Proceedings of Machine Learning Research vol 291:1–53, 2025 38th Annual Conference on Learning Theory

Logarithmic Width Suffices for Robust Memorization

Amitsour Egosi AMITSOUR.EGOSI@WEIZMANN.AC.IL
Weizmann Institute of Science

Gilad Yehudai GY2219@NYU.EDU
Courant Institute of Mathematical Sciences, New York University

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Nika Haghtalab and Ankur Moitra

Abstract

The memorization capacity of neural networks with a given architecture has been thoroughly studied in many works. Specifically, it is well-known that memorizing N samples can be done using a network of constant width, independent of N. However, the required constructions are often quite delicate. In this paper, we consider the natural question of how well feedforward ReLU neural networks can memorize *robustly*, namely while being able to withstand adversarial perturbations of a given radius. We establish both upper and lower bounds on the possible radius for general l_p norms, implying (among other things) that width *logarithmic* in the number of input samples is necessary and sufficient to achieve robust memorization (with robustness radius independent of N).

1. Introduction

The ability of neural networks to *memorize* labeled datasets is a central question in the study of their expressive power. Given some input domain $\mathcal{X}$, output domain $\mathcal{Y}$, and dataset size N, we say that a network memorizes datasets of size N, if for every labeled dataset $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, where $|\mathcal{D}| = N$, we can find parameters such that the resulting network $f : \mathcal{X} \to \mathcal{Y}$ perfectly fits the dataset (that is, $f(x) = y$ for every labeled pair $(x, y) \in \mathcal{D}$). The main question here – which has been studied in many recent works (see Section 2 for details) – is to characterize the size/architecture of the networks that have enough expressive power to memorize any dataset of a given size N.

However, merely fitting a given dataset is not enough for most tasks, and a desirable property for trained networks is that they remain robust to noise and minor modifications in the dataset. This robustness property allows neural networks to generalize from observed data points to unseen data points. Furthermore, neural networks have been shown to be vulnerable to adversarial attacks (Szegedy et al., 2013; Carlini and Wagner, 2017; Papernot et al., 2017; Athalye et al., 2018) in the form of slightly perturbed examples, where (in the context of visual data) the perturbation is often imperceptible to the human eye. Moreover, existing constructions of memorizing networks are often quite delicate, and not at all robust to such perturbations. This motivates the question of characterizing the networks that have enough capacity to *robustly* memorize a dataset. Concretely, considering datasets of the form $\{(x_1, y_1), \ldots, (x_N, y_N)\} \subset \mathbb{R}^d \times \{1, \ldots, C\}$ in a multiclass setting, and a robustness radius $\sigma > 0$, the problem we wish to study is the following: How large does a standard feedforward ReLU network f need to be, so that for any dataset of size N as above, there exists a choice of parameters such that $f(a_i) = y_i$ for every $a_i \in B_p^d(x_i, \sigma)$ (where $B_p^d(x_i, \sigma)$ is a ball of radius σ around x_i in l_p norm).

When considering the notion of the size of a network in the problem of robust memorization, one can define it in terms of depth, width or the total number of parameters of the network. Several works have observed empirically that wider networks tend to be more robust to adversarial perturbations (Madry et al., 2017; Wu et al., 2021; Zhu et al., 2022). This connection between the radius of robustness and the necessary width for robust memorization is still not well studied. A recent work (Yu et al., 2024) showed that width $k \geq d$ is necessary for *optimal* robust memorization in l_∞ norm. Optimality in their work requires the existence of a memorizing network of width k for *all* possible robustness radii, not accounting for possible finer relation between the width of the network and the robustness radius.

In this work we study this connection between width and robustness, and in particular we seek to determine what is the minimal width k required to ensure that for any dataset there exists a width k network that can memorize it with robustness radius σ. In the non-robust case, it is known that memorization can be achieved with constant-width networks (Park et al., 2021; Vardi et al., 2021). We show that for robust memorization, there exists a trade-off between the width k and radius σ. In our analysis we consider datasets with minimal l_2 distance of δ between different classes, called δ-separated datasets. This separation assumption is necessary for robust memorization, since the robustness radius is limited by the distance between differently-labeled points (see Remark 4 for more details). Our main contributions can be summarized as follows:

- We show bounds on the possible robustness radius σ in l_p norm for memorizing a δ-separated dataset of size N using a network of width k. Specifically, the following holds for some universal constants c_1, c_2 and for every $p \in [2, \infty]$:
  1. If $\frac{\sigma}{\delta} < \frac{c_1}{d^{1-1/p}} N^{-\frac{2}{k-6}}$, then any such dataset can be robustly memorized by a network of width k (Theorem 7).
  2. If $\frac{\sigma}{\delta} > c_2 N^{-\frac{2}{k}}$ then there exists a dataset that cannot be robustly memorized by any network of width k (Theorem 8).
- Both of the results above rely on a robust variant of the Johnson-Lindenstrauss Lemma that we develop (Theorem 13), which revolves around projecting high-dimensional points to a lower-dimensional subspace while maintaining separation between neighborhoods of data points, and may be of independent interest.
- The bounds above apply to the regime where the desired width is relatively small (less than the data dimension). In addition, we develop bounds for the more permissive regime where the width is larger than the data dimension (Theorem 5 and Theorem 6), extending the results of (Li et al., 2022; Yu et al., 2024) to other norms as well as to smaller widths.

The results above show that for guaranteeing robust memorization (with robustness radius independent of N and with width smaller than the dimension), a necessary and sufficient condition is that the width would depend logarithmically on N. Alternatively, if we wish to robustly memorize with constant width k independent of N, then the robustness parameter $\frac{\sigma}{\delta}$ necessarily decays polynomially in the dataset size N. This means that constructions similar to those from Park et al. (2021); Vardi et al. (2021), which achieve optimal memorization in terms of the number of parameters and with width independent of N, cannot achieve optimal robustness.

We note that the upper bound for robust memorization depends on the input dimension d, whereas the lower bound does not. This dependence introduces a gap between these bounds (which is discussed in Subsection 4.2.1). It remains a subject for future work to investigate tight bounds on the possible robustness radius σ.

2. Related Works

Memorization Memorization in neural networks is a well studied field with many established results. Baum (1988); Bubeck et al. (2020); Huang et al. (1991); Huang and Babri (1998); Sartori and Antsaklis (1991); Zhang et al. (2021) proved under different settings that $O(N)$ neurons and parameters are enough to memorize N data points. Huang (2003); Yun et al. (2019) improved these results and showed that $O(\sqrt{N})$ neurons are enough to memorize N points with a 3-layer neural networks, although the number of parameters is still $O(N)$. Park et al. (2021) gave the first sub-linear parameter memorization bound, with $N^{2/3}$ parameters to memorize N points. Finally, Vardi et al. (2021) proved that memorizing N points can be done using a network with $\tilde{O}(\sqrt{N})$ parameters. This is known to be optimal up to log terms due to VC dimension lower-bounds (Goldberg and Jerrum, 1995; Bartlett et al., 2019). Note that the width of the memorizing networks in Park et al. (2021); Vardi et al. (2021) is a universal constant, namely 12 in Vardi et al. (2021). Also, note that our results imply that the constructions from Park et al. (2021); Vardi et al. (2021) cannot achieve optimal robustness

Robust memorization Several works proved the existence of networks that memorize robustly using different methods. Yang et al. (2020); Bastounis et al. (2021) proved there exists locally Lipschitz classifiers, which implies some form of local robustness, although they did not give specific bounds on the size of the classifier. Li et al. (2022) showed the existence of robust memorization networks through VC dimension arguments. Most closely related to our work is Yu et al. (2024), which proves upper and lower bounds for robust memorization. In particular, they show that robust memorization with the optimal robust radius in l_∞ norm (including the constants) cannot be achieved if the width is smaller than the data dimension. We extend their result by showing the intricate trade-offs between the width of the network and the robustness radius.

Robustness and width Several papers observed empirically that there is a connection between the width of the neural network and its robustness properties. Madry et al. (2017) observed that wider networks tend to be more robust, even without adversarial training. Wu et al. (2021); Zhu et al. (2022) study the effect of the width on adversarial training, and provide theoretical justification in the NTK regime (Jacot et al., 2018; Allen-Zhu et al., 2019; Gao et al., 2019). Our work focuses on the expressive capacity required for robustness, rather than the optimization process which is studied in these works.

3. Preliminaries

Notations. For every $0 < p < \infty$ and $x \in \mathbb{R}^d$ denote $\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}$, and $\|x\|_\infty = \max_{1 \le i \le d} |x_i|$. For $0 < p < 1$ the function $\|\cdot\|_p$ is a quasi-norm, and for $1 \le p \le \infty$ it is the l_p norm with an induced metric dist_p. For all $0 < p \le \infty$ we define the l_p ball of radius r around x as $B_p^d(x, r) = \left\{x' \in \mathbb{R}^d \mid \|x' - x\|_p \le r\right\}$. Note that all balls in our work are closed balls. e is Euler's number $2.718...$.

For any $0 < p \leq \infty$ we will denote $c_p^+(d) = d^{\left[\frac{1}{2}-\frac{1}{p}\right]_+}$ and $c_p^-(d) = d^{\left[\frac{1}{2}-\frac{1}{p}\right]_-}$, where $[x]_+ = \max\{0, x\}$ (also called the ReLU activation) and $[x]_- = \min\{0, x\}$. Note that $c_p^+(d)$ is the radius of the l_2 ball that encloses the unit l_p ball, and that $c_p^-(d)$ is the radius of the l_2 ball that is inscribed in the unit l_p ball (see Lemma 70 in Appendix H.4). In these definitions, for $p = \infty$ we define $\frac{1}{\infty} = 0$. For additional notations used in the appendices, see Appendix A.

Neural Networks. In this paper, we focus on feedforward ReLU neural networks, defined as follows:

Definition 1 *Let $d \in \mathbb{N}_{\geq 2}$, $L \in \mathbb{N}$ and $d_0, d_1, ..., d_L \in \mathbb{N}$ with $d_L = 1, d_0 = d$, and let $W^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$, $b^{(l)} \in \mathbb{R}^{d_l}$ for all $1 \leq l \leq L$. Denote $T^{(l)}(x) = W^{(l)}x + b^{(l)}$. We will define a **feed forward ReLU neural network** to be $f : \mathbb{R}^{d_0} \longrightarrow \mathbb{R}$ given by*

$$f = T^{(L)} \circ [\cdot]_+ \circ T^{(L-1)} \circ ... \circ [\cdot]_+ \circ T^{(1)}$$

*where $[\cdot]_+$ is applied element-wise. We will say that the **depth** of f is $\mathcal{L}(f) := L$, the **architecture** of f is $\mathcal{A}(f) := (d_0, d_1, ..., d_L)$, and the **width** of f is $\mathcal{W}(f) := \max\{d_1, ..., d_{L-1}\}$.*

Data Assumptions and Robustness. Let $N, d \in \mathbb{N}_{\geq 2}$, $0 < \delta, \sigma$. We will use δ to denote the separation distance between different data classes and σ to denote the radius of robustness. Formally, we use the following definitions:

Definition 2 *Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subseteq \mathbb{R}^d \times [C]$ be a **dataset** of size N with C classes, comprised of **data points** x_i and **labels** y_i. We will denote by $\mathcal{D}_{d,N,C}$ the set of all such datasets. We say that a dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}$ is a δ-**separated** dataset for some $0 < \delta$, if $\min\{\|x_i - x_j\|_2 \mid y_i \neq y_j\} = \delta$, and denote by $\mathcal{D}_{d,N,C}(\delta)$ the set of all such datasets.*

Definition 3 *Let $\mathcal{D} \in \mathcal{D}_{d,N,C}$, $p \in (0, \infty]$ and $0 \leq \sigma$. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ (σ, p)-**robustly memorizes** the dataset $\mathcal{D}$ if for all $i \in [N]$ and $x \in B_p^d(x_i, \sigma)$ one has $f(x) = y_i$.*

4. Main Results

In this section, we present the main theorems that connect robust memorization and the width of the memorizing neural network, as well as proof sketches (with full proofs appearing in Appendix B). In our results in this section we will use the definition of δ-separated dataset from above, where for concreteness we measure separation in terms of the l_2 norm (see Appendix E for an extension to l_q norms for any $q \in [1, \infty]$). In the following, we let $N, d, C \in \mathbb{N}_{\geq 2}$, $k \in \mathbb{N}$, $0 < \delta, \sigma$ and $p \in (0, \infty]$.

Remark 4 (Robustness parameter $\frac{\sigma}{\delta}$ cannot exceed $\frac{1}{2c_p^+(d)}$) *Given some $0 < \delta$, we wish to find the maximal possible value of σ that allows for (σ, p)-robust memorization, of any δ-separated dataset, using a width k network. In the case of σ-neighborhoods with respect to the l_2 norm, the value of $\frac{\sigma}{\delta}$ must lie in the range $[0, \frac{1}{2})$. Indeed, if we allow $\frac{\delta}{2} \leq \sigma$ then the σ-neighborhood of two data points with different labels might intersect, so we cannot ensure robust memorization. Similarly, for general l_p norms, if we allow $\frac{\delta}{2} \leq c_p^+(d)\sigma$ then two l_2 balls of radius $c_p^+(d)\sigma$ might intersect, and so their enclosed l_p balls of radius σ might intersect. Therefore, the task of guaranteeing a (σ, p)-robust memorization for **every** possible δ-separated dataset can only be considered in the range $0 \leq \frac{\sigma}{\delta} < \frac{1}{2c_p^+(d)}$.*

4.1. Robust Memorization With Large Width

We first consider the easier case, where the desired width k can be larger than the data dimension d. In this case, for all values σ in the applicable range, one can (σ, p)-robustly memorize any δ-separated dataset with a width k network:

Theorem 5 *If $d + 6 \leq k$ and $\frac{\sigma}{\delta} < \frac{1}{2c_p^+(d)}$, then for every δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth $O\left(Nd\log_2\left(\frac{d}{1-\frac{2c_p^+(d)\sigma}{\delta}}\right)\right)$ that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.*

Note that as $\frac{\sigma}{\delta}$ approaches $\frac{1}{2c_p^+(d)}$ the depth of the network grows accordingly. If however $\frac{2c_p^+(d)\sigma}{\delta}$ is bounded from above by some universal constant, we obtain depth of $O(Nd\log_2(d))$. For the special case where $p \in \{1, \infty\}$ the range of the width in Theorem 5 can be improved and the log factor in the depth of the network can be removed:

Theorem 6 *Let $p \in \{1, \infty\}$. If $d + 4 \leq k$ and $\frac{\sigma}{\delta} < \frac{1}{2c_p^+(d)}$, then for every δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth $O(Nd)$ that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.*

Theorems 5 and 6 do not depend on the support of the dataset, and for fixed ratio σ/δ, the depth we obtain does not depend on δ. Furthermore, we allow for robust neighborhoods under l_p for any $p \in (0, \infty]$. Thus, our results extend the results in (Li et al., 2022, Theorem 2.2) and (Yu et al., 2024, Theorem B.6). We further extend Theorems 5 and 6 to allow for any choice of both separation and robustness norms in Appendix E.1.

