A Law of Robustness for Two-Layers Neural Networks

Sébastien Bubeck SEBUBECK@MICROSOFT.COM
Microsoft Research

Yuanzhi Li YUANZHIL@ANDREW.CMU.EDU
CMU

Dheeraj Nagaraj DHEERAJ@MIT.EDU
MIT

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We initiate the study of the inherent tradeoffs between the size of a neural network and its robustness, as measured by its Lipschitz constant. We make a precise conjecture that, for any Lipschitz activation function and for most datasets, any two-layers neural network with k neurons that perfectly fit the data must have its Lipschitz constant larger (up to a constant) than $\sqrt{n/k}$ where n is the number of datapoints. In particular, this conjecture implies that overparametrization is necessary for robustness, since it means that one needs roughly one neuron per datapoint to ensure a $O(1)$-Lipschitz network, while mere data fitting of d-dimensional data requires only one neuron per d datapoints. We prove a weaker version of this conjecture when the Lipschitz constant is replaced by an upper bound on it based on the spectral norm of the weight matrix. We also prove the conjecture in the high-dimensional regime $n \approx d$ (which we also refer to as the undercomplete case, since only $k \leq d$ is relevant here). Finally we prove the conjecture for polynomial activation functions of degree p when $n \approx d^p$. We complement these findings with experimental evidence supporting the conjecture.

Keywords: Neural Networks, Robustness, Memorization, Interpolation, Lipshitz extension

1. Introduction

We study two-layers neural networks with inputs in $\mathbb{R}^d$, k neurons, and Lipschitz non-linearity $\psi : \mathbb{R} \to \mathbb{R}$. These are functions of the form:

$$x \mapsto \sum_{\ell=1}^{k} a_\ell \psi(w_\ell \cdot x + b_\ell)\,, \tag{1}$$

with $a_\ell, b_\ell \in \mathbb{R}$ and $w_\ell \in \mathbb{R}^d$ for any $\ell \in [k]$. We denote by $\mathcal{F}_k(\psi)$ the set of functions of the form (1). When k is large enough and ψ is non-polynomial, this set of functions can be used to fit any given data set (Cybenko, 1989; Leshno et al., 1993). That is, given a data set $(x_i, y_i)_{i\in[n]} \in (\mathbb{R}^d \times \mathbb{R})^n$, one can find $f \in \mathcal{F}_k(\psi)$ such that

$$f(x_i) = y_i, \forall i \in [n]\,. \tag{2}$$

In a variety of scenarios one is furthermore interested in fitting the data *smoothly*. For example, in machine learning, the data fitting model f is used to make predictions at unseen points $x \notin$

$\{x_1, \ldots, x_n\}$. It is reasonable to ask for these predictions to be stable, that is a small perturbation of x should result in a small perturbation of $f(x)$.

A natural question is: how "costly" is this stability restriction compared to mere data fitting? In practice it seems much harder to find robust models for large scale problems, as first evidenced in the seminal paper (Szegedy et al., 2013). In theory the "cost" of finding robust models has been investigated from a computational complexity perspective in (Bubeck et al., 2019), from a statistical perspective in (Schmidt et al., 2018), and more generally from a model complexity perspective in (Degwekar et al., 2019; Raghunathan et al., 2019; Allen-Zhu and Li, 2020). We propose here a different angle of study within the broad model complexity perspective: does a model *have to* be larger for it to be robust? Empirical evidence (e.g., (Goodfellow et al., 2015; Madry et al., 2018)) suggests that bigger models (also known as "overparametrization") do indeed help for robustness.

Our main contribution is a conjecture (Conjecture 1 and Conjecture 2) on the precise tradeoffs between size of the model (i.e., the number of neurons k) and robustness (i.e., the Lipschitz constant of the data fitting model $f \in \mathcal{F}_k(\psi)$) for generic data sets. We say that a data set $(x_i, y_i)_{i\in[n]}$ is *generic* if it is i.i.d. with x_i uniform (or approximately so, see below) on the sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and y_i uniform on $\{-1, +1\}$. We give the precise conjecture in Section 2. We prove several weaker versions of Conjecture 1 and Conjecture 2 respectively in Section 4 and Section 3. We also give empirical evidence for the conjecture in Section 5.

A corollary of our conjecture. A key fact about generic data, established in Baum (1988); Yun et al. (2019); Bubeck et al. (2020), is that one can memorize arbitrary labels with $k \approx n/d$, that is merely one neuron per d datapoints. Our conjecture implies that for such optimal-size neural networks it is *impossible* to be robust, in the sense that the Lipschitz constant must be of order $\sqrt{d}$. The conjecture also states that to be robust (i.e. attain Lipschitz constant $O(1)$) one must *necessarily* have $k \approx n$, that is roughly each datapoint must have its own neuron. Therefore, we obtain a trade off between size and robustness, namely to make the network robust it needs to be d *times larger* than for mere data fitting. We illustrate these two cases in Figure 1. We train a neural network to fit generic data, and plot the maximum gradient over several randomly drawn points (a proxy for the Lipschitz constant) for various values of $\sqrt{d}$, when either $k = n$ (blue dots) or $k = \frac{10n}{d}$ (red dots). As predicted, for the large neural network ($k = n$) the Lipschitz constant remains roughly constant, while for the optimally-sized one ($k = \frac{10n}{d}$) the Lipschitz constant increases roughly linearly in $\sqrt{d}$.

Notation. For $\Omega \subset \mathbb{R}^d$ we define $\mathrm{Lip}_\Omega(f) = \sup_{x\neq x'\in\Omega} \frac{|f(x)-f(x')|}{\|x-x'\|}$ (if $\Omega = \mathbb{R}^d$ we omit the subscript and write $\mathrm{Lip}(f)$), where $\|\cdot\|$ denotes the Euclidean norm. For matrices we use $\|\cdot\|_{\mathrm{op}}, \|\cdot\|_{\mathrm{op},*}, \|\cdot\|_{\mathrm{F}}$ and $\langle\cdot,\cdot\rangle$ for respectively the operator norm, the nuclear norm (sum of singular values), the Frobenius norm, and the Frobenius inner product. We also use these notations for tensors of higher order, see Appendix A for more details on tensors. We denote $c > 0$ and $C > 0$ for universal numerical constants, respectively small enough and large enough, whose values can change in different occurences. Similarly, by $c_p > 0$ and $C_p > 0$ we denote constants depending only on the parameter p. We also write $\mathrm{ReLU}(t) = \max(t, 0)$ for the rectified linear unit.

