks at the very end is of much lower order). Each of these calls takes, up to lower order factors that will be absorbed elsewhere, $\widetilde{O}(d^2\log(1/\delta))$ time, so we conclude that FilteredPCA runs in time

$$\widetilde{O}(d^2\log(1/\delta)) \cdot \mathrm{poly}(e^{k^3S^2\Lambda^2/\varepsilon^2}, 2^{kS^3}, (B^{L+2}/\Lambda)^{kS^2})$$

as claimed. □

Remark 5.15 (Comparison to [CM20]). Here we briefly discuss what goes wrong if one simply tries mimicking the approach of [CM20]. Provided one has already recovered some (orthonormal) directions $w_1, ..., w_\ell$ spanning a subspace $W \subset V$, one would consider the matrix

$$\mathbf{M}_{\mathsf{CM}}^W \triangleq \Pi_{W^\perp} \mathop{\mathbb{E}}_{x,y} \Big[\mathbb{1}[|y| > \tau \wedge \|\Pi_W x\|^2 \le \alpha] \cdot (xx^\top - \mathrm{Id}) \Big] \Pi_{W^\perp}$$

for some $\alpha, \tau > 0$. The motivation for conditioning on $\|\Pi_W x\|^2 \le \alpha$ is that we now have

$$\langle \Pi_{V \setminus W}, \mathbf{M}_{\mathsf{CM}}^W \rangle = \mathop{\mathbb{E}}_{x,y} \Big[\mathbb{1}[|y| > \tau \wedge \|\Pi_W x\|^2 \le \alpha] \cdot (\|\Pi_{V \setminus W} x\|^2 - (k - \ell)) \Big],$$

and if one could choose τ strictly greater than the supremum of $|F(x)|$ over all x for which $\|\Pi_W x\|^2 \le \alpha$ and $\|\Pi_{V \setminus W} x\|^2 \le 2(k - \ell)$, then we would conclude that

$$\langle \Pi_{V \setminus W}, \mathbf{M}_{\mathsf{CM}}^W \rangle \ge (k - \ell) \cdot \mathbb{P}[|y| > \tau \wedge \|\Pi_W x\| \le \alpha] \tag{26}$$

and it would suffice to lower bound the probability on the right-hand side of (26). This is precisely the route taken by [CM20] for learning low-degree polynomials, but in the case of ReLU networks, it is not hard to devise functions F for which the probability on the right-hand side of (26) is zero for such choices of τ, e.g. if $d = k = 2$, $\ell = 1$, $v_1 = e_1$, and

$$F(x) \triangleq \phi(x/\alpha + y) - \phi(-x/\alpha + y).$$

References

[ADLS16] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Fast algorithms for segmented regression. In *International Conference on Machine Learning*, pages 2878–2886, 2016.

[AZL16] Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: Even faster svd decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*, pages 974–982, 2016.

[AZLL19] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pages 6158–6169, 2019.

[BB+18] Dmitry Babichev, Francis Bach, et al. Slice inverse regression with score functions. *Electronic Journal of Statistics*, 12(1):1507–1543, 2018.

[BG17] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 605–614, 2017.

[Bir40] Garrett Birkhoff. *Lattice theory*, volume 25. American Mathematical Soc., 1940.

[BJW19] Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory*, pages 195–268. PMLR, 2019.

[BR89] Avrim Blum and Ronald L Rivest. Training a 3-node neural network is np-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

[Bri12] David R Brillinger. A generalized linear model with gaussian regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012.

[CM20] Sitan Chen and Raghu Meka. Learning polynomials of few relevant dimensions. *arXiv preprint arXiv:2004.13748*, 2020.

[Dan17] Amit Daniely. Sgd learns the conjugate kernel class of the network. *CoRR*, abs/1702.08503, 2017.

[DGK+20] Ilias Diakonikolas, Surbhi Goel, Sushrut Karmalkar, Adam R Klivans, and Mahdi Soltanolkotabi. Approximation schemes for relu regression. In *Conference on Learning Theory*, 2020.