4.2. Robust Memorization With Small Width

We now turn to study the more challenging case where the desired width is smaller than the data dimension, which is our main contribution. The theorem below shows that in this regime, it is still possible to (σ, p)-robustly memorize any δ-separated dataset with a width k network, provided that the radius of robustness σ is small enough:

Theorem 7 *Suppose $7 \leq k \leq d + 5$ and*

$$\frac{\sigma}{\delta} \leq a_{p,d} N^{-\frac{2}{k-6}}\ , \quad \textit{where} \ \ a_{p,d} := \frac{1}{8\sqrt{e}} d^{-\frac{1}{2}+\left[\frac{1}{p}-\frac{1}{2}\right]_-}\ .$$

Then for every δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth $O(Nk\log_2(k))$ that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.

The amount by which $\frac{\sigma}{\delta}$ has to be small depends on the desired width k, input dimension d, the robustness metric l_p and on the dataset size N. The bound on $\frac{\sigma}{\delta}$ in Theorems 5 does not depend on N, and so one can then ask if the dependence on N in Theorem 7 can be improved. The next theorem shows that any improvement of the bound will still have a similar dependence on N, and that a bound of the form $\frac{\sigma}{\delta} < CN^{-\frac{2}{k}}$, is a necessary requirement for the case of small width k:

Theorem 8 *Suppose* $1 \leq k \leq d-1$ *and*

$$\frac{\sigma}{\delta} > b_{p,d} N^{-\frac{2}{k}} \;, \;\; \textit{where} \;\; b_{p,d} := 2416 d^{\left[\frac{1}{p}-\frac{1}{2}\right]_+} \;.$$

Then there exists a δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,2}(\delta)$, such that every neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and any depth cannot (σ, p)-robustly memorize the dataset $\mathcal{D}$.

From Theorem 7, we get that if

$$6 + \frac{2}{\log\left(\frac{\delta}{\sigma} a_{p,d}\right)} \log(N) < k \;, \tag{1}$$

then every δ-separated dataset of size N can be (σ, p)-robustly memorized by a width k neural network. On the other hand, from Theorem 8 we get that if

$$k < \frac{2}{\log\left(\frac{\delta}{\sigma} b_{p,d}\right)} \log(N) \;, \tag{2}$$

then, there exists a δ-separated dataset of size N that cannot be (σ, p)-robustly memorized by any width k neural network. Hence from Equations (1) and (2) we conclude the following corollary:

Corollary 9 *Let $p \in [2, \infty]$. There exists universal constants C_1, C_2 s.t. in the regime $k < d$,*

- *A width of $k > C_1 \frac{\log(N)}{\log\left(\frac{\delta}{\sigma}\right) + \log\left(\frac{1}{d}\right)}$ is sufficient for (σ, p)-robust memorization of every δ-separated dataset of size N.*
- *A width of $k > C_2 \frac{\log(N)}{\log\left(\frac{\delta}{\sigma}\right)}$ is necessary for (σ, p)-robust memorization of every δ-separated dataset of size N.*

We thus see that indeed in order to perform robust memorization with robustness radius independent of N and with width smaller than the data dimension d, a dependence logarithmic in N is both a necessary and a sufficient condition for the width.

Remark 10 (Fixed ratio k/d) *In the proof of Theorem 7 we are in fact proving a better bound of the form $\frac{\sigma}{\delta} \leq a_{p,d}\sqrt{k-6} \cdot N^{-\frac{2}{k-6}}$, which is of the order of $\sqrt{\frac{k}{d}} \cdot N^{-\frac{2}{k}}$ when $p = 2$. Therefore, in the regime where k/d is fixed (and $p = 2$), we get the following (for some constants C_1, C_2):*

- *A width of $k > C_1 \frac{\log(N)}{\log\left(\frac{\delta}{\sigma}\right)}$ is sufficient for (σ, p)-robust memorization of every δ-separated dataset of size N.*
- *A width of $k > C_2 \frac{\log(N)}{\log\left(\frac{\delta}{\sigma}\right)}$ is necessary for (σ, p)-robust memorization of every δ-separated dataset of size N.*

As discussed in the introduction, (Yu et al., 2024) showed that for $p = \infty$, achieving optimal robust memorization (i.e for every $\sigma < \frac{\delta}{2}$) is not possible when $k < d$. In contrast, Remark 10 implies that even when $c_1 d < k < d$, *nearly*-optimal robust memorization is still possible, i.e. for every $\sigma < c_2 \frac{\delta}{2}$ (for some universal constants $0 < c_1, c_2 < 1$).

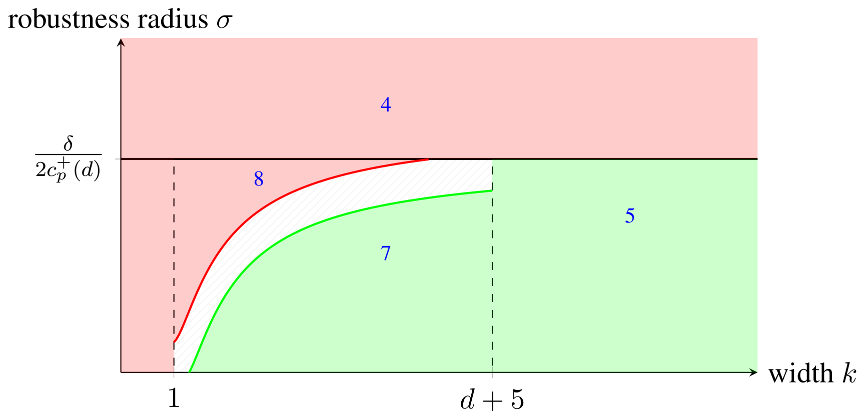


Figure 1: Illustration of main results describing regions where robust memorization is possible (green), not possible (red) and unknown (gray stripes). k is the width, σ the radius of robustness and δ the separation distance of the dataset of dimension d. Remark 4 and Theorems 5, 7, 8 are indicated in the regions that they discuss.

Remark 11 (Non-robust memorization) *In the case of non-robust memorization, i.e when $\sigma = 0$, we get from Theorem 7 that memorization is possible with networks whose width is a universal constant (namely, 7). This is consistent with previous results in Park et al. (2021); Vardi et al. (2021) about non-robust memorization.*

Theorems 7 and 8 can also be interpreted as results on the dependence between robustness radius and width. Fixing δ, d, p and N, we get bounds for the values of the radius σ for which robust memorization is always possible, as a function of the desired width k of the memorizing network. Both the upper bound from Theorem 7 (green curve in Figure 1) and the lower bound from Theorem 8 (red curve in Figure 1) are proportional to $N^{-\frac{2}{k}}$. The gap between them (gray stripes in Figure 1) stems from the difference between the terms $a_{p,d}$ and $b_{p,d}$.

4.2.1. THE GAP BETWEEN $a_{p,d}$ AND $b_{p,d}$

We proceed to discuss the gap between the upper bound in Theorem 7 and the lower bound in Theorem 8, and specifically the gap between the multiplicative factors $a_{p,d}$ and $b_{p,d}$. Note that by definition of $c_p^+(d), c_p^-(d)$ we have $a_{p,d} = \frac{1}{8\sqrt{e}} d^{-\frac{1}{2}+\left[\frac{1}{p}-\frac{1}{2}\right]_-} = \frac{1}{2c_p^+(d)} \cdot \frac{1}{\sqrt{16ed}}$, and that $b_{p,d} = 2416 d^{\left[\frac{1}{p}-\frac{1}{2}\right]_+} = \frac{1}{2c_p^-(d)} \cdot 4832$. The $c_p^+(d)$ factor in $a_{p,d}$ comes from the fact that we have to ensure that the *enclosing* l_2 balls of the l_p neighborhoods are disjoint (as discussed in Remark 4). On the other hand, the $c_p^-(d)$ factor in $b_{p,d}$ comes from the fact that we have to show that l_p neighborhoods in the constructed dataset intersect, and we do that by showing that the *inscribed* l_2 balls in them intersect. The gap between $a_{p,d}$ and $b_{p,d}$ is thus given by

$$\frac{b_{p,d}}{a_{p,d}} = c \cdot \frac{c_p^+(d)}{c_p^-(d)} \sqrt{d} = c \cdot \sqrt{d} \cdot d^{\left|\frac{1}{2}-\frac{1}{p}\right|}$$

for $c = 19328\sqrt{e}$. The need to reduce the dimension of the data with a linear map (see proof intuition in Subsection 4.3) introduces the dependence $d^{|\frac{1}{2}-\frac{1}{p}|}$ on d in the gap between the bounds, for all $p \neq 2$. In the special case that $p = 2$, this dependence vanishes and we are left with a gap of $\sqrt{d}$. The reason for this gap stems from the non-tightness of our robust variant of the Johnson-Lindenstrauss lemma, which will be presented in Section 5.

4.3. Proof Intuition

We now turn to provide a sketch for the proof of our main results.

We begin by discussing the proof of Theorem 5, which shows that for width larger than the dimension and sufficient depth, one can robustly memorize any δ-separated dataset for any applicable robustness radius. As discussed in Remark 4, since the separation is measured in l_2 norm, the l_p balls of radius σ around data points (from different classes) must be contained in disjoint l_2 balls of radius $r = c_p^+(d)\sigma$. Hence, a function that assigns each of these l_2 balls with its appropriate label will (σ, p)-robustly memorize the data. Given a collection of labeled l_2 balls in $\mathbb{R}^d$ we can perform this assignment using a function that computes the weighted sum of ball indicators $\sum_{i=1}^{N} y_i \cdot 1_{B_2^d(x_i,r)}$ over all N data points.

Since exact computation of the l_2 norm is not possible with ReLU networks, we first approximate the function $y_i \cdot 1_{B_2^d(x_i,r)}$ using the function

$$f_i(x) = \begin{cases} y_i & \|x - x_i\|_2 \le r \\ v(x) & r < \|x - x_i\|_2 \le r + w \\ 0 & r + w < \|x - x_i\|_2 \end{cases},$$

where $v(x)$ is some value bounded by y_i, and $w = \delta - 2r$. We then approximate f_i (and specifically $\|x - x_i\|_2$) using a ReLU network, by sequentially approximating for every $1 \le j \le d$ the square of each coordinate of the vector $x - x_i$. The resulting network completes the proof of Theorem 5. In the case that $p \in \{1, \infty\}$ we can replace the approximation of $\|x - x_i\|_2$ with a ReLU network that computes exactly the norm $\|x - x_i\|_p$, removing the logarithmic factor in the depth and yielding Theorem 6.

Performing the necessary computations for each of the coordinates $1 \le j \le d$ as above means that the width of the resulting network is at least d (computing sequentially, as we did, still requires the propagation of the input vector in each layer for future computations). Therefore, handling the regime where the width is smaller than d must involve some dimensionality reduction of the dataset. Specifically, if the desired width is k, then the first layer of any memorizing network must implement a linear mapping that reduces the dimension of the dataset to at most k dimensions (see Definition 1). If there is any hope to robustly memorize the dataset, this map cannot introduce an intersection between the (σ, p)-neighborhoods of points from different classes: Indeed, let T be a linear map of rank k such that there exists $(x_i, y_i), (x_j, y_j)$ with $y_i \neq y_j$ and $T(B_p^d(x_i, \sigma)) \cap T(B_p^d(x_j, \sigma)) \neq \emptyset$. Then any network whose first layer is T, cannot (σ, p)-robustly memorize the dataset.

To ensure that in general $T(B_p^d(x_i, \sigma))$, $T(B_p^d(x_j, \sigma))$ are disjoint we have to require that $T(B_2^d(x_i, r))$, $T(B_2^d(x_j, r))$ are disjoint (recall that $r = c_p^+(d)\sigma$). The first step is thus to characterize the conditions that guarantee the ability or lack thereof to linearly map any δ-separated dataset to k dimensions while preserving separation and avoiding intersection of the images of the

l_2 neighborhoods. In Section 5 we discuss the existence of such mappings. In the positive direction we find conditions that ensure the existence of such a map T. Normalizing T appropriately, we obtain a map T' that shares the properties of T and also satisfies that $T'(B_2^d(x_i, r))$ is an l_2 ball in $\mathbb{R}^k$. Composing this map T' with the network from Theorem 5 (where now the dimension of the data is k) yields Theorem 7. In the negative direction, we find conditions that allow us to construct a δ-separated dataset that no T of rank k can preserve, establishing the proof of Theorem 8.

5. Preserving Linear Maps

As discussed at the end of the previous section, a key tool that we need in order to establish robust memorization with small width is the existence of a linear transformation, which maps a given δ-separated dataset into a lower-dimensional subspace, while preserving a separation between the neighborhoods of points. More formally, we define a *preserving* linear map into k dimensions (with respect to the l_2 norm) as follows:

Definition 12 *Let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset, and let $\sigma < \frac{\delta}{2}$. We say that a linear function $T : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ (σ, k)-**preserves** $\mathcal{D}$ if*

$$\|a - a'\|_2 \leq \|T(a) - T(a')\|_2$$

for every $(x, y), (x', y') \in \mathcal{D}$ with $y \neq y'$ and every $a \in B_2^d(x, \sigma), a' \in B_2^d(x', \sigma)$.

Note that Definition 12 requires only a lower bound on the norm of $T(a - a')$, without an upper bound. This is different from the usual notions of approximate isometry (as used, for example, in the celebrated Johnson-Lindenstrauss lemma), and results from the fact that we are only interested in preventing unwanted intersections. As a result, the definition does not involve a scaling factor ϵ of the type $\epsilon\|a - a'\|_2 \leq \|T(a) - T(a')\|_2$ since one can always scale the linear map by $1/\epsilon$ to obtain a map that satisfies the above definition. We further discuss the connection between our results and the Johnson-Lindenstrauss lemma in Subsection 5.2.