Generic data. We give some flexibility in our definition of "generic data" in order to focus on the essence of the problem, rather than technical details. Namely, in addition to the spherical model mentioned above, where x_i is i.i.d. uniform on the sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$, we also consider the very closely related model where x_i is i.i.d. from a centered Gaussian with covariance $\frac{1}{d}\mathrm{I}_d$ (in particular $\mathbb{E}[\|x_i\|^2] = 1$, and in fact $\|x_i\|$ is tightly concentrated around 1). In both cases we

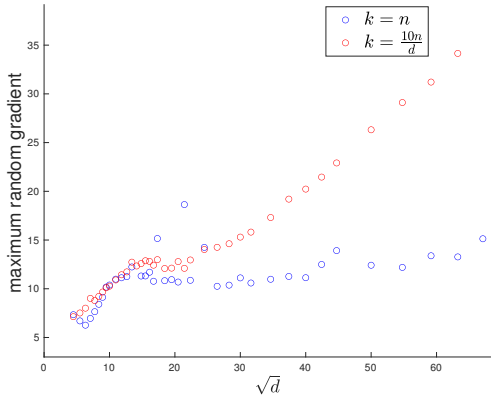

Figure 1: See Section 5 for the details of this experiment.

consider y_i to be i.i.d. random signs. We say that a property holds with high probability for *generic data*, if it holds with high probability either for the spherical model or for the Gaussian model.

2. A conjectured law of robustness

Our main contribution is the following conjecture, which asserts that, on generic data sets, increasing the size of a network is necessary to obtain robustness:

Conjecture 1 *For generic data sets, with high probability*[1]*, any $f \in \mathcal{F}_k(\psi)$ fitting the data*[2] *(i.e., satisfying (2)) must also satisfy:*

$$\mathrm{Lip}_{\mathbb{S}^{d-1}}(f) \geq c\sqrt{\frac{n}{k}}\,.$$

Note that for generic data, with high probability (for $n = \mathrm{poly}(d)$), there exists a smooth interpolation. Namely there exists $g : \mathbb{R}^d \to \mathbb{R}$ with $g(x_i) = y_i, \forall i \in [n]$ and $\mathrm{Lip}(g) = O(1)$. This follows easily from the fact that with high probability (for large d) one has $\|x_i - x_j\| \geq 1, \forall i \neq j$. Conjecture 1 puts restrictions on how smoothly one can interpolate data with small neural networks. A striking consequence of the conjecture is that for a two-layers neural network $f \in \mathcal{F}_k(\psi)$ to be as robust as this function g (i.e., $\mathrm{Lip}(f) = O(1)$) and fit the data, one must have $k = \Omega(n)$, i.e., roughly one neuron per data point. On the other hand with that many neurons it is quite trivial to smoothly interpolate the data, as we explain in Section 3.3. Thus the conjecture makes a strong statement that essentially the trivial smooth interpolation is the best thing one can do. In addition to making the prediction that one neuron per datapoint is necessary for optimal smoothness, the conjecture also gives a precise prediction on the possible tradeoff between size of the network and its robustness. We also conjecture that this whole range of tradeoffs is actually achievable:

1. We do not quantify the "with high probability" in our conjecture. We believe the conjecture to be true except for an event of exponentially small probability with respect to the sampling of a generic data set, but even proving that the statement is true with strictly positive probability would be extremely interesting.
2. We expect the same lower bound to hold even if one only asks f to approximately fit the data. In fact our provable variants of Conjecture 1 are based proofs that are robust to only assuming an approximately fitting f.

Conjecture 2 *Let n, d, k be such that $C \cdot \frac{n}{d} \leq k \leq C \cdot n$ and $n \leq d^C$ where C is an arbitrarily large constant in the latter occurence. There exists ψ such that, for generic data sets, with high probability, there exists $f \in \mathcal{F}_k(\psi)$ fitting the data (i.e., satisfying (2)) and such that*

$$\mathrm{Lip}_{\mathbb{S}^{d-1}}(f) \leq C\sqrt{\frac{n}{k}}\,.$$

The condition $k \leq C \cdot n$ in Conjecture 2 is necessary, for any interpolation of the data must have Lipschitz constant at least a constant. The other condition on k, namely $k \geq C \cdot \frac{n}{d}$, is also necessary, for that many neurons is needed to merely guarantee the existence of a data-fitting neural network with k neurons (see Baum (1988); Yun et al. (2019); Bubeck et al. (2020)). Finally the condition $n \leq d^C$ is merely used to avoid explicitly stating a logarithmic term in our conjecture (indeed, equivalently one can replace this condition by adding a multiplicative polylogarithmic term in d in the claimed inequality).

Our results around Conjecture 2 (Section 3). We prove Conjecture 2 for both the optimal smoothness regime (which is quite straightforward, see Section 3.3) and for the optimal size regime (here more work is needed, and we use a certain tensor-based construction, see Section 3.4). In the latter case we only prove approximate data fitting (mostly to simplify the proofs), and more importantly we need to assume that n is of order d^p for some even integer p. It would be interesting to generalize the proof to any n. While the conjecture remains open between these two extreme regimes, we do give a construction in Section 3.3 which has the correct qualitative behavior (namely increasing k improves the Lipschitz constant), albeit the scaling we obtain is n/k instead of $\sqrt{n/k}$, see Theorem 3.

Our results around Conjecture 1 (Section 4). We prove a weaker version of Conjecture 1 where the Lipschitz constant on the sphere is replaced by a proxy involving the spectral norm of the weight matrix, see Theorem 5. We also prove the conjecture in the optimal size regime, specifically when $n = d^p$ for an integer p and one uses a polynomial activation function of degree p, see Theorem 8. For $p = 1$ (i.e., $n \approx d$) we in fact prove the conjecture for abritrary non-linearities, see Theorem 6.

Further open problems. Our proposed law of robustness is a first mathematical formalization of the broader phenomenon that "overparametrization in neural networks is necessary for robustness". Ideally one would like a much more refined understanding of the phenomenon than the one given in Conjecture 1. For example, one could imagine that in greater generality, the law would read $\mathrm{Lip}_\Omega(f) \geq F(k, (x_i, y_i)_{i\in[n]}, \Omega)$. That is, we would like to understand how the achievable level of smoothness depends on the particular data set at hand, but also on the set where we expect to be making predictions. Another direction to generalize the law would be to extend it to multi-layers neural networks. In particular one could imagine the most general law would replace the parameter k (number of neurons) by the type of architecture being used and in turn predict the best architecture for a given data set and prediction set. At the other end of the spectrum, it would also be interesting to understand the law in more restrictive models, such as kernels or even neural tangent kernels. We note however that this is far from an easy question, for example the case of a power activation function is equivalent to a kernel corresponding to the embedding $x \mapsto x^{\otimes p}$, and as we explain in Section 4.3 this is already a quite non-trivial case of our conjecture (indeed, the number of neurons in this case corresponds to the rank of a certain tensor of order p, and the tensor rank is a notoriously difficult notion to work with). Finally note that our proposed law apply to *all* neural networks, but

it would also be interesting to understand how the law interacts with algorithmic considerations (for example in Section 5 we use Adam Kingma and Ba (2015) to find a set of weights that qualitatively match Conjecture 2).