[DH18] Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pages 1887–1930, 2018.

[DKKZ20] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539, 2020.

[DLT18] Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[DMN19] Anindya De, Elchanan Mossel, and Joe Neeman. Is your function low dimensional? In *Conference on Learning Theory*, pages 979–993, 2019.

[DV20] Amit Daniely and Gal Vardi. Hardness of learning neural networks with natural weights. *arXiv preprint arXiv:2006.03177*, 2020.

[GGJ+20] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. *arXiv preprint arXiv:2006.12011*, 2020.

[GK19] Surbhi Goel and Adam R Klivans. Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory*, pages 1470–1499, 2019.

[GKK19] Surbhi Goel, Sushrut Karmalkar, and Adam Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems*, pages 8584–8593, 2019.

[GKKT17] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. In *Conference on Learning Theory*, pages 1004–1042. PMLR, 2017.

[GKLW18] Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. In *International Conference on Learning Representations*, 2018.

[GKM18] Surbhi Goel, Adam R. Klivans, and Raghu Meka. Learning one convolutional layer with overlapping patches. In Jennifer G. Dy and Andreas Krause 0001, editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1778–1786. PMLR, 2018.

[GLM18] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[GMOV18] Weihao Gao, Ashok Vardhan Makkuva, Sewoong Oh, and Pramod Viswanath. Learning one-hidden-layer neural networks under general input distributions. *CoRR*, abs/1810.04133, 2018.

[GRS18] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[GS19] Navin Goyal and Abhishek Shetty. Non-gaussian component analysis using entropy methods. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 840–851, 2019.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[JSA15] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv*, pages arXiv–1506, 2015.

[KS09] Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

[Li91] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[Li92] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[LMZ20] Yuanzhi Li, Tengyu Ma, and Hongyang R. Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In Jacob D. Abernethy and Shivani Agarwal 0001, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2613–2682. PMLR, 2020.

[LSSS14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, pages 855–863, 2014.

[LY17] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 597–607, 2017.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[MR18] Pasin Manurangsi and Daniel Reichman. The computational complexity of training relu (s). *arXiv preprint arXiv:1810.04207*, 2018.

[Ovc02] Sergei Ovchinnikov. Max-min representation of piecewise linear functions. *Contributions to Algebra and Geometry*, 43(1):297–302, 2002.

[PV16] Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

[RST09] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.

[Sha18] Ohad Shamir. Distribution-specific hardness of learning neural networks. *Journal of Machine Learning Research*, 19(32):1–29, 2018.

[SSSS17] Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3067–3075, 2017.

[SVWX17] Le Song, Santosh Vempala, John Wilmes, and Bo Xie. On the complexity of learning neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5520–5528, 2017.

[Tia17] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3404–3413. PMLR, 2017.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[Vu06] VH Vu. On the infeasibility of training neural networks with small mean-squared error. *IEEE Transactions on Information Theory*, 44(7):2892–2900, 2006.

[VW19] Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. In *COLT*, volume 99, 2019.

[VX11] Santosh S Vempala and Ying Xiao. Structure from local optima: Learning subspace juntas via higher order pca. *arXiv preprint arXiv:1108.3329*, 2011.

[ZLJ16] Yuchen Zhang, Jason D Lee, and Michael I Jordan. L1-regularized neural networks are improperly learnable in polynomial time. In *33rd International Conference on Machine Learning, ICML 2016*, pages 1555–1563. International Machine Learning Society (IMLS), 2016.

[ZPS17] Qiuyi Zhang, Rina Panigrahy, and Sushant Sachdeva. Electron-proton dynamics in deep learning. *CoRR*, abs/1702.00458, 2017.

[ZSJ+17] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149, 2017.

[ZYWG19] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1524–1534. PMLR, 2019.