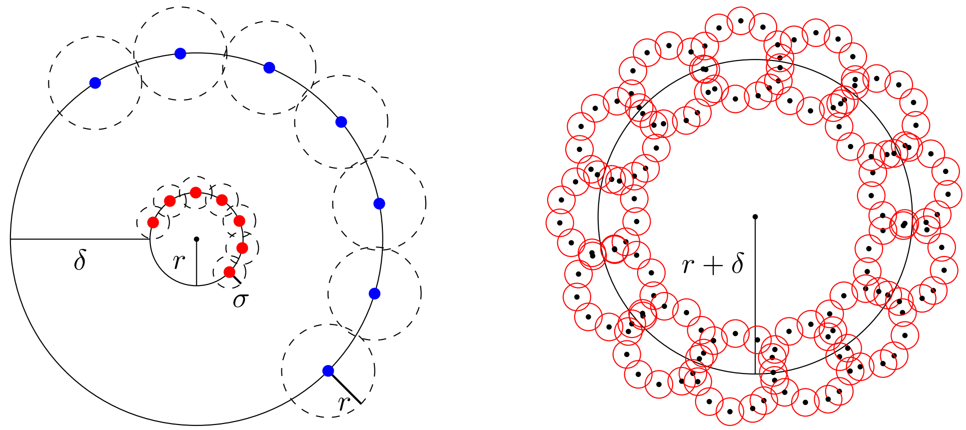
We are interested in the problem of determining the conditions under which such a linear map exists. Concretely, our problem can be formulated in the following manner:

> *Let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset. Under what conditions on N, δ, σ, d, k can we guarantee that there exists a linear map $T : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ that (σ, k)-preserves $\mathcal{D}$?*

5.1. Conditions for the Existence of a Preserving Linear Map

Following the discussion in Remark 4 we know that since we deal with l_2 neighborhoods of radius σ, the ratio $\frac{\sigma}{\delta}$ can only be considered in the range $[0, \frac{1}{2})$. Any σ such that $\frac{1}{2} \leq \frac{\sigma}{\delta}$ would cause intersecting l_2 neighborhoods, and thus any linear transformation will not be preserving as defined above. On the other hand, when $\frac{\sigma}{\delta} = 0$, any finite dataset has a $(0, k)$-preserving map for every k (since then it suffices that distinct data points have distinct images, which is always possible). Thus, the question is which values of $\frac{\sigma}{\delta} \in (0, \frac{1}{2})$ allow for the existence of (σ, k)-preserving linear maps.

The main result of this section is the following theorem, which provides an almost tight characterization:



(a) An illustration of the dataset. Points from the first class are colored red, and points from the second class are colored blue.

(b) The set of points $x'_j - x_i$ and their σ-neighborhoods colored red.

Figure 2: A dataset that cannot be (σ, k)-preserved.

Theorem 13 *Let $N, d, C \in \mathbb{N}_{\geq 2}$ such that $1 \leq k \leq d-1$ and let $\sigma < \frac{\delta}{2}$. There exists universal constants C_1, C_2 such that*

1. *If $\frac{\sigma}{\delta} < C_1 \sqrt{\frac{k}{d}} N^{-\frac{2}{k}}$ then every δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$, has a (σ, k)-preserving linear map.*

2. *If $\frac{\sigma}{\delta} > C_2 N^{-\frac{2}{k}}$ then there exists a δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,2}(\delta)$, for which no (σ, k)-preserving linear map exists.*

We present here a brief sketch of the main proof ideas. The full proof follows immediately from Theorems 18 and 19 in Appendix B.2.2.

We first discuss the positive result (in item 1 of Theorem 13). Given a δ-separated dataset $\mathcal{D}$, we consider the collection of normalized differences:

$$S = \left\{ \frac{a - a'}{\|a - a'\|_2} \mid a \in B_2^d(x, \sigma), a' \in B_2^d(x', \sigma) \text{ s.t } (x, y), (x', y') \in \mathcal{D} \text{ with } y \neq y' \right\} .$$

Note that by definition, a linear map T will (σ, k)-preserve $\mathcal{D}$ if and only if $1 \leq \|Ts\|_2$ for every $s \in S$. We then show using a probabilistic argument that if $\frac{\sigma}{\delta}$ is small enough, there exists some orthogonal projection matrix P of rank k such that $\epsilon \leq \|Ps\|_2$ for every $s \in S$, where ϵ is some value in $[0, \frac{1}{2}]$. Taking $T = \frac{1}{\epsilon} P$ proves item 1 of Theorem 13.

For the negative result in item 2 of Theorem 13, we construct a δ-separated dataset that cannot be (σ, k)-preserved. To do so, consider some origin-centered $(k+1)$-dimensional ball embedded in $\mathbb{R}^d$, which we will denote as $\tilde{B}_2^{k+1}(0, r)$ (for some $r > 0$). When $\frac{\sigma}{\delta}$ is big enough, there are enough points to construct the following δ-separated dataset, containing 2 classes: One class, $\{x_1, ..., x_{N/2}\}$, will be the centers of a σ-cover of the boundary $\partial \tilde{B}_2^{k+1}(0, r)$ (red points in Figure 2(a)), and the other class $\{x'_1, ..., x'_{N/2}\}$ will be the centers of an r-cover of the boundary of a larger $(k+1)$-dimensional embedded ball $\tilde{B}_2^{k+1}(0, r + \delta)$ (blue points in Figure 2(a)). These two classes comprise together a δ-separated dataset since $\delta \leq \|x_i - x'_j\|_2$ for any i, j. Now, define the collection of σ-neighborhoods (see Figure 2(*b*))

$$U = \left\{ a'_j - a_i \mid a'_j \in B_2^d(x'_j, 0), a_i \in B_2^d(x_i, \sigma) \right\} .$$

By the construction of U, it follows that for every point in the boundary $x \in \partial \tilde{B}_2^{k+1}(0, r+\delta)$, there exists some $u \in U$ such that $u \in \text{Span}\{x\}$. On the other hand, since $\partial \tilde{B}_2^{k+1}(0, r+\delta)$ is the boundary of a $(k+1)$-dimensional ball embedded in $\mathbb{R}^d$, it follows from dimensionality considerations that any $d-k$ dimensional subspace K must intersect some point in the ball's boundary. Taken together, we obtain that for every $d-k$ dimensional subspace K, there exists some $u \in U \cap K$. Let T be a linear map of rank k, and denote $K = \text{Ker}T$. We conclude from the above that there exists some $u = a'_j - a_i$ such that $u \in K$ and so by definition T does not (σ, k)-preserve $\mathcal{D}$.

As explained in the proof intuition of our main results (Subsection 4.3), Theorems 7 and 8 are proved by using the existence or lack-thereof of a preserving map T. The $\sqrt{\frac{k}{d}}$ gap between item 1 and 2 of Theorem 13 is thus the reason for the $\sqrt{d}$ gap between Theorem 7 and Theorem 8 that is mentioned in Subsection 4.2.1. We refer the reader to Appendix D for further discussion of this $\sqrt{\frac{k}{d}}$ gap.

5.2. Comparison to the Johnson–Lindenstrauss Lemma

The problem of finding a preserving linear map as in Theorem 13 bears similarities to the problem addressed by the Johnson–Lindenstrauss (JL) lemma (see Dasgupta and Gupta (2003)). Informally, the JL lemma states that a high-dimensional datasets can be embedded into a subspace of much lower dimension while approximately preserving distances. Formally:

Theorem 14 (JL lemma) *Let $X \subset \mathbb{R}^d$ with $|X| = N$, and let $0 < \epsilon < 1$. If $\frac{C_{JL} \ln(N)}{\epsilon^2} < k$ (where C_{JL} is some universal constant), then there exists a linear map $T : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ such that for every $x, x' \in X$*

$$(\sqrt{1-\epsilon})\|x-x'\|_2 \leq \|T(x)-T(x')\|_2 \leq (\sqrt{1+\epsilon})\|x-x'\|_2 . \tag{3}$$

A map that satisfies Eq. (3) is called a JL map.

The JL lemma provides conditions for a (σ, k)-preserving map of any δ-separated dataset in the case where $\sigma = 0$. However, satisfying the JL map condition alone (as defined in the theorem) is not enough to ensure (σ, k)-preservability for general $0 \leq \sigma < \frac{\delta}{2}$. This is because a JL map provides approximate isometry for the data points themselves but can still stretch the space in a way that turns their σ-neighborhoods into hyper-ellipsoids that intersect. Indeed, below is a simple example of a dataset where for every $\sigma > 0$ there exists a JL map that is not (σ, k)-preserving (see Figure 3 for an illustration):

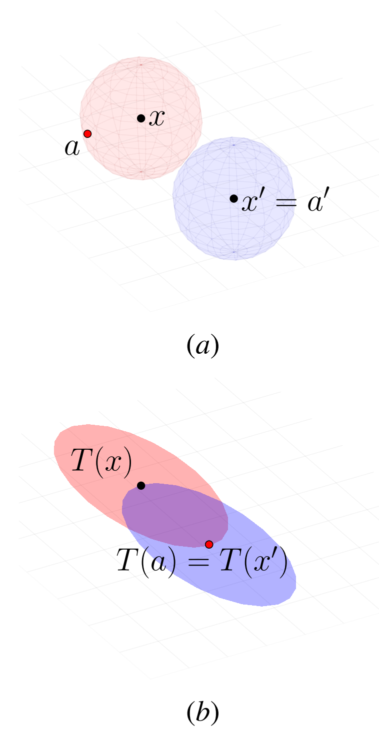


Figure 3: The dataset $\mathcal{D}$ before (top) and after (bottom) applying T. Distance between the data points x, x' is preserved but the images of their σ-neighborhoods intersect.

Example 1 *Let $x = 0, x' = e_1 \in \mathbb{R}^d$, $\mathcal{D} = \{(x, 0), (x', 1)\}$.*
Note that $\mathcal{D}$ is a 1-separated dataset and let $0 < \sigma < 1/2$. Let $0 < \epsilon < 1$ and $\frac{C_{JL} \ln(2)}{\epsilon^2} < k \leq d-1$. Define $T : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ by

$$T\left(\sum_{i=1}^{d} \alpha_i e_i\right) = \left(\alpha_1 - \frac{1}{\sigma}\alpha_2\right) e_1 + \sum_{i=2}^{k} \alpha_{i+1} e_i \ .$$

Then, $\|T(x') - T(x)\|_2 = \|e_1\|_2 = \|x'\|_2 = \|x' - x\|_2$ and so T is a JL map. However, T does not (σ, k)-preserve $\mathcal{D}$: Taking $a = -\sigma e_2$ and $a' = x'$, we have $a \in B_2^d(x, \sigma)$, $a' \in B_2^d(x', \sigma)$, yet $T(a') - T(a) = T(a' - a) = T\left(e_1 + \sigma e_2\right) = (1 - \frac{1}{\sigma}\sigma)e_1 = 0$.

Example 1 shows that simply using the standard formulation of the JL lemma is not enough to obtain a (σ, k)-preserving map. This is because, a priori, the JL map obtained from the JL lemma does not preserve neighborhoods. Our proof for item 1 of Theorem 13 is a variant of the proof of the JL lemma, that enables us to guarantee also a bound on the distortion of neighborhoods.

6. Conclusion

In this paper, we showed that for the task of robust memorization of datasets of size N, there exists a trade-off between the radius of robustness σ and the width k of the memorizing network. We showed that in the regime where the width is less than the data dimension, robust memorization can only be done with a robustness radius of $N^{-\frac{2}{k}}$ (up to a constant), and in particular achieving the optimal robust memorization capacity (up to a constant) can only be done with width k logarithmic in N. This is in contrast to the non-robust case where constant width is sufficient (see Vardi et al. (2021)).

To obtain our bounds, we develop a robust variant of the Johnson-Lindenstrauss lemma (Theorem 13) with almost tight bounds. An interesting question for future work is the tightness of our bounds in Theorem 13. Namely, establishing whether the radius of robustness has to depend on the data dimension d in order to guarantee the ability to linearly reduce the dataset to k dimensions in a robust manner. Answering this question would provide a complete and full characterization of the relation between width and robustness radius under the l_2 norm.

Another possible direction for future work is to study the relation between robustness and number of parameters. In our work we focused mostly on the width, and for a constant width our construction requires $O(N)$ parameters. In the non-robust case, $\tilde{O}(\sqrt{N})$ parameters are sufficient and necessary up to logarithmic terms (see Vardi et al. (2021)), motivating the question of what is the optimal number of parameters for *robust* memorization.

Finally, it would be interesting to study the extent to which trained neural networks (using standard optimization methods) can robustly memorize datasets, and in particular whether a network width logarithmic with N is still sufficient, as in our existence results.

Acknowledgments

This research is supported in part by European Research Council (ERC) grant 754705, and by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Alexander Bastounis, Anders C Hansen, and Verner Vlačić. The mathematics of adversarial attacks in ai–why deep learning is unstable despite the existence of stable neural networks. *arXiv preprint arXiv:2109.06098*, 2021.

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

Károly Böröczky and Gergely Wintsche. Covering the sphere by equal spherical balls. *Discrete and Computational Geometry: The Goodman-Pollack Festschrift*, pages 235–251, 2003.

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.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. Deep neural network approximation theory. *arXiv preprint arXiv:1901.02220*, 2019.

G. B. Folland. *A course in abstract harmonic analysis, Second edition*. CRC Press, 2015.

Peter Frankl and Hiroshi Maehara. Some geometric applications of the beta distribution. *Annals of the Institute of Statistical Mathematics*, 42:463–474, 1990.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Paul W Goldberg and Mark R Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2-3):131–148, 1995.

Boris Hanin and Mark Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.

Guang-Bin Huang. Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE transactions on neural networks*, 14(2):274–281, 2003.

Guang-Bin Huang and Haroon A Babri. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE transactions on neural networks*, 9(1):224–229, 1998.

Shih-Chi Huang, Yih-Fang Huang, et al. Bounds on the number of hidden neurons in multilayer perceptrons. *IEEE transactions on neural networks*, 2(1):47–55, 1991.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Konrad Knopp. *Theorie und Anwendung der unendlichen Reihen*. Springer, 1964. (English translation: Blackie, 1951 & Dover, reprint, 1990).

Binghui Li, Jikai Jin, Han Zhong, John Hopcroft, and Liwei Wang. Why robust generalization in deep learning is difficult: Perspective of expressive power. *Advances in Neural Information Processing Systems*, 35:4370–4384, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *Preprint, arXiv:1706.06083*, 2017.