3. Smooth interpolation

We start with a warm-up in Section 3.1 where we discuss the simplest case of interpolation with a linear model ($k = 1, n \leq d$) and in Section 3.2 for the optimal smoothness regime ($k = n$). We generalize the construction of Section 3.2 in Section 3.3 to obtain the whole range of tradeoffs between k and $\mathrm{Lip}(f)$, albeit with a suboptimal scaling, see Theorem 3. We also generalize the linear model calculations of Section 3.1 in Section 3.4 to obtain the optimal size regime for larger values of n via a certain tensor construction.

3.1. The simplest case: optimal size regime when $n \leq c \cdot d$

Let us consider $k = 1$, $n \leq c \cdot d$ and $\psi(t) = t$. Thus we are trying to find $w \in \mathbb{R}^d$ such that $w \cdot x_i = y_i$ for all $i \in [n]$, or in other words $Xw = Y$ with X the $n \times d$ matrix whose i^{th} row is x_i, and $Y = (y_1, \ldots, y_n)$. The smoothest solution to this system (i.e., the one minimizing $\|w\|$) is

$$w = X^\top (XX^\top)^{-1} Y \,,$$

Note that

$$\mathrm{Lip}(x \mapsto w \cdot x) = \|w\| = \sqrt{w^\top w} = \sqrt{Y^\top (XX^\top)^{-1} Y} \,.$$

Using [Theorem 5.58, Vershynin (2012)] one has with probability at least $1 - \exp(C - cd)$ (and using that $n \leq c \cdot d$) that

$$XX^\top \succeq \frac{1}{2} I_n \,,$$

and thus $\mathrm{Lip}(x \mapsto w \cdot x) \leq \sqrt{2} \cdot \|Y\| = \sqrt{2n}$. This concludes the proof sketch of Conjecture 2 for the simplest case $k = 1$ and $n \leq d$.

3.2. Another simple case: optimal smoothness regime

Next we consider the optimal smoothness regime in Conjecture 2, namely $k = n$. First note that, for generic data and $n = \mathrm{poly}(d)$, with high probability the caps $C_i := \left\{x \in \mathbb{S}^{d-1} : x_i \cdot x \geq 0.9\right\}$ are disjoint sets and moreover they each contain a single data point (namely x_i). With a single ReLU unit it is then easy to make a smooth function (10-Lipschitz) which is 0 outside of C_i and equal to $+1$ at x_i (in other words the neuron activates for a single data point), namely $x \mapsto 10 \cdot \mathrm{ReLU}\left(x_i \cdot x - 0.9\right)$. Thus one can fit the entire data set with the following ReLU network which is 10-Lipschitz on the sphere:

$$f(x) = \sum_{i=1}^{n} 10 y_i \cdot \mathrm{ReLU}\left(x_i \cdot x - 0.9\right) \,.$$

This concludes the proof of Conjecture 2 for the optimal smoothness regime $k = n$.

3.3. Intermediate regimes via ReLU networks

We now combine the two constructions above (the linear model of Section 3.1 and the "isolation" strategy of Section 3.2) to give a construction that can trade off size for robustness (albeit not optimally according to Conjecture 2), see Appendix C for the proof.

Theorem 3 *Let n, d, k be such that $C \cdot \frac{n \log(n)}{d} \leq k \leq C \cdot n$. For generic data sets, with probability at least $1 - 1/n^C$, there exists $f \in \mathcal{F}_k(\mathrm{ReLU})$ fitting the data (i.e., satisfying (2)) and such that*

$$\mathrm{Lip}_{\mathbb{S}^{d-1}}(f) \leq C \cdot \frac{n \log(d)}{k} .$$

3.4. Optimal size networks via tensor interpolation

In this section we essentially prove Conjecture 2 in the optimal size regime (namely $k \cdot d \approx n$), with three caveats:

1. We allow a slack of a $\log n$ factor by considering $k \cdot d = Cn \log(n)$ instead of the optimal $k \cdot d = Cn$ as in Baum (1988); Bubeck et al. (2020).

2. We only prove approximate fit rather than exact fit. It is likely that with more work one can use the core of our argument to obtain exact fit. For that reason we did not make any attempt to optimize the dependency on ε in Theorem 4. For instance one could probably obtain $\log(1/\varepsilon)$ rather than $1/\mathrm{poly}(\varepsilon)$ dependency by using an iterative scheme that fits the residuals, as in (Bresler and Nagaraj, 2020; Bubeck et al., 2020).

3. We assume that n is of order d^p for some even integer p. While it might be that one can apply the same proof for odd integers, the whole construction crucially relies on p being an even integer as we essentially do a linear regression over the feature embedding $x \mapsto x^{\otimes p}$. A possible approach to extend the proof to other values of n would be use the scheme of Section 3.3 with the linear regression there replaced by the tensor regression used below.

Theorem 4 *Fix $\varepsilon > 0$, p an even integer, and let $\psi(t) = t^p$. Let n, d, k be such that $n \log(n) = \varepsilon^2 \cdot d^p$ and $k = C_p \cdot d^{p-1}$. Then for generic data, with probability at least $1 - 1/n^C$, there exists $f \in \mathcal{F}_k(\psi)$ such that*

$$|f(x_i) - y_i| \leq C_p \cdot \varepsilon \,, \forall i \in [n] \,, \tag{3}$$

and

$$\mathrm{Lip}_{\mathbb{S}^{d-1}}(f) \leq C_p \sqrt{\frac{n}{k}} .$$

Proof We propose to approximately fit with the following neural network:

$$f(x) = \sum_{i=1}^{n} y_i (x_i \cdot x)^p .$$

Naively one might think that this neural network requires n neurons. However, it turns out that one can always decompose a symmetric tensor of order p into $k = 2^p d^{p-1}$ rank-1 symmetric tensors of order p, so that in fact $f \in \mathcal{F}_k(\psi)$. For $p = 2$ this simply follows from eigendecomposition and for general p we give a simple proof in [Appendix A, Lemma 11].