A Deferred Proofs

A.1 Concentration for Piecewise Linear Functions

Lemma A.1. *For any $\delta > 0$ and any $t \le \Lambda^2 k$, the following holds. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a Λ-Lipschitz kicker with relevant subspace V of dimension k. Then for samples $x_1, ..., x_N \sim \mathcal{N}(0, \mathrm{Id})$, where $N = \Theta\left((\mu + \Lambda^2 k)^2 \log(1/\delta)/t^2\right)$, the empirical estimate $\widehat{\sigma}^2 \triangleq \frac{1}{N}\sum_i F(x_i)^2$ satisfies*

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,\mathrm{Id})}[F(x)^2] - \widehat{\sigma}^2 \right| \le t$$

with probability at least $1 - \delta$.

Proof. As F is Λ-Lipschitz and continuous piecewise-linear, by Theorem 4.9 and Lemma 4.5 it has a lattice polynomial representation $\max_{j\in[m]} \min_{i\in\mathcal{I}_j} \langle u_i, \cdot\rangle$ for some clauses $\{\mathcal{I}_j\}$ and vectors $\{u_i\}$ for which $\|u_i\| \le \Lambda$. In particular, by Cauchy-Schwarz, $|F(x)| \le \Lambda\|x\|$ for all x. Now define the function $G(x) \triangleq F(x)^2 - \mu$ where $\mu \triangleq \mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{Id})}[F(x)^2]$. We can therefore naively upper bound the moments of G by

$$\mathbb{E}[|G|^t]^{1/t} \le \mu + \mathbb{E}[F^{2t}]^{1/t} \le \mu + \Lambda^2 \cdot \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Pi_V)}[\|x\|^{2t}]^{1/t} \le \mu + O(\Lambda^2 k)\cdot(t-1)$$

for all $t \ge 2$, where the last step follows by standard hypercontractivity. Furthermore, $\mathbb{E}[|G|] \le 2\mu$. For $x \sim \mathcal{N}(0, \mathrm{Id})$, $G(x)$ is therefore a sub-exponential, mean-zero random variable with sub-exponential norm $K \triangleq O(\mu + \Lambda^2 k)$, so by Fact 3.7 and the bound on t in the hypothesis, for $N = \Theta(K^2 \log(1/\delta)/t^2)$, the claim follows. □

A.2 Representing Boolean Functions as ReLU Networks

Lemma A.2. *For any function $F : \{\pm 1\}^n \to \{\pm 1\}$, there exists a set of weight matrices $\mathbf{W}_0, ..., \mathbf{W}_{n-1}$ for which $F(x) = \mathbf{W}_{n-1}\phi(\mathbf{W}_{n-2}\phi(\cdots\phi(\mathbf{W}_0 x)\cdots))$ for all $x \in \{\pm 1\}^n$.*

Proof. From the Fourier expansion of F as $F(x) = \sum_S \widehat{F}[S]\prod_{i\in S} x_i$, we see that it suffices to show how to represent any Fourier basis function $\prod_{i\in S} x_i$ with a ReLU network with depth n. We first show how to represent the function $x_1 x_2$. Observe that for any $x_1, x_2 \in \{\pm 1\}$, we have that

$$x_1 \cdot x_2 = \phi(x_1 + x_2) + \phi(-x_1 - x_2) - \phi(x_2) - \phi(-x_2), \tag{27}$$

which is a two-layer neural network. Suppose inductively that for some $1 \le m < n$, there exist weight matrices $\mathbf{W}'_0, \ldots, \mathbf{W}'_{m-1}$ for which $\prod_{i=1}^m x_i = \mathbf{W}'_{m-1}\phi(\mathbf{W}'_{m-2}\phi(\cdots\phi(\mathbf{W}'_0 x)\cdots))$ for all

$x \in \{\pm 1\}^n$. Then to compute $\prod_{i=1}^{m+1} x_i$, we can use (27) to conclude that

$$\prod_{i=1}^{m+1} x_i = \phi\left(\prod_{i=1}^{m} x_i + x_{m+1}\right) + \phi\left(-\prod_{i=1}^{m} x_i - x_{m+1}\right) - \phi(x_{m+1}) - \phi(-x_{m+1}).$$

It is clear that this can be represented as a ReLU network with depth $m+1$. □