Pertti Mattila. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*. Number 44. Cambridge university press, 1999.

Márton Naszódi. On some covering problems in geometry. *Proceedings of the American Mathematical Society*, 144(8):3555–3562, 2016.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

Sejun Park, Jaeho Lee, Chulhee Yun, and Jinwoo Shin. Provable memorization via deep neural networks using sub-linear parameters. In *Conference on Learning Theory*, pages 3627–3661. PMLR, 2021.

Jan Hendrik Rolfes and Frank Vallentin. Covering compact metric spaces greedily. *Acta Mathematica Hungarica*, 155:130–140, 2018.

Michael A Sartori and Panos J Antsaklis. A simple method to derive bounds on the size and to train multilayer neural networks. *IEEE transactions on neural networks*, 2(4):467–471, 1991.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *Preprint, arXiv:1312.6199*, 2013.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. On the optimal memorization power of relu neural networks. *arXiv preprint arXiv:2110.03187*, 2021.

Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems*, 34: 7054–7067, 2021.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Lijia Yu, Xiao-Shan Gao, and Lijun Zhang. Optimal robust memorization with relu neural networks. In *The Twelfth International Conference on Learning Representations*, 2024.

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 15558–15569, 2019.

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

Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). *Advances in neural information processing systems*, 35:36094–36107, 2022.

Appendix

Table of Contents

Appendix A. Additional Notations

For $1 \leq k \leq d \in \mathbb{N}$ we will denote by $\mathrm{Gr}_{d,k}$ the real Grassmannian manifold which is the set of all k-dimensional vector subspaces of $\mathbb{R}^d$. For any vector subspace W there exists a unique orthogonal projection matrix onto it, that is a matrix $P \in M_d(\mathbb{R})$ with $P = P^2 = P^\top$ and $\mathrm{Im}(P) = W$. We will denote this matrix by P_W. We can thus think of $\mathrm{Gr}_{d,k}$ as the set of such projections. More generally, we will denote $\mathrm{End}_{d,k} = \{M \in M_d(\mathbb{R}) \mid \mathrm{rk}(M) = k\}$ for the set of rank k matrices, and note that $\mathrm{Gr}_{d,k} = \{P \in \mathrm{End}_{d,k} \mid P = P^2 = P^\top\}$.

Denote by $\mathbb{S}^{d-1} = \left\{x \in \mathbb{R}^d \mid \|x\|_2 = 1\right\}$ the $(d-1)$-dimensional unit sphere in $\mathbb{R}^d$. One can equip the sphere with a metric structure using the geodesic metric $\mathrm{dist}_{\mathrm{arc}}$ given by the angle between two points, $\mathrm{dist}_{\mathrm{arc}}(a, b) = \arccos\langle a, b\rangle$. We denote by

$$B^{d-1}_{\mathrm{arc}}(x, \varphi) = \left\{x' \in \mathbb{S}^{d-1} \mid \mathrm{dist}_{\mathrm{arc}}(x', x) \leq \varphi\right\}$$

the metric ball in this metric space, sometimes called a geodesic ball, cap or spherical cap.

We will denote by ν_d the unique Haar probability measure of the orthogonal group $O(d)$ and by $\mu_{d-1}, \gamma_{d,k}$ the unique $O(d)$-invariant probability measures of $\mathbb{S}^{d-1}$ and $\mathrm{Gr}_{d,k}$ respectively. For the formal definitions see Subsection H.1.

Appendix B. Proofs of the Main Results

B.1. Robust Memorization With Large Width

Proof [Proof of Theorem 5] Denote $r = c_p^+(d)\sigma$ and let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset (separation under l_2 norm). We know that $r < \delta/2$ and so there exists some $0 < w$ such that $\delta = 2r + w$. Denote $\tilde{\epsilon} = \frac{w^2}{4d(w+2r)^2} = \frac{(\delta-2r)^2}{4d\delta^2}$ then $0 < \tilde{\epsilon} < 1/2$ and so from Lemma 41 we get that there exists a neural network $g_{\tilde{\epsilon},2} : \mathbb{R} \to \mathbb{R}$ with width 3 and depth $O\left(\log_2(\tilde{\epsilon}^{-1})\right) = O\left(\log_2\left(\frac{d\delta}{\delta-2r}\right)\right)$ such that $|g_{\tilde{\epsilon},2}(\alpha) - \alpha^2| \leq \epsilon$ for every $\alpha \in [0, 1]$. Using this network, from Lemma 38 we get that for every $(x_i, y_i) \in \mathcal{D}$ there exists a neural network $f_{x_i,w,2} : \mathbb{R}^d \to \mathbb{R}$ with width $W = d + 2 + \mathcal{W}(g_{\tilde{\epsilon},2}) = d + 5$, and depth $L = O\left(d\mathcal{L}(g_{\tilde{\epsilon},2})\right) = O\left(d\log_2\left(\frac{d\delta}{\delta-2r}\right)\right)$, such that for all $x \in \mathbb{R}^d$ we have $f_{x_i,w,2}(x) \leq y_i$ and

$$f_{x_i,w,2}(x) = \begin{cases} y_i & \|x - x_i\|_2 \leq r \\ 0 & r + w \leq \|x - x_i\|_2 \end{cases}.$$

Finally, because $\mathcal{D}$ is δ-separated (under the l_2 norm), from Theorem 37 there exists a neural network $F_{d,\delta,r,2} : \mathbb{R}^d \to \mathbb{R}$ with width $d + 6$ and depth $O\left(Nd\log_2\left(\frac{d\delta}{\delta-2r}\right)\right)$ that $(r, 2)$-robustly memorizes the dataset $\mathcal{D}$. Define $f = F_{d,\delta,r,2}$ and let $x \in B_p^d(x_i, \sigma)$. Then, by definition of $c_p^+(d)$ and Lemma 70, we have $x \in B_2^d(x_i, r)$ so $f(x) = F_{d,\delta,r,2}(x) = y_i$ and hence f indeed (σ, p)-robustly memorizes $\mathcal{D}$. Now $d + 6 \leq k$ and so by padding each hidden layer of f with $k - (d + 6)$ neurons we obtain f with width k and depth

$$O\left(Nd\log_2\left(\frac{d\delta}{\delta - 2r}\right)\right) = O\left(Nd\log_2\left(\frac{d\delta}{\delta - 2c_p^+(d)\sigma}\right)\right)$$
$$= O\left(Nd\log_2\left(d\left(1 - \frac{2c_p^+(d)\sigma}{\delta}\right)^{-1}\right)\right)$$

■

Proof [Proof of Theorem 6] We prove for $p = 1$ and $p = \infty$:

- Case $p = 1$: Follow the proof of Theorem 5 where instead of Lemma 41 use the identity map $g_1(\alpha) = \alpha$ (with width and depth of 1) to obtain from Lemma 38 indicators $f_{x_i,w,1}$. Because $\mathcal{D}$ is δ-separated (under the l_2 norm) it satisfies $\delta \leq \|x_i - x_j\|_1$ for every x_i, x_j with $y_i \neq y_j$ and so we can use Theorem 37 with the $f_{x_i,w,1}$'s we have in order to get a neural network $F_{d,\delta,\sigma,1} : \mathbb{R}^d \to \mathbb{R}$ with width $d + 4$ and depth $O(Nd)$ that $(\sigma, 1)$-robustly memorizes the dataset $\mathcal{D}$. Define $f = F_{d,\delta,\sigma,1}$ then $f(x) = F_{d,\delta,\sigma,1}(x) = y_i$. Now $d + 4 \leq k$ and so by padding each hidden layer of f with $k - (d + 4)$ neurons we obtain f with width k and depth $O(Nd)$.

- Case $p = \infty$: Denote $\tau = \delta / c_p^+(d)$ and let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset. From Lemma 39 we get that for every $(x_i, y_i) \in \mathcal{D}$ there exists a neural network $f_{x_i,w,\infty} : \mathbb{R}^d \to \mathbb{R}$ with width $W = d+3$, and depth $L = O(d)$, such that for all $x \in \mathbb{R}^d$ we have $f_{x_i,w,\infty}(x) \leq y_i$ and
$$f_{x_i,w,\infty}(x) = \begin{cases} y_i & \|x - x_i\|_\infty \leq \sigma \\ 0 & \sigma + w \leq \|x - x_i\|_\infty \end{cases},$$
for every $0 < w$. Finally, because $\mathcal{D}$ is δ-separated (under the l_2 norm) we get from the definition of τ and from Lemma 69 that $\tau \leq \|x_i - x_j\|_\infty$ for every x_i, x_j with $y_i \neq y_j$ and so from Theorem 37 there exists a neural network $F_{d,\tau,\sigma,\infty} : \mathbb{R}^d \to \mathbb{R}$ with width $d + 4$ and depth $O(Nd)$ that (σ, ∞)-robustly memorizes the dataset $\mathcal{D}$. Define $f = F_{d,\tau,\sigma,\infty}$ and let $x \in B_\infty^d(x_i, \sigma)$ then $f(x) = F_{d,\tau,\sigma,\infty}(x) = y_i$. Now $d + 4 \leq k$ and so by padding each hidden layer of f with $k - (d + 4)$ neurons we obtain f with width k and depth $O(Nd)$.

■

B.2. Robust Memorization With Small Width

B.2.1. Robust Memorization and Preservability

Definition 15 *Let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset, and let $\sigma < \frac{\delta}{2}$. We say that $\mathcal{D}$ is (σ, k)-**preservable** if there exists some $M \in End_{d,k}$ that (σ, k)-preserves $\mathcal{D}$. If, furthermore, $M = \frac{1}{\epsilon}P$ for some $P \in Gr_{d,k}$ we say that $\mathcal{D}$ is (σ, ϵ, k)-**orthogonally preservable**.*

Robust memorization with width smaller than the data dimension is possible only when the data is preservable as the next two theorems show (see Subsection B.3 for proofs):

Theorem 16 *Let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset such that $\mathcal{D}$ is (σ', k)-preservable under a map M with $\|M\|_2 < \frac{\delta}{2\sigma'}$, where $\sigma' = c_p^+(d)\sigma$. Then there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width $k+6$ and depth $O\left(Nk\log_2\left(\frac{k}{1-\|M\|_2\frac{2\sigma'}{\delta}}\right)\right)$ that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.*

Theorem 17 *Let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset such that $\mathcal{D}$ is not (σ', k)-preservable, where $\sigma' = c_p^-(d)\sigma$. Then every neural network f with width $\leq k$ cannot (σ, p)-robustly memorize the dataset $\mathcal{D}$.*

B.2.2. CHARACTERIZATION OF PRESERVABILITY

Theorems 16 and 17 highlight the connection between the preservability of a dataset in dimension k, and the ability to robustly memorize it with a network of width k. Therefore, we look for criteria to ensure (σ, k)-preservability for some general σ, δ and k.

Theorem 18 *If $\frac{2\sigma}{\delta} \leq \frac{1}{2}\sqrt{\frac{k}{de}}N^{-\frac{2}{k}}$, then every δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ is (σ, ϵ, k)-orthogonally preservable with $\epsilon = \frac{1}{2}\sqrt{\frac{k}{de}}N^{-\frac{2}{k}}$.*

The proof can be found in Subsection C.1

Theorem 19 *If $\frac{2\sigma}{\delta} > 4832N^{-\frac{2}{k}}$ then there exists a δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,2}(\delta)$ which is not (σ, k)-preservable.*

The proof can be found in Subsection C.2

B.2.3. PROOFS OF THEOREM 7 AND THEOREM 8

Using Theorems 18 and 19 we can now prove the main results:

Proof [Proof of Theorem 7] Let $\frac{\sigma}{\delta} \leq \frac{1}{8c_p^+(d)\sqrt{e}}\sqrt{\frac{k-6}{d}}N^{-\frac{2}{k-6}}$ and let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset. Denote $\sigma' := c_p^+(d)\sigma$, then from Theorem 18 we have that $\mathcal{D}$ is $(\sigma', \epsilon, k-6)$-orthogonally preservable with $\epsilon = \frac{1}{2}\sqrt{\frac{k-6}{de}}N^{-\frac{2}{k-6}}$. Note that $\frac{1}{\epsilon} < \frac{\delta}{2\sigma'}$ and so from Theorem 16 and Lemma 21 we conclude that there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth

$$O\left(N(k-6)\log_2\left(\frac{k-6}{1-\frac{2\sigma'}{\epsilon\delta}}\right)\right) = O\left(Nk\log_2\left(\frac{k}{1-\frac{2\sigma'}{\epsilon\delta}}\right)\right),$$

that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.

Now $\frac{2c_p^+(d)\sigma}{\delta} \leq \frac{1}{2}\epsilon$ so $\frac{2\sigma'}{\epsilon\delta} \leq \frac{1}{2}$ and hence the depth of f is $O\left(Nk\log_2(k)\right)$. The theorem follows by noting that $\frac{1}{8\sqrt{e}}d^{-\frac{1}{2}+\left[\frac{1}{p}-\frac{1}{2}\right]_-} \leq \frac{1}{8c_p^+(d)\sqrt{e}}\sqrt{\frac{k-6}{d}}$. ■

Proof [Proof of Theorem 8] Let $\frac{\sigma}{\delta} > \frac{2416}{c_p^-(d)} N^{-\frac{2}{k}}$ and denote $\sigma' := c_p^-(d)\sigma$, then from Theorem 19 we get that there exists a δ-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,2}(\delta)$ which is not (σ', k)-preservable. Finally, by Theorem 17 we conclude that there isn't a neural network f with width equal to k that (σ, p)-robustly memorizes the dataset $\mathcal{D}$. The theorem follows by noting that $2416 d^{\left[\frac{1}{p}-\frac{1}{2}\right]_+} = \frac{2416}{c_p^-(d)}$ ■

B.3. Proofs of Theorem 16 and Theorem 17

Proof [Proof of Theorem 16] Denote the elements in $\mathcal{D}$ by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \in \mathcal{D}_{d,N,C}(\delta)$. We know that $\mathcal{D}$ is (σ', k)-preservable under the map M. By definition of (σ', k)-preservability and Lemma 21, $\sigma' < \frac{\delta}{2}$ and there exists $P \in \text{Gr}_{d,k}$ such that for every $a_i \in B_2^d(x_i, \sigma'), a_j \in B_2^d(x_j, \sigma')$ with $y_i \neq y_j$ one has

$$\epsilon \|a_i - a_j\|_2 \leq \|P(a_i - a_j)\|_2$$

with $\epsilon = \frac{1}{\|M\|_2}$. We look at the data points of $\mathcal{D}$ projected by P. Denote $\mathcal{D}' = \{(x_i', y_i)\}_{i=1}^N$ where $x_i' = P(x_i)$ then $\mathcal{D}' \in \mathcal{D}_{k,N,C}$. For any $y_i \neq y_j$ and $a_i' \in B_2^k(x_i', \sigma')$, $a_j' \in B_2^k(x_j', \sigma')$ we get from Lemma 71 that there are $a_i \in B_2^d(x_i, \sigma'), a_j \in B_2^d(x_j, \sigma')$ such that $Pa_i = a_i', Pa_j = a_j'$. Hence

$$\|a_i' - a_j'\|_2 = \|P(a_i - a_j)\|_2 \geq \epsilon \|a_i - a_j\|_2 \geq \epsilon(\delta - 2\sigma') . \tag{4}$$

Denote $\tau = \min\left\{\|x_i' - x_j'\|_2 \mid y_i \neq y_j\right\}$ (note that $0 < \epsilon\delta \leq \tau$). Assume that $\tau/2 \leq \sigma'$ then there exists $y_i \neq y_j$ and $a_i' \in B_2^k(x_i', \sigma')$, $a_j' \in B_2^k(x_j', \sigma')$ such that $a_i' = a_j'$. From Eq. (4) we get that $\epsilon(\delta - 2\sigma') \leq 0$ so $\delta \leq 2\sigma'$ which is a contradiction, and so $\sigma' < \tau/2$.