One also has by applying [Appendix B, Lemma 13] with $\tau = C_p \log(n)$ and doing an union bound, that with probability at least $1 - 1/n^C$, for any $j \in [n]$,

$$\left| \sum_{i=1, i \neq j}^{n} y_i (x_i \cdot x_j)^p \right| \leq C_p \sqrt{\frac{n \log(n)}{d^p}} \leq C_p \varepsilon \, .$$

In particular this proves (3).

Thus it only remains to estimate the Lipschitz constant, which by [Appendix A, Lemma 10] is reduced to estimating the operator norm of the tensor $\sum_{i=1}^n y_i x_i^{\otimes p}$. We do so in [Appendix B, Lemma 14]. ■

4. Provable weaker versions of Conjecture 1

Conjecture 1 can be made weaker along several directions. For example the quantity of interest $\mathrm{Lip}_{\mathbb{S}^{d-1}}(f)$ can be replaced by various upper bound proxies for the Lipschitz constant. A mild weakening would be to replace it by the Lipschitz constant on the whole space (we shall in fact only consider this notion here). A much more severe weakening is to replace it by a quantity that depends on the spectral norm of the weight matrix (essentially ignoring the pattern of activation functions). For the latter proxy we actually give a complete proof, see Theorem 5, which in particular formally proves that "overparametrization is a law of robustness for generic data sets". Other interesting directions to weaken the conjecture include specializing it to common activation functions, or simply having a smaller lower bound on the Lipschitz constant. In Section 4.2 we prove the conjecture when n is replaced by d in the lower bound. We say that this inequality is in the "very high-dimensional case", in the sense that it matches the conjecture for $n \approx d$ (alternatively we also refer to it as the "undercomplete case", in the sense that only $k \leq d$ is relevant in this very high-dimensional scenario). In the moderately high-dimensional case ($n \gg d$) the proof strategy we propose in Section 4.2 cannot work. In Section 4.3 we give another argument for the latter case, specifically in the optimal size regime (i.e., $k \cdot d \approx n$) and for a power activation function, see Theorem 7. We generalize this to polynomial activation functions in Section 4.4. In the specific case of a quadratic activation function we also show a lower bound that applies for any k and which is in fact larger than the one given in Conjecture 1, see Theorem 9 in Section 4.5.

4.1. Spectral norm proxy for the Lipschitz constant

We can rewrite (1) as

$$f(x) = a^\top \psi(Wx + b) \, , \tag{4}$$

where $a = (a_1, \ldots, a_k) \in \mathbb{R}^k$, $b = (b_1, \ldots, b_k) \in \mathbb{R}^k$, $W \in \mathbb{R}^{k \times d}$ is the matrix whose ℓ^{th} row is w_ℓ, and ψ is extended from $\mathbb{R} \to \mathbb{R}$ to $\mathbb{R}^k \to \mathbb{R}^k$ by applying it coordinate-wise. We prove here the following:

Theorem 5 *Assume that ψ is L-Lipschitz. For $f \in \mathcal{F}_k(\psi)$ one has*

$$\mathrm{Lip}(f) \leq L \cdot \|a\| \cdot \|W\|_{\mathrm{op}} \, . \tag{5}$$

For a generic data set, if $f(x_i) = y_i, \forall i \in [n]$ and f has no bias terms (i.e., $b = 0$ in (4)), then with positive probability one has:

$$L \cdot \|a\| \cdot \|W\|_{\mathrm{op}} \geq \sqrt{\frac{n}{k}} \,. \tag{6}$$

Note that we prove the inequality (6) only with positive probability (i.e., there exists a data set where the inequality is true), but in fact it is easy to derive the statement with high probability using classical concentration inequalities.

Proof Since $\psi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, we have:

$$f(x)-f(x') \leq \|a\| \cdot \|\psi(Wx+b)-\psi(Wx'+b)\| \leq L \cdot \|a\| \cdot \|Wx-Wx'\| \leq L \cdot \|a\| \cdot \|W\|_{\mathrm{op}} \cdot \|x-x'\| \,,$$

which directly proves (5).

Next, following the proof of [Proposition 1, Bubeck et al. (2020)] one obtains that for a generic data set, with positive probability, one has (without bias terms):

$$\sum_{\ell=1}^{k} |a_\ell| \cdot \|w_\ell\| \geq \frac{\sqrt{n}}{L} \,.$$

It only remains to observe that:

$$\frac{\sqrt{n}}{L} \leq \sum_{\ell=1}^{k} |a_\ell| \cdot \|w_\ell\| \leq \sqrt{\sum_{\ell=1}^{k} |a_\ell|^2 \cdot \sum_{\ell=1}^{k} \|w_\ell\|^2} = \|a\| \cdot \|W\|_{\mathrm{F}} \leq \sqrt{k} \cdot \|a\| \cdot \|W\|_{\mathrm{op}} \,,$$

which concludes the proof of (6). ■

4.2. Undercomplete case

Next we prove the conjecture in the high dimensional case $n \approx d$. More precisely we replace n by d in the conjectured lower bound. Importantly note that the resulting lower bound then becomes non-trivial only in the regime $k \leq d$ (the "undercomplete case").

We consider in fact a slightly more general scenario than interpolation with a neural network, namely we simply assume that one interpolates the data with a function $f(x) = g(Px)$ where P is a linear projection on a k-dimensional subspace (this clearly generalizes $f \in \mathcal{F}_k(\psi)$, in fact it even allows for the non-linearity ψ to depend on the data[3], or to have a different non-linearity for each neuron).

Theorem 6 *Let $n \geq d$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function such that $f(x_i) = y_i, \forall i \in [n]$ and moreover $f(x) = g(Px)$ for some differentiable function $g : \mathbb{R}^k \to \mathbb{R}$ and matrix $P \in \mathbb{R}^{k \times d}$. Then, for generic data, with probability at least $1 - \exp(C - cd)$ one must have*

$$\mathrm{Lip}(f) \geq c\sqrt{\frac{d}{k}} \,.$$

3. It would be interesting to study whether allowing data-dependent non-linearities could affect the conclusion of our conjectures. Such study would need to crucially rely on having only one hidden layer, as it is known from the Kolmogorov-Arnold theorem that with two hidden layers and data-dependent non-linearities one can obtain perfect approximation properties with $k \leq d$ (albeit the non-linearities are non-smooth).

Proof Let us modify g so that P is simply an orthogonal projection operator (i.e., $PP^\top = \mathrm{I}_k$). Let us also assume for sake of notational simplicity that we have a balanced data set of size $2n$, that is with: $y_1, \dots y_n = +1$ and $y_{n+1}, \dots, y_{2n} = -1$. Let us denote $x_i' = x_i - x_{n+i}$ for $i \in [n]$. The sequence x_i' is i.i.d. and satisfies $\mathbb{E}[x_i' x_i'^\top] = \frac{2}{d}\mathrm{I}_d$.

Now observe that on the segment $[x_i, x_{n+i}]$ (whose length is less than 2), the function f changes value from $+1$ to -1, and thus there exists $z_i \in [x_i, x_{n+i}]$ such that:

$$1 \leq |\nabla f(z_i) \cdot (x_i - x_{n+i})| = |\nabla f(z_i) \cdot x_i'| \,.$$

Moreover one has (using that $\nabla f(x) = P^\top \nabla g(x)$, and thus $\|\nabla g(x)\| = \|P \nabla f(x)\| \leq \mathrm{Lip}(f)$)

$$|\nabla f(z_i) \cdot x_i'| = |\nabla g(Pz_i) \cdot (Px_i')| \leq \mathrm{Lip}(f) \cdot \|Px_i'\| \,.$$

Combining the two above displays one has:

$$\frac{n}{\mathrm{Lip}(f)} \leq \sum_{i=1}^n \|Px_i'\| \leq \sqrt{n \sum_{i=1}^n \|Px_i'\|^2} = \sqrt{n \sum_{i=1}^n x_i'^\top P^\top P x_i'} = \sqrt{n \langle \sum_{i=1}^n x_i' x_i'^\top, P^\top P \rangle_{\mathrm{HS}}} \,.$$

Using [Theorem 5.39, Vershynin (2012)] (specifically (5.23)) we know that with probability at least $1 - \exp(C - cd)$ we have $\left\|\sum_{i=1}^n x_i' x_i'^\top\right\|_{\mathrm{op}} \leq C\frac{n}{d}$ (here we use $n \geq d$ too). Moreover we have $\|P^\top P\|_{\mathrm{op},*} = \mathrm{Tr}(P^\top P) = \mathrm{Tr}(PP^\top) = k$. Thus we have $\langle \sum_{i=1}^n x_i' x_i'^\top, P^\top P \rangle_{\mathrm{HS}} \leq C\frac{n \cdot k}{d}$ so that with the above display one obtains $\frac{n}{\mathrm{Lip}(f)} \leq n\sqrt{\frac{Ck}{d}}$, which concludes the proof. ■

4.3. Power activation

We prove here the conjecture for the power activation function $\psi(t) = t^p$ with p an integer and with no bias terms (we deal with general polynomials, including with bias, in Section 4.4). Without bias such a network can be written as:

$$f(x) = \sum_{\ell=1}^k a_\ell (w_\ell \cdot x)^p = \langle T, x^{\otimes p} \rangle \,, \tag{7}$$

where $T = \sum_{\ell=1}^k a_\ell w_\ell^{\otimes p}$. As we already saw in the proof of Theorem 4 (see specifically [Appendix A, Lemma 11]), without loss of generality we have $k \leq C_p d^{p-1}$. We now prove that tensor networks of the form (7) cannot obtain a Lipschitz constant[4] better than $\sqrt{n/d^{p-1}}$, in accordance with Conjecture 1 for full rank tensors (where $k \approx d^{p-1}$).

Theorem 7 *Assume that we have a tensor T of order p such that $\langle T, x_i^{\otimes p} \rangle = y_i, \forall i \in [n]$. Then, for generic data, with probability at least $1 - C\exp(-c_p d)$, one must have*

$$\|T\|_{\mathrm{op}} \geq c_p \sqrt{\frac{n}{d^{p-1}}} \,.$$

4. Note that without loss of generality one can assume T to be symmetric, since we only consider how it acts on $x^{\otimes p}$. For symmetric tensors one has that the Lipschitz constant on the unit ball is lower bounded by the operator norm of T thanks to (9)

Proof Denoting $\Omega = \sum_{i=1}^n y_i x_i^{\otimes p}$, we have (using $y_i^2 = 1$ for the first equality and [Appendix A, Lemma 12] for the last inequality):

$$n = \langle T, \Omega \rangle \leq \|\Omega\|_{\mathrm{op}} \cdot \|T\|_{\mathrm{op},*} \leq d^{p-1} \cdot \|\Omega\|_{\mathrm{op}} \cdot \|T\|_{\mathrm{op}} \,. \tag{8}$$

Thus we obtain $\|T\|_{\mathrm{op}} \geq \frac{n}{d^{p-1} \cdot \|\Omega\|_{\mathrm{op}}}$, and it only remains to apply [Appendix B, Lemma 14] which states that with probability at least $1 - C\exp(-c_p d)$ one has $\|\Omega\|_{\mathrm{op}} \leq C_p \sqrt{\frac{n}{d^{p-1}}}$. ■

4.4. Polynomial activation

We now observe that one can generalize Theorem 7 to handle biases (the parameters b_l in 1), and in fact even general polynomial activation function. Indeed, observe that any polynomial of $\langle w, x \rangle + b$ must also be a polynomial in $\langle w, x \rangle$, albeit with different coefficients.

Theorem 8 *Let $\psi(t) = \sum_{q=0}^p \alpha_q t^q$ and assume that we have $f \in \mathcal{F}_k(\psi)$ such that $f(x_i) = y_i, \forall i \in [n]$. Then, for generic data, with probability at least $1 - C\exp(-c_p d)$ one must have*

$$\mathrm{Lip}_{\{x : \|x\| \leq 1\}}(f) \geq c_p \sqrt{\frac{n}{d^{p-1}}} \,.$$

Proof Note that for $f \in \mathcal{F}_k(\psi)$ there exists tensors $T_0, \ldots, T_p$, such that T_q is a tensor of order q, and f can be written as:

$$f(x) = \sum_{q=0}^p \langle T_q, x^{\otimes q} \rangle \,.$$

Now let us define $\Omega_q = \sum_{i=1}^n y_i x_i^{\otimes q}$, and observe that

$$n = \sum_{i=1}^n y_i f(x_i) = \sum_{q=0}^p \langle T_q, \Omega_q \rangle \,,$$

and thus there exists $q \in \{1, \ldots, p\}$ such that $\langle T_q, \Omega_q \rangle \geq c_p n$ (we ignore the term $q = 0$ by considering the largest balanced subset of the data, i.e. we assume $\sum_{i=1}^n y_i = 0$). Now one can repeat the proof of Theorem 7 to obtain that with probability at least $1 - C\exp(-c_p d)$, one has $\|T_q\|_{\mathrm{op}} \geq c_p \sqrt{\frac{n}{d^{p-1}}}$. It only remains to observe that the Lipschitz constant of f on the unit ball is lower bounded by $\|T_q\|_{\mathrm{op}}$.