This means that $\mathcal{D}' \in \mathcal{D}_{k,N,C}(\tau)$ and $\frac{\sigma'}{\tau} < \frac{1}{2} = \frac{1}{2c_2^+(k)}$. Denote $k' = k + 6$ then by Theorem 5 there exists a neural network $f' : \mathbb{R}^k \to \mathbb{R}$ with width k' and depth

$$O\left(Nk \log_2\left(\frac{k}{1 - \frac{2c_2^+(k)\sigma'}{\tau}}\right)\right) = O\left(Nk \log_2\left(\frac{k}{1 - \frac{2\sigma'}{\tau}}\right)\right)$$

that $(\sigma', 2)$-robustly memorizes the dataset $\mathcal{D}'$. Note that $0 < \epsilon\delta \leq \tau$ and that $\frac{1}{\epsilon} = \|M\|_2 < \frac{\delta}{2\sigma'}$ so $0 < \frac{2\sigma'}{\tau} \leq \frac{2\sigma'}{\epsilon\delta} < 1$ and the depth of f' is $O\left(Nk \log_2\left(\frac{k}{1 - \frac{2\sigma'}{\epsilon\delta}}\right)\right)$.

We define the function $f = f' \circ P$ (where we think of P now as a $k \times d$ matrix). Then $f' : \mathbb{R}^k \to \mathbb{R}$ and $f : \mathbb{R}^d \to \mathbb{R}$ have the same width and depth. Let us show that f indeed (σ, p)-robustly memorizes the dataset $\mathcal{D}$:

Let $i \in [N]$ and $x \in B_p^d(x_i, \sigma)$ then by Lemma 70 $x \in B_2^d\left(x_i, c_p^+(d)\sigma\right)$ and so because $\|Pv\|_2 \leq \|v\|_2$ for every v, we get

$$P(x) \in B_2^k\left(P(x_i), c_p^+(d)\sigma\right) = B_2^k\left(x_i', \sigma'\right) .$$

But f' as shown above $(\sigma', 2)$-robustly memorizes the dataset $\mathcal{D}'$ so $f'(Px) = y_i$ from which we conclude that $f(x) = y_i$, and so f indeed (σ, p)-robustly memorizes the dataset $\mathcal{D}$. Furthermore, the width of f is $k + 6$ and its depth is $O\left(Nk \log_2\left(\frac{k}{1 - \|M\|_2 \frac{2\sigma'}{\delta}}\right)\right)$. ■

Proof [Proof of Theorem 17] Let $f = T^{(L)} \circ [\cdot]_+ \circ T^{(L-1)} \circ ... \circ [\cdot]_+ \circ T^{(1)}$ be a neural network with width $\mathcal{W}(f) \leq k$ and architecture $\mathcal{A}(f) = (d_0, d_1, ..., d_L)$. By the definition of width $\mathcal{W}$, we have $d_1 \leq k$ and there exists some $W^{(1)} \in M_{d_1 \times d}(\mathbb{R})$ and $b^{(1)} \in \mathbb{R}^{d_1}$ such that $T^{(1)}x = W^{(1)}x + b^{(1)}$. Denote $M = \begin{bmatrix} W^{(1)} \\ 0 \end{bmatrix} \in M_d(\mathbb{R})$, then $M \in \text{End}_{d,d_1}$. Now, $\mathcal{D}$ is not (σ', k)-preservable and hence not (σ', d_1)-preservable, where $\sigma' = c_p^-(d)\sigma$. Hence, from Lemma 20 there exists some $(x_i, y_i), (x_j, y_j) \in \mathcal{D}$ with $y_i \neq y_j$ and some $a_i \in B_2^d(x_i, \sigma'), a_j \in B_2^d(x_j, \sigma')$ such that $M(a_i - a_j) = 0$. By Lemma 70 we have $a_i \in B_p^d(x_i, \sigma), a_j \in B_p^d(x_j, \sigma)$ with $M(a_i - a_j) = 0$. Hence $W^{(1)}a_i = W^{(1)}a_j$ so $f(a_i) = f(a_j)$ and we conclude that f cannot (σ, p)-robustly memorizes the dataset $\mathcal{D}$. ■

The following is a useful consequence of the lack of preservability.

Lemma 20 *If $\mathcal{D}$ is not (σ, k)-preservable, then for every $M \in \text{End}_{d,k}$ there exists some $a_i \in B_2^d(x_i, \sigma), a_j \in B_2^d(x_j, \sigma)$ with $y_i \neq y_j$ such that $M(a_i - a_j) = 0$.*

Proof [Proof of Lemma 20] Assume that there exists M such that for every $a_i \in B_2^d(x_i, \sigma), a_j \in B_2^d(x_j, \sigma)$ with $y_i \neq y_j$ we have $0 < \|M(a_i - a_j)\|_2$. Let i, j such that $y_i \neq y_j$. The Minkowski difference $B_{i,j} := B_2^d(x_i, \sigma) - B_2^d(x_j, \sigma)$ is compact and $\|M(\cdot)\|_2$ is continuous and positive on $B_{i,j}$ so it obtains a minimum $0 < t_{i,j}$. Since there are finitely many $t_{i,j}$ we can denote $0 < t = \min \{t_{i,j} \mid y_i \neq y_j\}$. Similarly $\|\cdot\|_2$ is continuous and positive on $B_{i,j}$ so it obtains a maximum $0 < \tau'_{i,j}$. Denote $0 < \tau' = \max \left\{\tau'_{i,j} \mid y_i \neq y_j\right\}$. Let $(x_i, y_i), (x_j, y_j) \in \mathcal{D}$ with $y_i \neq y_j$ and let $a_i \in B_2^d(x_i, \sigma), a_j \in B_2^d(x_j, \sigma)$. Then

$$\begin{aligned}\|M(a_i - a_j)\|_2 &\geq t \\ &= t\frac{\|a_i - a_j\|_2}{\|a_i - a_j\|_2} \\ &\geq \frac{t}{\tau'}\|a_i - a_j\|_2\end{aligned}$$

so if we define $M' = \frac{\tau'}{t}M$ we get $M' \in \text{End}_{d,k}$ and it (σ, k)-preserves $\mathcal{D}$ which is a contradiction. ■

Lemma 21 *$\mathcal{D}$ is (σ, ϵ, k)-orthogonally preservable if and only if it is (σ, k)-preservable under some M with $\|M\|_2 = \frac{1}{\epsilon}$.*

Proof [Proof of Lemma 21] If $\mathcal{D}$ is (σ, ϵ, k)-orthogonally preservable under P, define $M = \frac{1}{\epsilon}P$ then $M \in \text{End}_{d,k}$ and indeed $\mathcal{D}$ is (σ, k)-preservable under M with $\|M\|_2 = \frac{1}{\epsilon}$.
In the other direction, if $\mathcal{D}$ is (σ, k)-preservable under M, denote by $M = U\Sigma V^\top$ the singular value decomposition of M, where $U, V \in O(d)$, $\Sigma = \text{diag}(s_1, ..., s_k, 0, ..., 0) \in M_d(\mathbb{R})$ and $0 < s_k \leq ... \leq s_1$ are the singular values of M. Note that $\|\Sigma x\|_2 \leq s_1\|P_{W_0}x\|_2$ for any x where

$P_{W_0} = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}$. Thus, for any x

$$\begin{aligned} \|x\|_2 \le \|Mx\|_2 \Leftrightarrow & \quad \|x\|_2 \le \|U\Sigma V^\top x\|_2 \\ \Leftrightarrow & \quad \|x\|_2 \le \|\Sigma V^\top x\|_2 \\ \Rightarrow & \quad \|x\|_2 \le s_1 \|P_{W_0} V^\top x\|_2 \\ \Leftrightarrow & \quad \|x\|_2 \le s_1 \|P_{VW_0} x\|_2 \end{aligned}$$

where the last equivalence follows from Lemma 72 and the fact that $V^\top = V^{-1}$. Hence, if we define $W = VW_0$ then $P_W \in \mathrm{Gr}_{d,k}$ and for every $a_i \in B_2^d(x_i, \sigma), a_j \in B_2^d(x_j, \sigma)$ with $y_i \neq y_j$ one has

$$\frac{1}{s_1} \|a_i - a_j\|_2 \le \|P_W (a_i - a_j)\|_2 \, .$$

So $\mathcal{D}$ is (σ, ϵ, k)-orthogonally preservable with $\epsilon = \frac{1}{s_1} = \frac{1}{\|M\|_2}$ ■

Appendix C. Proof of Characterization of Preservability

C.1. Proof of Theorem 18

We begin by proving the following lemma which is a modification of (Mattila, 1999, Lemma 3.11):

Lemma 22 *For any $x \in \mathbb{S}^{d-1}$ and $0 < r < 1$ one has*

$$\gamma_{d,k}\left(\{P \in Gr_{d,k} \mid \|Px\|_2 \le r\}\right) = \mu_{d-1}\left(\left\{y \in \mathbb{S}^{d-1} \mid y_1^2 + ... + y_k^2 \le r^2\right\}\right) \, .$$

Proof [Proof of Lemma 22] Let $V_0 = \text{Span}\{e_{k+1}, ..., e_d\}$, we have that:

$$\begin{aligned} &\gamma_{d,k}\left(\{W \in \mathrm{Gr}_{d,k} \mid \|P_W x\|_2 \le r\}\right) \\ &= \gamma_{d,k}\left(\left\{W \in \mathrm{Gr}_{d,k} \mid \mathrm{dist}_2(x, W^\perp) \le r\right\}\right) && \text{(definition of orthogonal projection)} \\ &= \gamma_{d,d-k}\left(\left\{W^\perp \in \mathrm{Gr}_{d,d-k} \mid \mathrm{dist}_2(x, W^\perp) \le r\right\}\right) && \text{(Lemma 53)} \\ &= \gamma_{d,d-k}\left(\{V \in \mathrm{Gr}_{d,d-k} \mid \mathrm{dist}_2(x, V) \le r\}\right) \\ &= \nu_d\left(\{g \in O(d) \mid \mathrm{dist}_2(x, gV_0) \le r\}\right) && \text{(Lemma 52)} \\ &= \nu_d\left(\left\{g \in O(d) \mid \mathrm{dist}_2(g^{-1}x, V_0) \le r\right\}\right) && (O(d) \text{ preserves } l_2 \text{ norm}) \\ &= \mu_{d-1}\left(\left\{y \in \mathbb{S}^{d-1} \mid \mathrm{dist}_2(y, V_0) \le r\right\}\right) && \text{(Lemma 52)} \\ &= \mu_{d-1}\left(\left\{y \in \mathbb{S}^{d-1} \mid \|P_{V_0^\perp} y\|_2 \le r\right\}\right) && \text{(definition of orthogonal projection)} \\ &= \mu_{d-1}\left(\left\{y \in \mathbb{S}^{d-1} \mid y_1^2 + ... + y_k^2 \le r^2\right\}\right) && \text{(definition of } V_0) \end{aligned}$$

■

Lemma 23 *For any $x \in \mathbb{S}^{d-1}$ one has*

$$\left\{P \in Gr_{d,k} \mid \exists b \in B_{arc}^{d-1}(x,\varphi) \ s.t \ \|Pb\|_2 \le \epsilon\right\} \subseteq \left\{P \in Gr_{d,k} \mid \|Px\|_2 \le \epsilon + \sin\varphi\right\}.$$

Proof [Proof of Lemma 23] Let P be in the set on the left-hand side, and denote $W = Im(P) \cap \mathbb{S}^{d-1}$, then there exists $b \in B_{\text{arc}}^{d-1}(x,\varphi)$ such that $\|Pb\|_2 \le \epsilon$. Now, $\text{dist}_{\text{arc}}(b, W)$ (the angle between b, Pb) satisfies

$$\cos(\text{dist}_{\text{arc}}(b,W)) = \frac{\langle b, Pb\rangle}{\|b\|_2\|Pb\|_2} = \frac{\|Pb\|_2^2}{1 \cdot \|Pb\|_2} = \|Pb\|_2 \le \epsilon.$$

Similarly $\|P(x)\|_2 = \cos(\text{dist}_{\text{arc}}(x,W))$. From the triangle inequality we have $\text{dist}_{\text{arc}}(b,W) \le \text{dist}_{\text{arc}}(b,x) + \text{dist}_{\text{arc}}(x,W) \le \varphi + \text{dist}_{\text{arc}}(x,W)$ and so

$$\begin{aligned}
\|P(x)\|_2 &= \cos(\text{dist}_{\text{arc}}(x,W)) \\
&\le \cos(\text{dist}_{\text{arc}}(b,W) - \varphi) \\
&= \cos(\text{dist}_{\text{arc}}(b,W))\cos(\varphi) + \sin(\text{dist}_{\text{arc}}(b,W))\sin(\varphi) \\
&\le \cos(\text{dist}_{\text{arc}}(b,W)) + \sin(\varphi) \\
&\le \epsilon + \sin(\varphi) \ .
\end{aligned}$$

■

Lemma 24 *For any $0 < r < 1$ one has*

$$\mu_{d-1}\left(\left\{y \in \mathbb{S}^{d-1} \mid y_1^2 + ... + y_k^2 \le \frac{k}{d}r^2\right\}\right) < e^{\frac{k}{2}} r^k .$$

Proof [Proof of Lemma 24] From (Dasgupta and Gupta, 2003, Lemma 2.2), we get that

$$\mu_{d-1}\left(\left\{y \in \mathbb{S}^{d-1} \mid y_1^2 + ... + y_k^2 \le \frac{k}{d}r^2\right\}\right) \le r^k \cdot \left(1 + \frac{k(1-r^2)}{d-k}\right)^{(d-k)/2} .$$

Now, $x \mapsto \left(1 + \frac{k(1-x^2)}{d-k}\right)^{(d-k)/2}$ is decreasing in $[0,1]$ so for any $k < d$

$$\begin{aligned}
\left(1 + \frac{k(1-r^2)}{d-k}\right)^{(d-k)/2} &\le \left(1 + \frac{k}{d-k}\right)^{(d-k)/2} \\
&= \left(1 + \frac{k/2}{(d-k)/2}\right)^{(d-k)/2} \\
&\le e^{\frac{k}{2}} ,
\end{aligned}$$

which finishes the proof.