As we mentioned in Section 4.3, without loss of generality we can assume T_q is symmetric, and thus by (9) there exists $x \in \mathbb{S}^{d-1}$ such that $\|T_q\|_{\mathrm{op}} = \langle T_q, x^{\otimes q} \rangle$. Now consider the univariate polynomial $P(t) = f(tx)$. By Markov brothers' inequality one has $\max_{t \in [-1,1]} P(t) \geq |P^{(q)}(0)| = q! \cdot |\langle T_q, x^{\otimes q} \rangle| = q! \cdot \|T_q\|_{\mathrm{op}}$, thus concluding the proof. ■

4.5. Quadratic activation

In Section 4.3 we obtained a lower bound for tensor networks that match Conjecture 1 only when the rank of the corresponding tensor is maximal. Here we show that for quadratic networks (i.e., $p = 2$) we can match Conjecture 1, and in fact even obtain a better bound, for any rank k:

Theorem 9 *Assume that we have a matrix $T \in \mathbb{R}^{d \times d}$ with rank k such that:*

$$\langle T, x_i^{\otimes 2} \rangle = y_i, \forall i \in [n] \, .$$

Then, for generic data, with probability at least $1 - C\exp(-cd)$, one must have

$$\|T\|_{\mathrm{op}} \geq c\frac{\sqrt{nd}}{k} \ \ (\geq c\sqrt{n/k}) \, .$$

Proof The proof is exactly the same as for Theorem 7, except that in (8), instead of using Lemma 12 we use the fact that for a matrix T of rank k one has:

$$\|T\|_{\mathrm{op},*} \leq k \cdot \|T\|_{\mathrm{op}} \, .$$

■

5. Experiments

We consider a generic dataset from the Gaussian model (i.e., $x_1, \ldots, x_n$ i.i.d. from $\mathcal{N}(0, \frac{1}{d} I_d)$ and labels $y_1, \ldots, y_n$ i.i.d from the uniform distribution over $\{-1, 1\}$ and independent of $x_1, \ldots, x_n$). For various values of (n, d, k) we train two-layers neural networks with k ReLU units and batch normalization (see Ioffe and Szegedy (2015)) between the linear layer and ReLU layer, using the Adam optimizer (Kingma and Ba, 2015) on the least squares loss. We keep the values of (n, k, d) where the network successfully memorizes the random labels (possibly after a rounding to $\{-1, +1\}$, and such that prior to rounding the least squares loss is at most some small value ε to be specified later). Given a triple (n, d, k), suppose the output of the trained network is $f_{n,d,k} : \mathbb{R}^d \to \mathbb{R}$. We then generate $z_1, \ldots, z_T$ (where $T = 1000$) i.i.d from the distribution $\mathcal{N}(0, \frac{1}{d} I_d)$, independently of everything else and define the "maximum random gradient" to be $\max_{i \in [T]} \|\nabla f_{n,k,d}(z_i)\|$ (it is our proxy for the true Lipschitz constant $\sup_{z \in \mathbb{S}^{d-1}} \|\nabla f_{n,d,k}(z)\|$). Our experimental results are as follows:

Experiment 1: We ran experiments with n between 100 and 2000, d between ~ 50 and $\sim n$, and k between ~ 10 and $\sim n$ (we also choose $\varepsilon = 0.02$ for the thresholding). In Figure 2 we give a scatter plot of $\left(\sqrt{\frac{n}{k}}, \max_{i \in [T]} \|\nabla f_{n,k,d}(z_i)\|\right)$, and as predicted we see a linear trend, thus providing empirical evidence for Conjecture 1.

Experiment 2: In this experiment, we investigate the two extreme cases $k \sim n$ and $k \sim n/d$. We fix $n = 10^4$ and sweep the value of d between 10 to 5000 (we also choose $\varepsilon = 0.1$ for the thresholding). In the first case, we let $k = n$ and in the second case we let $k = 10n/d$. In Figure 3 we plot $\sqrt{d}$ versus the maximum random gradient (as defined above) for both cases. We observe a linear dependence between the maximum gradient value and $\sqrt{d}$ when we have $k = 10n/d$, and roughly a constant maximum gradient value when $k = n$, thus providing again evidence for Conjecture 1

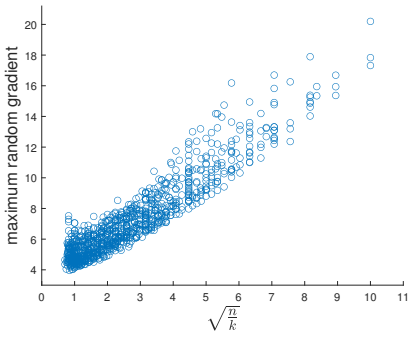

Figure 2: Scatter plot of maximum random gradient with respect to $\sqrt{\frac{n}{k}}$ with 906 data points (Experiment 1)

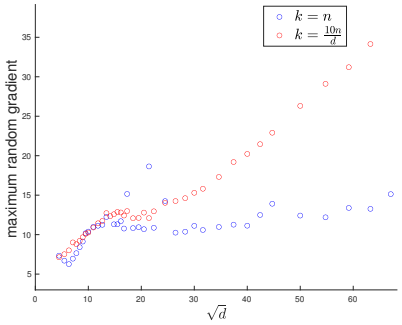

Figure 3: Scatter plot of maximum random gradient with respect to $\sqrt{d}$ in optimal smoothness (blue) and optimal size (red) regimes (Experiment 2)

6. Acknowledgements

This work was partly done while Y. Li and D. Nagaraj were visiting Microsoft Research.

References

James Alexander and André Hirschowitz. Polynomial interpolation in several variables. *Journal of Algebraic Geometry*, 4(2):201–222, 1995.

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190*, 2020.

Eric B Baum. On the capabilities of multilayer perceptrons. *Journal of complexity*, 4(3):193–215, 1988.

Guy Bresler and Dheeraj Nagaraj. A corrective view of neural networks: Representation, memorization and learning. *arXiv preprint arXiv:2002.00274*, 2020.

Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840, 2019.

Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and weights size for memorization with two-layers neural networks. arXiv preprint arXiv:2006.02855, 2020.

Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Akshay Degwekar, Preetum Nakkiran, and Vinod Vaikuntanathan. Computational limitations in robust classification and win-win results. volume 99 of *Proceedings of Machine Learning Research (COLT)*, pages 994–1028, 2019.

Shmuel Friedland and Lek-Heng Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281, 2018.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

Zhening Li, Yuji Nakatsukasa, Tasuku Soma, and André Uschmajew. On orthogonal tensors and best rank-one approximation ratio. *SIAM Journal on Matrix Analysis and Applications*, 39(1): 400–425, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 2004.

Grigoris Paouris, Petros Valettas, and Joel Zinn. Random version of dvoretzkys theorem in lpn. *Stochastic Processes and their Applications*, 127(10):3187–3227, 2017.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. In *International Conference on Learning Representations*, 2019.

Bruce Arie Reznick. *Sum of even powers of real linear forms*, volume 463. American Mathematical Soc., 1992.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Practice*, pages 210–268. Cambridge University Prteess, 2012.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small relu networks are powerful memorizers: a tight analysis of memorization capacity. In *Advances in Neural Information Processing Systems*, pages 15532–15543, 2019.

Appendix A. Results on tensors

A tensor of order p is an array $T = (T_{i_1,\dots,i_p})_{i_1,\dots,i_p \in [d]}$. The Frobenius inner product for tensors is defined by:

$$\langle T, S \rangle = \sum_{i_1,\dots,i_p=1}^{d} T_{i_1,\dots,i_p} S_{i_1,\dots,i_p} \,,$$

with the corresponding norm $\|\cdot\|_{\mathrm{F}}$. A tensor is said to be of rank 1 if it can be written as:

$$T = u_1 \otimes \ldots \otimes u_p \,,$$

for some $u_1, \dots, u_p \in \mathbb{R}^d$. The operator norm $\|\cdot\|_{\mathrm{op}}$ is defined by:

$$\|T\|_{\mathrm{op}} = \sup_{S \text{ rank } 1, \|S\|_{\mathrm{F}} \leq 1} \langle T, S \rangle \,.$$

For symmetric tensors (i.e., such that the entries of the array are invariant under permutation of the p indices), Banach's Theorem (see e.g., [(2.32), Nemirovski (2004)]) states that in fact one has

$$\|T\|_{\mathrm{op}} = \sup_{x \in \mathbb{S}^{d-1}} \langle T, x^{\otimes p} \rangle \,. \tag{9}$$

We refer to Friedland and Lim (2018) for more details and background on tensors. We now list a couple of useful results, with short proofs.

Lemma 10 *For a tensor T of order p, one has*

$$\mathrm{Lip}_{\mathbb{S}^{d-1}}(x \mapsto \langle T, x^{\otimes p} \rangle) \leq p \cdot \|T\|_{\mathrm{op}} \,.$$

Proof One has for any $x, y \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
\left| \langle T, x^{\otimes p} \rangle - \langle T, y^{\otimes p} \rangle \right| &\leq \sum_{q=1}^{p} \left| \langle T, x^{\otimes p-q+1} \otimes y^{\otimes q-1} \rangle - \langle T, x^{\otimes p-q} \otimes y^{\otimes q} \rangle \right| \\
&\leq p \cdot \|x - y\| \cdot \sup_{x^1,\dots,x^p \in \mathbb{S}^{d-1}} \left| \langle T, \otimes_{q=1}^{p} x^q \rangle \right| \\
&= p \cdot \|x - y\| \cdot \|T\|_{\mathrm{op}} \,.
\end{aligned}$$

■

Lemma 11 *For any tensor T of order p, there exists $w_1, \dots, w_{2^p d^{p-1}} \in \mathbb{R}^d$ and $\xi_1, \dots, \xi_{2^p d^{p-1}} \in \{-1, +1\}$ such that for all $x \in \mathbb{R}^d$,*

$$\langle T, x^{\otimes p} \rangle = \sum_{\ell=1}^{2^p d^{p-1}} \xi_\ell \cdot (w_\ell \cdot x)^p \,.$$

Results like Lemma 11 go back at least to Reznick (1992). In fact much more precise results on minimal decomposition in rank-1 tensors are known thanks to the work of Alexander and Hirschowitz (1995). We refer to (Comon et al., 2008) for more discussion on this topic.

Proof First note that trivially T can be written as:

$$T = \sum_{i_1,\dots,i_{p-1}=1}^{d} e_{i_1} \otimes \dots \otimes e_{i_{p-1}} \otimes T[i_1,\dots,i_{p-1},1:d]\,. \tag{10}$$

Thus one only needs to prove that a function of the form $x \mapsto \prod_{q=1}^{p}(w_q \cdot x)$ can be written as the sum of 2^p functions of the form $(w' \cdot x)^p$. To do so note that, with ε_q i.i.d. random signs,

$$\mathbb{E}\left[\prod_{q=1}^{p}\varepsilon_q \cdot \left(\sum_{q=1}^{p}\varepsilon_q w_q \cdot x\right)^p\right] = \mathbb{E}\left[\prod_{q=1}^{p}\varepsilon_q \cdot \sum_{q_1,\dots,q_p=1}^{p}\left(\prod_{r=1}^{p}\varepsilon_{q_r} w_{q_r} \cdot x\right)\right] = p!\prod_{q=1}^{p}(w_q \cdot x)\,.$$

■

Lemma 12 *For any tensor T of order p one has:*

$$\|T\|_{\mathrm{op},*} \leq d^{p-1} \cdot \|T\|_{\mathrm{op}}\,.$$

The above result and its proof are directly taken from Li et al. (2018). We only repeat the argument here for sake of completeness.

Proof Note that the decomposition (10) is orthogonal, and thus for any tensor S of order p one has:

$$\begin{aligned}
\langle T, S\rangle &\leq \sqrt{d^{p-1} \cdot \sum_{i_1,\dots,i_{p-1}=1}^{d} \langle e_{i_1} \otimes \dots \otimes e_{i_{p-1}} \otimes T[i_1,\dots,i_{p-1},1:d], S\rangle^2} \\
&\leq \sqrt{d^{p-1} \cdot \|S\|_{\mathrm{op}}^2 \cdot \sum_{i_1,\dots,i_{p-1}=1}^{d} \|T[i_1,\dots,i_{p-1},1:d]\|^2} \\
&= d^{\frac{p-1}{2}} \cdot \|S\|_{\mathrm{op}} \cdot \|T\|_F\,.
\end{aligned}$$

Thus one has $\|T\|_{\mathrm{op},*} \leq d^{\frac{p-1}{2}} \cdot \|T\|_F$. By duality one also has $\|T\|_{\mathrm{op}} \geq d^{-\frac{p-1}{2}} \cdot \|T\|_F$, which concludes the proof.