■

Lemma 25 *Let* $\sin\varphi < \frac{1}{2}\sqrt{\frac{k}{de}}m^{-\frac{1}{k}}$*, and denote* $\epsilon = \frac{1}{2\sqrt{e}}m^{-\frac{1}{k}}$*, then for any* $x \in \mathbb{S}^{d-1}$

$$\gamma_{d,k}\left(\left\{P \in Gr_{d,k} \mid \exists b \in B_{arc}^{d-1}(x,\varphi) \ s.t \left\|\sqrt{\frac{d}{k}}Px\right\|_2 \leq \epsilon\right\}\right) < m^{-1}$$

Proof [Proof of Lemma 25] We have

$$\begin{aligned}
&\gamma_{d,k}\left(\left\{P \in \mathrm{Gr}_{d,k} \mid \exists b \in B_{\mathrm{arc}}^{d-1}(x,\varphi) \ \text{s.t} \left\|\sqrt{\frac{d}{k}}Px\right\|_2 \leq \epsilon\right\}\right) \\
&\leq \gamma_{d,k}\left(\left\{P \in \mathrm{Gr}_{d,k} \mid \|Px\|_2 \leq \sqrt{\frac{k}{d}}\epsilon + \sin\varphi\right\}\right) && \text{(Lemma 23)} \\
&= \gamma_{d,k}\left(\left\{P \in \mathrm{Gr}_{d,k} \mid \|Px\|_2 \leq \sqrt{\frac{k}{d}}\left(\epsilon + \sqrt{\frac{d}{k}}\sin\varphi\right)\right\}\right) \\
&= \mu_{d-1}\left(\left\{y \in \mathbb{S}^{d-1} \mid y_1^2 + ... + y_k^2 \leq \frac{k}{d}\left(\epsilon + \sqrt{\frac{d}{k}}\sin\varphi\right)^2\right\}\right) && \text{(Lemma 22)} \\
&\leq e^{\frac{k}{2}}\left(\epsilon + \sqrt{\frac{d}{k}}\sin\varphi\right)^k && \text{(Lemma 24)} \\
&= e^{\frac{k}{2}}\left(\frac{1}{2\sqrt{e}}m^{-\frac{1}{k}} + \sqrt{\frac{d}{k}}\sin\varphi\right)^k \\
&< e^{\frac{k}{2}}\left(\frac{1}{2\sqrt{e}}m^{-\frac{1}{k}} + \sqrt{\frac{d}{k}}\frac{1}{2}\sqrt{\frac{k}{de}}m^{-\frac{1}{k}}\right)^k && \text{(choice of } \varphi\text{)} \\
&= e^{\frac{k}{2}}\left(\frac{1}{\sqrt{e}}m^{-\frac{1}{k}}\right)^k \\
&= m^{-1}.
\end{aligned}$$

Where we can use Lemma 24 because $\epsilon + \sqrt{\frac{d}{k}}\sin\varphi \leq \frac{1}{\sqrt{e}}m^{-\frac{1}{k}} \leq \frac{1}{\sqrt{e}}2^{-\frac{1}{k}} < 1$. ■

Lemma 26 *Let* $\mathcal{V} \subset \mathbb{S}^{d-1}$ *with* $|\mathcal{V}| = m$*. Denote* $\epsilon = \frac{1}{2}\sqrt{\frac{k}{de}}m^{-\frac{1}{k}}$ *and let* $\sin\varphi < \epsilon$*, then there exists some* $P \in Gr_{d,k}$ *such that for every* $u \in \bigcup_{v\in\mathcal{V}} B_{arc}^{d-1}(v,\varphi)$ *one has* $\epsilon \leq \|Pu\|_2$.

Proof [Proof of Lemma 26] Denote $A_v = \left\{P \in \mathrm{Gr}_{d,k} \mid \exists b \in B_{\mathrm{arc}}^{d-1}(v,\varphi) \ \text{s.t} \ \|Pb\|_2 \leq \epsilon\right\}$, then

$$\gamma_{d,k}\left(\bigcup_{v\in\mathcal{V}} A_v\right) \leq \sum_{v\in\mathcal{V}}\gamma_{d,k}(A_v) < \sum_{v\in\mathcal{V}}\frac{1}{m} = m\frac{1}{m} = 1 = \gamma_{d,k}(\mathrm{Gr}_{d,k}) \ ,$$

where the second inequality follows from Lemma 25. Therefore there exists some $P \in \mathrm{Gr}_{d,k}$ such that for every $u \in \bigcup_{v\in\mathcal{V}} B_{\mathrm{arc}}^{d-1}(v,\varphi)$ one has $\epsilon \leq \|Pu\|_2$. ■

Proof [Proof of Theorem 18] Let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta)$ be a δ-separated dataset. Define

$$\mathcal{V} = \left\{ \frac{x_i - x_j}{\|x_i - x_j\|_2} \mid y_i \neq y_j \right\}$$

then $\mathcal{V} \subset \mathbb{S}^{d-1}$ with $|\mathcal{V}| = m$ for some $m \leq N^2$. Denote $\varphi = \sin^{-1}\left(\frac{2\sigma}{\delta}\right)$ then $\sin\varphi < \frac{1}{2}\sqrt{\frac{k}{de}}N^{-\frac{2}{k}} \leq \frac{1}{2}\sqrt{\frac{k}{de}}m^{-\frac{1}{k}}$. Hence, by Lemma 26 there exists some $P \in \mathrm{Gr}_{d,k}$ such that for every $u \in \bigcup_{v\in\mathcal{V}} B_{\text{arc}}^{d-1}(v,\varphi)$ one has $\frac{1}{2}\sqrt{\frac{k}{de}}m^{-\frac{1}{k}} \leq \|Pu\|_2$ so $\frac{1}{2}\sqrt{\frac{k}{de}}N^{-\frac{2}{k}} \leq \|Pu\|_2$. By the definition of $\mathcal{V}, \delta, \varphi$ and B_{arc}^{d-1} we conclude that for every $a_i \in B_2^d(x_i,\sigma), a_j \in B_2^d(x_j,\sigma)$ with $y_i \neq y_j$ one has

$$\begin{aligned}\frac{a_i - a_j}{\|a_i - a_j\|_2} &\in B_{\text{arc}}^{d-1}\left(\frac{x_i - x_j}{\|x_i - x_j\|_2}, \sin^{-1}\left(\frac{2\sigma}{\|x_i - x_j\|_2}\right)\right) \\ &\subseteq B_{\text{arc}}^{d-1}\left(\frac{x_i - x_j}{\|x_i - x_j\|_2}, \sin^{-1}\left(\frac{2\sigma}{\delta}\right)\right) ,\end{aligned}$$

and so

$$\epsilon\|a_i - a_j\|_2 \leq \|P(a_i - a_j)\|_2 ,$$

which completes the proof. ■

C.2. Proof of Theorem 19

In order to prove Theorem 19 we show that for a big enough geodesic radius φ we can always cover $\mathbb{S}^k$ with geodesic balls $B_{\text{arc}}^k(\varphi)$. We do so using the covering number τ (defined in Definition 59) and covering density ϑ (defined in Definition 60). This cover will enable us to construct a non preservable dataset.

We begin by proving the following lemma

Lemma 27 *Let $r \leq R$, $k \in \mathbb{N}$ and denote $\Delta_k = \sqrt{2}\left(5k\ln(k+1)\sqrt{2\pi(k+1)}\right)^{1/k}$. We have that if $\left(\frac{\Delta_k R}{r}\right)^k < m$, then there exists a set $\mathcal{V} = \{v_1, ..., v_m\} \subseteq \partial B_2^{k+1}(0,R)$ such that*

$$\partial B_2^{k+1}(0,R) \subseteq \bigcup_{i=1}^{m} B_2^{k+1}(v_i,r) .$$

Proof [Proof of Lemma 27] Let $2\sin(\varphi/2) = r/R$. We have

$$\begin{aligned}
&\tau\left(\partial B_2^{k+1}(0,R), \mathcal{G}_{O(k+1)} B_2^{k+1}(Re_1, r)\right) \\
&= \tau\left(\mathbb{S}^k, \mathcal{G}_{O(k+1)} B_2^{k+1}(e_1, r/R)\right) \\
&= \tau\left(\mathbb{S}^k, \mathcal{G}_{O(k+1)} B_{\text{arc}}^k(\varphi)\right) \\
&= \frac{\vartheta\left(\mathbb{S}^k, \mathcal{G}_{O(k+1)} B_{\text{arc}}^k(\varphi)\right)}{\mu_k\left(B_{\text{arc}}^k(\varphi)\right)} \\
&\leq \frac{5k\ln(k+1)}{\mu_k\left(B_{\text{arc}}^k(\varphi)\right)} && \text{(Lemma 62)} \\
&\leq \frac{5k\ln(k+1)\sqrt{2\pi(k+1)}}{\sin^k\varphi} && \text{(Lemma 63)} \\
&= \frac{2^{\frac{k}{2}} 5k\ln(k+1)\sqrt{2\pi(k+1)}}{(\sqrt{2}\sin\varphi)^k} \\
&= \frac{\Delta_k^k}{(\sqrt{2}\sin\varphi)^k} \\
&\leq \frac{\Delta_k^k}{(2\sin(\varphi/2))^k} && (2\sin(\varphi/2) \leq \sqrt{2}\sin\varphi) \\
&= \frac{\Delta_k^k}{(r/R)^k} \\
&< m\ .
\end{aligned}$$

We conclude that there exists a set $\mathcal{V}$ with $|\mathcal{V}| = m$ which satisfies the requirements of the lemma. ■

Lemma 28 *Let* $0 \leq \sigma < \frac{\delta}{2}$ *then*

$$\frac{1}{2}\left(\sigma + \sqrt{\sigma^2 + 4\sigma\delta}\right) \leq \sqrt{2\sigma\delta}$$

Proof [Proof of Lemma 28] Assume $0 < \sigma$ (otherwise the claim is trivial). Note that

$$\begin{aligned}
\frac{1}{2}\left(\sigma + \sqrt{\sigma^2 + 4\sigma\delta}\right) \leq \sqrt{2\sigma\delta} &\iff 2\sigma^2 + 4\sigma\delta + 2\sigma\sqrt{\sigma^2 + 4\sigma\delta} \leq 8\sigma\delta \\
&\iff \sigma^2 + \sigma\sqrt{\sigma^2 + 4\sigma\delta} \leq 2\sigma\delta \\
&\iff \sigma + \sqrt{\sigma^2 + 4\sigma\delta} \leq 2\delta\ ,
\end{aligned}$$

and indeed

$$\sigma + \sqrt{\sigma^2 + 4\sigma\delta} < \left(\frac{\delta}{2}\right) + \sqrt{\left(\frac{\delta}{2}\right)^2 + 4\left(\frac{\delta}{2}\right)\delta} = 2\delta\ ,$$

so we are done. ■

We are now ready to prove Theorem 19:

Proof [Proof of Theorem 19] Denote $r = \sqrt{2\sigma\delta}$ then from Lemma 28 we have that

$$r \geq \frac{1}{2}\left(\sigma + \sqrt{\sigma^2 + 4\sigma\delta}\right) \ ,$$

and so $0 \leq r^2 - \sigma r - \sigma\delta$, from which we conclude that

$$\frac{r+\delta}{r} \leq \frac{r}{\sigma} \ . \tag{5}$$

We also have

$$\left(\frac{\sigma}{r}\right)^2 = \left(\frac{\sigma}{\sqrt{2\sigma\delta}}\right)^2 = \frac{\sigma}{2\delta} \ . \tag{6}$$

Now,

$$\begin{aligned}
\frac{2\sigma}{\delta} > 4832 N^{-\frac{2}{k}} &\implies \frac{2\sigma}{\delta} > 16\left(\sqrt{2}\cdot 10\sqrt{\pi}\ln 2\right)^2 N^{-\frac{2}{k}} \\
&\implies \frac{2\sigma}{\delta} > 16\Delta_k^2 N^{-\frac{2}{k}} && \text{(Definition of } \Delta_k \text{ and Lemma 75)} \\
&\implies \frac{\sigma}{2\delta} > 4\Delta_k^2 N^{-\frac{2}{k}} \\
&\implies \left(\frac{\sigma}{r}\right)^2 > 4\Delta_k^2 N^{-\frac{2}{k}} && \text{(Eq. (6))} \\
&\implies \frac{r}{\sigma} < \frac{1}{2\Delta_k} N^{\frac{1}{k}} \\
&\implies \frac{r}{\sigma} < \frac{1}{\Delta_k}\left(\frac{N}{2}\right)^{\frac{1}{k}} && \left(\frac{1}{2} \leq \frac{1}{2^{\frac{1}{k}}} \text{ for all } 1 \leq k\right) \\
&\implies \left(\frac{\Delta_k r}{\sigma}\right)^k < \frac{N}{2} \ ,
\end{aligned}$$

so by Lemma 27 there exists some $a_1, ..., a_{N/2}$ points on the sphere $\partial B_2^{k+1}(0, r)$ such that

$$\partial B_2^{k+1}(0, r) \subseteq \bigcup_{i=1}^{N/2} B_2^{k+1}\left(a_i, \sigma\right) \ .$$

By Lemma 73 we have $\partial B_2^{k+1}(0, r) \subseteq \bigcup\limits_{i=1}^{N/2} B_2^{k+1}\left(-a_i, \sigma\right)$ hence for any x' we have

$$\partial B_2^{k+1}(x', r) \subseteq \bigcup_{i=1}^{N/2} B_2^{k+1}\left(x' - a_i, \sigma\right) \ .$$

Now, by Eq. (5) we have $\frac{r+\delta}{r} \leq \frac{r}{\sigma}$ and so $\left(\frac{\Delta_k(r+\delta)}{r}\right)^k \leq \left(\frac{\Delta_k r}{\sigma}\right)^k < \frac{N}{2}$. Therefore, by Lemma 27 there exists some $b_1, ..., b_{N/2}$ points on the sphere $\partial B_2^{k+1}(0, r+\delta)$ such that

$$\partial B_2^{k+1}(0, r+\delta) \subseteq \bigcup_{i=1}^{N/2} B_2^{k+1}\left(b_i, r\right) \ .$$

We now define the dataset $\mathcal{D} = \{(a_1,1),...,(a_{N/2},1),(b_1,2),...,(b_{N/2},2)\}$. Note that by definition $\delta \le \|a_i - b_j\|_2$ and so $\mathcal{D} \in \mathcal{D}_{d,N,2}(\delta)$.

We now show that $\mathcal{D}$ is not (σ,k)-preservable. Let $M \in \text{End}_{d,k}$ and denote $K = \ker M$, then $K \in \text{Gr}_{d,d-k}$. Now $\mathbb{R}\partial B_2^{k+1}(0,r+\delta) \in \text{Gr}_{d,k+1}$ and since $d-k \ge d-(k+1)+1$ we get from Lemma 74 that there exists some $0 \ne x \in K \cap \mathbb{R}\partial B_2^{k+1}(0,r+\delta)$. Since $0 \ne x \in \mathbb{R}\partial B_2^{k+1}(0,r+\delta)$ we can write $x = t_1 u$ for some $u \in \partial B_2^{k+1}(0,r+\delta)$ and $t_1 \ne 0$.

Now, $u \in \partial B_2^{k+1}(0,r+\delta) \subseteq \bigcup_{i=1}^{N/2} B_2^{k+1}(b_i,r)$ hence there exists some b_j such that $u \in B_2^{k+1}(b_j,r)$. Therefore, there exists some $v \in \partial B_2^{k+1}(b_j,r)$ such that $u = t_2 v$ for some $t_2 \ne 0$. Now $Mv = Mt_2^{-1}u = Mt_2^{-1}t_1^{-1}x = t_2^{-1}t_1^{-1}Mx = 0$ so $v \in K$. But $\partial B_2^{k+1}(b_j,r) \subseteq \bigcup_{i=1}^{N/2} B_2^{k+1}(b_j - a_i,\sigma)$ and so there exists some a_i such that $v \in B_2^{k+1}(b_j - a_i,\sigma)$. Equivalently, there exists some $\alpha_i \in B_2^{k+1}(a_i,\sigma)$ such that $v = b_j - \alpha_i$. But $v \in K$ so $Mv = 0$. Treating $\mathbb{R}^{k+1}$ as a subspace of $\mathbb{R}^d$, we conclude that there exists some $\alpha_i \in B_2^d(a_i,\sigma)$ and some $\beta_j = b_j \in B_2^d(b_j,\sigma)$ such that

$$\|M(\beta_j - \alpha_i)\|_2 = 0 < \delta - \sigma \le \|\beta_j - \alpha_i\|_2$$

and so by definition $\mathcal{D}$ is not (σ,k)-preservable. ■

Appendix D. Tighter Bounds

For a δ-separated dataset $\mathcal{D}$ denote the set $X_{\mathcal{D}} = \left\{\frac{x-x'}{\|x-x'\|_2} \mid (x,y),(x',y') \in \mathcal{D} \text{ with } y \ne y'\right\}$. From Definition 12, it follows that the problem introduced in Section 5 is equivalent to the following problem: Under what conditions, for any δ-separated dataset $\mathcal{D}$ there exists a rank k map $M \in \text{End}_{d,k}$ such that

$$\forall v \in \bigcup_{v \in X_{\mathcal{D}}} B_{\text{arc}}^{d-1}(v,r_v) \quad \text{we have } 1 \le \|Mv\|_2 \, .$$

Here the radius r_v for $v = \frac{x-x'}{\|x-x'\|_2}$ is given by $r_v = \sin^{-1}\left(\frac{2\sigma}{\|x-x'\|_2}\right) \le \sin^{-1}\left(\frac{2\sigma}{\delta}\right)$. Note that since the set $\bigcup_{v \in X_{\mathcal{D}}} B_{\text{arc}}^{d-1}(v,r_v)$ is compact, there exists such M if and only if $Mv \ne 0$ on this set (see e.g Lemma 20). We conclude that the problem introduced in Section 5 can be solved given a solution to the following more general problem on the sphere: Under what conditions, for any fixed set $X \subset \mathbb{S}^{d-1}$ of size $|X| = m$ there exists a rank k map $M \in \text{End}_{d,k}$ such that

$$\forall x \in X^{(\varphi)}, Mx \ne 0. \quad \text{where} \quad X^{(\varphi)} := \bigcup_{x \in X} B_{\text{arc}}^{d-1}(x,\varphi) \, .$$

Or, equivalent, that there exists some subspace $U \in \text{Gr}_{d,d-k}$ (the kernel of M) such that $X^{(\varphi)} \cap U = \emptyset$. We define the sets that share this property:

Definition 29 *Let $X \subset \mathbb{S}^{d-1}$. We say that X is $(d-k,\varphi)$-**hitting** if for every $U \in Gr_{d,d-k}$ one has $X^{(\varphi)} \cap U \ne \emptyset$. We will denote by $\mathcal{X}_{\varphi,d-k}$ the set of all $(d-k,\varphi)$-hitting sets, and define the $(d-k,\varphi)$-**hitting number** to be*

$$m_{\varphi,d-k} = \min\{|X| \mid X \in \mathcal{X}_{\varphi,d-k}\} \, .$$

If $m < m_{\varphi,d-k}$ then any $X \subset \mathbb{S}^{d-1}$ with $|X| = m$ will satisfy $X^{(\varphi)} \cap U = \emptyset$ for some $U \in \mathrm{Gr}_{d,d-k}$. On the other hand, if $m_{\varphi,d-k} \leq m$ then there exists some $X \subset \mathbb{S}^{d-1}$ with $|X| = m$ such that for every $U \in \mathrm{Gr}_{d,d-k}$ one has $X^{(\varphi)} \cap U \neq \emptyset$. We conclude that our problem reduces to finding the value of $m_{\varphi,d-k}$. Namely, given φ, d, k, what is the minimal number of spherical caps on $\mathbb{S}^{d-1}$ of radius φ that are required in order to intersect every $d-k$-dimensional subspace non-trivially?

Using covering arguments it follows from Subsection C.2 that this number has an upper bound

$$m_{\varphi,d-k} \leq \left(\frac{C_2}{\varphi}\right)^k .$$

For the lower bound we inspect the contribution made by each point $x \in X$ separately. Namely, for each $x \in X$ we define the set $A_{d,d-k}(x,\varphi) = \left\{U \in \mathrm{Gr}_{d,d-k} \mid U \cap x^{(\varphi)} \neq \emptyset\right\}$. It follows from the definition that the covering number of $\mathrm{Gr}_{d,d-k}$ by translated copies of $A_{d,d-k}(x,\varphi)$ is exactly $m_{\varphi,d-k}$ (translation is done with respect to the transitive action of $O(d)$. See Definition 59 in Appendix H.3). From the proof in Subsection C.1 we get that the measure of $A_{d,d-k}(x,\varphi)$ is given by

$$\gamma_{d,d-k}(A_{d,d-k}(x,\varphi)) = \mu_{d-1}\left(y \in \mathbb{S}^{d-1} \mid y_1^2 + \ldots + y_k^2 \leq \sin^2(\varphi)\right) .$$

But $Z = y_1^2 + \ldots + y_k^2 \sim \mathrm{Beta}\left(\frac{k}{2}, \frac{d-k}{2}\right)$ (see (Frankl and Maehara, 1990, Corollary 1.1)) so the measure is given by the CDF:

$$\gamma_{d,d-k}(A_{d,d-k}(x,\varphi)) = \mathrm{CDF}_Z(\sin^2\varphi) = I_{\sin^2\varphi}\left(\frac{k}{2}, \frac{d-k}{2}\right)$$

where $I_r(a,b)$ is the regularized incomplete beta function. Note that the inverse of the measure, $\gamma_{d,d-k}(A_{d,d-k}(x,\varphi))^{-1}$, decreases to 1 as a function of d at a rate bounded by $(d/k)^{-\frac{k}{2}}$. To bound $m_{\varphi,d-k}$ from bellow we showed using the union bound that for small enough m, no set X of size m will result in a cover of $\mathrm{Gr}_{d,d-k}$. The above discussion shows that this small enough m decreases at a rate proportional to $(d/k)^{-\frac{k}{2}}$, and so the bound we get is

$$\left(\frac{k}{d}\right)^{k/2}\left(\frac{C_1}{\varphi}\right)^k \leq m_{\varphi,d-k} .$$

This probabilistic approach will always provide a lower bound that depends on d since the measure of each $A_{d,d-k}(x,\varphi)$ depends on d. Still, it is not clear whether this dependence on d is unavoidable. One could argue that since the problem involves only subspaces of co-dimension k, the number of spherical caps required to hit them should not depend on the ambient dimension d. Granted, the spherical caps on $\mathbb{S}^{d-1}$ depend on d, but as we saw in the upper bound, there are configurations of hitting caps that do not take advantage of the ambient dimension d. The question is whether such configurations are necessary, i.e is it true that caps of higher dimension cannot provide a more efficient configuration of a hitting set. It remain a subject for future work to investigate tight bounds for this problem.

Appendix E. Separation in l_q Norm

In Section 4 we obtained results that connect robust memorization and the width of the memorizing network, for any δ-separated datasets. Our definition of separation used the l_2 norm. In this section we aim to extend these results and consider datasets where separation is measured in an arbitrary l_q norm. All the proofs for this section appear in Section F.

In the following, we let $N, d, C \in \mathbb{N}_{\geq 2}$, $k \in \mathbb{N}$, $0 < \delta, \sigma, p \in (0, \infty]$, $q \in [1, \infty]$. Note that as before, p can be smaller than 1. This is because it is used to define the geometric shape of the robust neighborhood of the data points, and for this purpose the properties of quasi-norm suffice. On the other hand, q is used to define a norm that measures the separation distance between data points, and so it has to remain in the range $[1, \infty]$.

We will denote $c^+_{p,q}(d) = d^{\left[\frac{1}{q}-\frac{1}{p}\right]_+}$ and $c^-_{p,q}(d) = d^{\left[\frac{1}{q}-\frac{1}{p}\right]_-}$. As before, it follows from Lemma 70 that $c^+_{p,q}(d)$ is the radius of the l_q ball that encloses the unit l_p ball, and $c^-_{p,q}(d)$ is the radius of the l_q ball that is inscribed in the unit l_p ball.

Definition 30 *We say that a dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}$ is a (δ, q)-**separated** dataset, if we have $\delta = \min\{\|x_i - x_j\|_q \mid y_i \neq y_j\}$, and denote by $\mathcal{D}_{d,N,C}(\delta, q)$ the set of all such datasets.*

As in Section 4, given some $0 < \delta$ and $q \in [1, \infty]$ we wish to find the maximal possible value of σ that allows for (σ, p)-robust memorization, of any (δ, q)-separated dataset, using a width k network. When $q = 2$ we saw that the applicable range of σ was:

$$0 \leq \frac{\sigma}{\delta} < \frac{1}{2c^+_{p,2}(d)} .$$

For general $q \in [1, \infty]$, using the same reasoning we get that (σ, p)-robust memorization of every (δ, q)-separated dataset can only be considered in the range

$$0 \leq \frac{\sigma}{\delta} < \frac{1}{2c^+_{p,q}(d)} .$$

E.1. Robust Memorization With Large Width

For brevity we will use the notation $\lambda = \left(1 - \frac{2c^+_{p,q}(d)\sigma}{\delta}\right)^{-1}$ in the following theorems. In the case that the desired width k is bigger than the dimension of the data d we have:

Theorem 31 *Let $p \in (0, \infty]$, $q \in (1, \infty) \setminus \mathbb{N}$. If $d + 12 \leq k$ and*

$$\frac{\sigma}{\delta} < \frac{1}{2c^+_{p,q}(d)}$$

then, for every (δ, q)-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta, q)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth

$$O\left(Nd^{1+\frac{1}{q}}\lambda q\left(\log_2(dq\lambda^q) + q\right)\right)$$

that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.

In the case that $q \in \mathbb{N}$ an improved result can be obtained:

Theorem 32 *Let $p \in (0, \infty]$, $q \in \mathbb{N}_{\geq 2}$. If $d + 9 \leq k$ and*

$$\frac{\sigma}{\delta} < \frac{1}{2c^{+}_{p,q}(d)}$$

then, for every (δ, q)-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta, q)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth

$$O\left(Ndq \log_2\left(dq\lambda^q\right)\right)$$

that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.

Note that for $q = 2$ we recover Theorem 5. In the special case that $q \in \{1, \infty\}$ the range of the width can be improved, and the log term in the depth can be removed:

Theorem 33 *Let $p \in (0, \infty]$, $q \in \{1, \infty\}$. If $d + 4 \leq k$ and*

$$\frac{\sigma}{\delta} < \frac{1}{2c^{+}_{p,q}(d)}$$

then, for every (δ, q)-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta, q)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth $O(Nd)$ that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.

Note that if we allow the range of the width k to be $3d + 1 \leq k$ then in the case that $p = q \in \{1, \infty\}$, adjusting the construction in the proof of Theorem 33 according to Remark 40, the depth can be improved to be $O(N)$, thus recovering the result in (Yu et al., 2024, Theorem 4.8).