■

Appendix B. Results on random tensors

Lemma 13 *For any fixed $x \in \mathbb{S}^{d-1}$ and generic data, with probability at least $1 - C\exp(-c_p\tau)$ one has:*

$$\left|\sum_{i=1}^{n} y_i (x_i \cdot x)^p\right| \leq C_p\sqrt{\frac{n\tau}{d^p}}\,.$$

Proof Using [Theorem 1, Paouris et al. (2017)] one has, for any fixed $x \in \mathbb{S}^{d-1}$ and $\tau \leq n$,

$$\mathbb{P}\left(\left|d^{p/2}\sum_{i=1}^{n}|x_i \cdot x|^p - n\sigma_p\right| > C_p\sqrt{n\tau}\right) \leq C\exp(-c_p\tau)\,,$$

where σ_p denotes the p^{th} moment of the standard Gaussian. Let us denote $n^+ = |\{i \in [n] : y_i = +1\}$ and $T^+ = \sum_{i:y_i=+1} x_i^{\otimes p}$, and similarly for n^-, T^-. Now with probability $1 - C\exp(-c\tau)$ (with respect to the randomness of the $y_i's$) we have

$$|n^+ - n^-| \leq \sqrt{n\tau}\,.$$

Thus combining the two above displays we obtain with probability at least $1 - C\exp(-c_p\tau)$,

$$d^{p/2}\left|\sum_{i:y_i=+1}|x_i \cdot x|^p - \sum_{i:y_i=-1}|x_i \cdot x|^p\right| \leq C_p\sqrt{n\tau} + \sigma_p|n^+ - n^-| \leq C_p\sqrt{n\tau}\,,$$

■

Lemma 14 *For generic data, with probability at least* $1 - C\exp(-c_p d)$ *one has:*

$$\left\|\sum_{i=1}^{n} y_i x_i^{\otimes p}\right\|_{\mathrm{op}} \leq C_p\sqrt{\frac{n}{d^{p-1}}}\,.$$

Proof Let $\mathcal{N}$ be an $\frac{1}{2p}$-net of $\mathcal{S}^{d-1}$ (in particular $|\mathcal{N}| \leq C_p^d$). By an union bound and Lemma 13 one has:

$$\mathbb{P}\left(\exists x \in \mathcal{N}_\varepsilon \;:\; \left|\sum_{i=1}^{n} y_i|x_i \cdot x|^p\right| > C_p\sqrt{\frac{n}{d^{p-1}}}\right) \leq C\exp(-c_p d)\,, \tag{11}$$

Let $T = \sum_{i=1}^{n} y_i x_i^{\otimes p}$. Note that T is symmetric, and thus thanks to (9) and Lemma 10, one has:

$$\|T\|_{\mathrm{op}} \leq \max_{x\in\mathcal{N}}\langle T, x^{\otimes p}\rangle + \frac{1}{2}\|T\|_{\mathrm{op}}\,,$$

and in particular $\|T\|_{\mathrm{op}} \leq 2\max_{x\in\mathcal{N}}\langle T, x^{\otimes p}\rangle$, which together with (11) concludes the proof. ■

Appendix C. Proof of Theorem 3

Let $m = \frac{n}{k}$ (by assumption $m \leq c \cdot \frac{d}{\log(n)}$) and assume it is an integer. Let us choose m points with the same label, say it is the points $x_1, \ldots, x_m$ with label $+1$. As in Section 3.1 let $w \in \mathbb{R}^d$ be the minimal norm vector that satisfy $w \cdot x_i = 1$, and thus as we proved there with probability at least $1 - \exp(C - cd)$ one has $\|w\| \leq \sqrt{2m}$. Crucially for the end of the proof, also note that the distribution of w is rotationally invariant. Next observe that with probability at least $1 - 1/n^C$ (with respect to the sampling of $x_{m+1}, \ldots, x_n$) one has $\max_{i\in\{m+1,\ldots,n\}}|w \cdot x_i| \leq C \cdot \|w\|\sqrt{\frac{\log(n)}{d}} \leq \frac{1}{2}$.

In particular the cap $\mathcal{C} := \{x \in \mathbb{S}^{d-1} : w \cdot x \geq \frac{1}{2}\}$ contains $x_1, \ldots, x_m$ but does not contain any x_i, $i > m$. Thus the neuron

$$x \mapsto 2 \cdot \mathrm{ReLU}\left(w \cdot x - \frac{1}{2}\right) ,$$

computes the value 1 at points $x_1, \ldots, x_m$ and the value 0 at the rest of the training set.

One can now repeat this process, and build the neurons $w_1, \ldots, w_k$ (all with norm $\leq \sqrt{2m}$), so that (with well-chosen signs $\xi_\ell \in \{-1, 1\}$) the data is perfectly fitted by the function:

$$f(x) = \sum_{\ell=1}^{k} 2 \cdot \xi_\ell \cdot \mathrm{ReLU}\left(w_\ell \cdot x - \frac{1}{2}\right) .$$

It only remains to estimate the Lipschitz constant. Note that if a point $x \in \mathbb{S}^{d-1}$ activates a certain subset $A \subset \{1, \ldots, k\}$ of the neurons, then the gradient at this point is $\sum_{\ell \in A} w'_\ell$ with $w'_\ell = 2\xi_\ell w_\ell$. Using that the $w_{\ell'}$ are rotationally invariant, one also has with probability at least $1 - Cn\exp(-cd)$ that $\left\|\sum_{\ell \in A} w'_\ell\right\|^2 \leq C \cdot |A| \cdot m$ for all $A \subset \{1, \ldots, k\}$. Thus it only remains to control how large A can be. We show below that $|A| \leq Cm\log(d)$ with probability at least $1 - C\exp(-cd\log(d))$ which will conclude the proof.

If x activates neuron ℓ then $w_\ell \cdot x \geq \frac{1}{2} \geq \frac{\|w_\ell\|}{4\sqrt{m}}$. Now note that for any fixed $x \in \mathbb{S}^{d-1}$ and fixed $A \subset [k]$, $\mathbb{P}\left(\forall \ell \in A, w_\ell \cdot x \geq \frac{\|w_\ell\|}{4\sqrt{m}}\right) \leq C\exp\left(-c|A|\frac{d}{m}\right)$, so that

$$\mathbb{P}\left(\exists A \subset [k] : |A| = a \text{ and } \forall \ell \in A, w_\ell \cdot x \geq \frac{\|w_\ell\|}{4\sqrt{m}}\right) \leq \exp\left(Ca\log(k) - ca\frac{d}{m}\right) .$$

In particular we conclude that with $a = Cm\log(d)$ the probability that a fixed point on the sphere activates more than a neuron is exponentially small in $d\log(d)$ (recall that $m\log(k) \leq cd$ by assumption). Thus we can conclude via an union bound on an ε-net that the same holds for the entire sphere simultaneously. This concludes the proof.