In the special case that $p = q \in \mathbb{N}_{\geq 2}$, (Yu et al., 2024, Theorem B.6) obtained the following result:

> Let $p = q \in \mathbb{N}_{\geq 2}$. If $k > cd$ for some universal constant c and
>
> $$\frac{\sigma}{\delta} < \frac{1}{2}$$
>
> then, for every (δ, q)-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta, q)$, such that $\mathcal{D} \subseteq [-\Delta, \Delta]^d$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth
>
> $$O\left(Nq \log_2\left(qd\left(\frac{\Delta}{\delta/2 - \sigma}\right)^q\right)\right) \tag{7}$$
>
> that (σ, q)-robustly memorizes the dataset $\mathcal{D}$.

By changing the range of k from $d + 9 \leq k$ to $8d + 1 \leq k$, and adjusting our construction for Theorem 32 according to Remark 40, we would get the following result:

> Let $p = q \in \mathbb{N}_{\geq 2}$. If $k \geq 8d + 1$ and
>
> $$\frac{\sigma}{\delta} < \frac{1}{2}$$
>
> then, for every (δ, q)-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta, q)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth
>
> $$O\left(Nq \log_2\left(qd\left(\frac{\delta/2}{\delta/2 - \sigma}\right)^q\right)\right) \tag{8}$$
>
> that (σ, q)-robustly memorizes the dataset $\mathcal{D}$.

The depth in Eq. (7) depends on the global spread of the data through the quantity Δ, whereas the depth in Eq. (8) is favorable since it is only affected by the local structure through the relation between δ and σ. In particular, for any two datasets with domains Δ_1, Δ_2, that share values for δ, σ we will obtain memorizing networks of the same complexity regardless of the size of the domains Δ_1 and Δ_2 of the datasets.

We conclude by remarking that in all of these cases, if one does not care about the complexity of the depth, a network with near optimal width always exists:

Proposition 34 *Let $p \in (0, \infty]$, $q \in [1, \infty]$. If $d + 1 \leq k$ and*

$$\frac{\sigma}{\delta} < \frac{1}{2c^{+}_{p,q}(d)}$$

then, for every (δ, q)-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta, q)$, such that $\mathcal{D} \subseteq B^d_2(0, \Delta)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth

$$O\left(\frac{C\Delta}{\delta - 2c^{+}_{p,q}(d)\sigma}\right)^{d+1}$$

that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.

A proof sketch for Proposition 34 appears in G.

E.2. Robust Memorization With Small Width

In the case where the desired width is smaller than the data dimension we can obtain similarly to Theorem 7 the following result for general l_q norm:

Theorem 35 *Let $p \in (0, \infty]$, $q \in [1, \infty]$ and denote $a_{p,q,d} = \frac{1}{8\sqrt{e}} d^{-\frac{1}{2} + \left[\frac{1}{2} - \frac{1}{q}\right]_{-} + \left[\frac{1}{p} - \frac{1}{2}\right]_{-}}$.*
If $7 \leq k \leq d + 5$ and

$$\frac{\sigma}{\delta} \leq a_{p,q,d} N^{-\frac{2}{k-6}}$$

then, for every (δ, q)-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta, q)$, there exists a neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and depth $O\left(Nk \log_2(k)\right)$ that (σ, p)-robustly memorizes the dataset $\mathcal{D}$.

Note that for $q = 2$ we recover Theorem 7. Similar to Theorem 8, we also have a lower bound:

Theorem 36 *Let $p \in (0, \infty]$, $q \in [1, \infty]$ and denote $b_{p,q,d} = 2416 d^{\left[\frac{1}{2} - \frac{1}{q}\right]_{+} + \left[\frac{1}{p} - \frac{1}{2}\right]_{+}}$.*
If $1 \leq k \leq d - 1$ and

$$\frac{\sigma}{\delta} > b_{p,q,d} N^{-\frac{2}{k}}$$

then, there exists a (δ, q)-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,2}(\delta, q)$ such that every neural network $f : \mathbb{R}^d \to \mathbb{R}$ with width k and any depth cannot (σ, p)-robustly memorize the dataset $\mathcal{D}$.

Again, for $q = 2$ we recover Theorem 8. To prove the bounds in Theorems 35, 36 we reduce (resp. enlarge) δ by a factor of $c^{-}_{q,2}(d)$ (resp. $c^{+}_{q,2}(d)$) so that the relation between the norms l_2, l_q obtained from Lemma 69 would enable us to deal with datasets separated under l_2 norm and then use Theorems 7, 8. This scaling of δ results in constants $a_{p,q,d} = a_{p,d} \cdot c^{-}_{q,2}(d) = a_{p,d} \cdot d^{\left[\frac{1}{2}-\frac{1}{q}\right]_-}$, and $b_{p,q,d} = b_{p,d} \cdot c^{+}_{q,2}(d) = b_{p,d} \cdot d^{\left[\frac{1}{2}-\frac{1}{q}\right]_+}$ where $a_{p,d}, b_{p,d}$ are the constant in Theorems 7, 8 respectively. The gap between $a_{p,q,d}$ and $b_{p,q,d}$ is thus given by

$$\frac{b_{p,q,d}}{a_{p,q,d}} = \frac{b_{p,d} \cdot d^{\left[\frac{1}{2}-\frac{1}{q}\right]_+}}{a_{p,d} \cdot d^{\left[\frac{1}{2}-\frac{1}{q}\right]_-}} = c \cdot d^{\frac{1}{2}} \cdot d^{\left|\frac{1}{2}-\frac{1}{p}\right|} \cdot d^{\left|\frac{1}{2}-\frac{1}{q}\right|}$$

for $c = 19328\sqrt{e}$. When $1 \leq p$ we have $d^{\left|\frac{1}{2}-\frac{1}{p}\right|} \cdot d^{\left|\frac{1}{2}-\frac{1}{q}\right|} \leq d$ and so the gap is bounded by $c \cdot d^{\frac{3}{2}}$. In particular, this would be the case when $p = q$ (since $q \in [1, \infty]$).

Appendix F. Proofs for Section E

F.1. Robust Memorization With Large Width

Proof [Proof of Theorem 31] Denote $r = c^{+}_{p,q}(d)\sigma$ and let $\mathcal{D} \in \mathcal{D}_{d,N,C}(\delta, q)$ be a (δ, q)-separated dataset. We know that $r < \delta/2$ and so there exists some $0 < w$ such that $\delta = 2r + w$. Denote $\tilde{\epsilon} = \frac{w^q}{4d(w+2r)^q} = \frac{(\delta-2r)^q}{4d\delta^q}$ then $0 < \tilde{\epsilon} < 1/2$, $q \in (1, \infty) \setminus \mathbb{N}$ and so from Lemma 44 we get that there exists a neural network $g_{\tilde{\epsilon},q} : \mathbb{R} \to \mathbb{R}$ with width 9 and depth

$$O\left(q\tilde{\epsilon}^{-\frac{1}{q}}\left(\log_2(q\tilde{\epsilon}^{-1}) + q\right)\right) = O\left(d^{\frac{1}{q}}\lambda q\left(\log_2(dq\lambda^q) + q\right)\right) ,$$

(where $\lambda = \left(1 - \frac{2r}{\delta}\right)^{-1}$) such that $|g_{\tilde{\epsilon},q}(\alpha) - \alpha^q| \leq \tilde{\epsilon}$ for every $\alpha \in [0, 1]$. Using this network, from Lemma 38 we get that for every $(x_i, y_i) \in \mathcal{D}$ there exists a neural network $f_{x_i,w,q} : \mathbb{R}^d \to \mathbb{R}$ with width $W = d+2+\mathcal{W}(g_{\tilde{\epsilon},q}) = d+11$, and depth $L = O\left(d\mathcal{L}(g_{\tilde{\epsilon},q})\right) = O\left(d^{1+\frac{1}{q}}\lambda q\left(\log_2(dq\lambda^q) + q\right)\right)$, such that for all $x \in \mathbb{R}^d$ we have $f_{x_i,w,q}(x) \leq y_i$ and

$$f_{x_i,w,q}(x) = \begin{cases} y_i & \|x - x_i\|_q \leq r \\ 0 & r + w \leq \|x - x_i\|_q \end{cases} ,$$

Finally, because $\mathcal{D}$ is (δ, q)-separated, from Theorem 37 there exists a neural network $F_{d,\delta,r,q} : \mathbb{R}^d \to \mathbb{R}$ with width $d + 12$ and depth $O\left(Nd^{1+\frac{1}{q}}\lambda q\left(\log_2(dq\lambda^q) + q\right)\right)$ that (r, q)-robustly memorizes the dataset $\mathcal{D}$. Define $f = F_{d,\delta,r,q}$ and let $x \in B^d_p(x_i, \sigma)$. Then, by definition of $c^{+}_{p,q}(d)$ and Lemma 70, we have $x \in B^d_q(x_i, r)$ and so $f(x) = F_{d,\delta,r,q}(x) = y_i$. Now $d + 12 \leq k$ and so by padding each hidden layer of f with $k - (d + 12)$ neurons we obtain f with width k and depth

$$O\left(Nd^{1+\frac{1}{q}}\lambda q\left(\log_2(dq\lambda^q) + q\right)\right)$$

that (σ, p)-robustly memorizes the dataset $\mathcal{D}$. ■

Proof [Proof of Theorem 32] Following the exact same proof as the proof of Theorem 31 where instead of Lemma 44 we use Lemma 42, we obtain a neural network $F_{d,\delta,r,q} : \mathbb{R}^d \to \mathbb{R}$ with width $d + 9$ and depth $O\left(Ndq\log_2(dq\lambda^q)\right)$ that (r, q)-robustly memorizes the dataset $\mathcal{D}$. Define $f = F_{d,\delta,r,q}$. Now $d+9 \leq k$ and so by padding each hidden layer of f with $k-(d+9)$ neurons we obtain f with width k and depth $O\left(Ndq\log_2(dq\lambda^q)\right)$ that (σ,p)-robustly memorizes the dataset $\mathcal{D}$. ■

Proof [Proof of Theorem 33] We prove for $q=1$ and $q=\infty$:

- Case $q=1$: Follow the proof of Theorem 31 where instead of Lemma 44 use the identity map $g_1(\alpha)=\alpha$ with width and depth of 1. We obtain a neural network $F_{d,\delta,r,1}:\mathbb{R}^d\to\mathbb{R}$ with width $d+4$ and depth $O(Nd)$ that $(r,1)$-robustly memorizes the dataset $\mathcal{D}$. Define $f=F_{d,\delta,r,1}$. Now $d+4\leq k$ and so by padding each hidden layer of f with $k-(d+4)$ neurons we obtain f with width k and depth $O(Nd)$ that (σ,p)-robustly memorizes the dataset $\mathcal{D}$.

- Case $q=\infty$: Follow the proof of Theorem 31 where instead of using Lemma 44 and Lemma 38 use Lemma 39. We get that for every $(x_i,y_i)\in\mathcal{D}$ and every $0<w$, there exists a neural network $f_{x_i,w,\infty}:\mathbb{R}^d\to\mathbb{R}$ with width $W=d+3$, and depth $L=O(d)$, such that for all $x\in\mathbb{R}^d$ we have $f_{x_i,w,\infty}(x)\leq y_i$ and

$$f_{x_i,w,\infty}(x)=\begin{cases}y_i & \|x-x_i\|_\infty\leq r\\ 0 & r+w\leq\|x-x_i\|_\infty\end{cases},$$

  Finally, because $\mathcal{D}$ is (δ,∞)-separated, from Theorem 37 we obtain a neural network $F_{d,\delta,r,\infty}:\mathbb{R}^d\to\mathbb{R}$ with width $d+4$ and depth $O(Nd)$ that (r,∞)-robustly memorizes the dataset $\mathcal{D}$. Define $f=F_{d,\delta,r,1}$ and conclude that f has width k and depth $O(Nd)$ and it (σ,p)-robustly memorizes the dataset $\mathcal{D}$.

■

F.2. Robust Memorization With Small Width

Proof [Proof of Theorem 35] Let $\frac{\sigma}{\delta}\leq\frac{c^-_{q,2}(d)}{2c^+_{p,2}(d)}\sqrt{\frac{k-6}{16ed}}N^{-\frac{2}{k-6}}$ and let $\mathcal{D}\in\mathcal{D}_{d,N,C}(\delta,q)$ be a (δ,q)-separated dataset. Denote $\sigma':=c^+_{p,2}(d)\sigma$ and $\delta'=c^-_{q,2}(d)\delta$, then $\frac{2\sigma'}{\delta'}\leq\frac{1}{4\sqrt{e}}\sqrt{\frac{k-6}{d}}N^{-\frac{2}{k-6}}$ and from Lemma 69 $\mathcal{D}\in\mathcal{D}_{d,N,C}(\delta',2)$. Therefore, from Theorem 18 we have that $\mathcal{D}$ is $(\sigma',\epsilon,k-6)$-orthogonally preservable with $\epsilon=\frac{1}{2\sqrt{e}}\sqrt{\frac{k-6}{d}}N^{-\frac{2}{k-6}}$. Note that $\frac{1}{\epsilon}<\frac{\delta'}{2\sigma'}$ and so from Theorem 16 and Lemma 21 we conclude that there exists a neural network $f:\mathbb{R}^d\to\mathbb{R}$ with width k and depth

$$O\left(Nk\log_2\left(\frac{k}{1-\frac{2\sigma'}{\epsilon\delta'}}\right)\right)$$

that (σ,p)-robustly memorizes the dataset $\mathcal{D}$.

Now $\frac{2c^+_{p,2}(d)\sigma}{c^-_{q,2}(d)\delta}\leq\frac{1}{2}\epsilon$ so $\frac{2\sigma'}{\epsilon\delta'}\leq\frac{1}{2}$ and hence the depth of f is $O\left(Nk\log_2(k)\right)$. The theorem follows by noting that $\frac{1}{8\sqrt{e}}d^{-\frac{1}{2}+\left[\frac{1}{2}-\frac{1}{q}\right]_-+\left[\frac{1}{p}-\frac{1}{2}\right]_-}\leq\frac{c^-_{q,2}(d)}{2c^+_{p,2}(d)}\sqrt{\frac{k-6}{16ed}}$. ■

Proof [Proof of Theorem 36] Let $\frac{\sigma}{\delta}>\frac{c^+_{q,2}(d)}{c^-_{p,2}(d)}2416N^{-\frac{2}{k}}$ and denote $\sigma':=c^-_{p,2}(d)\sigma$ and $\delta'=c^+_{q,2}(d)\delta$, then $\frac{2\sigma'}{\delta'}>4832N^{-\frac{2}{k}}$ and so from Theorem 19 we get that there exists a $(\delta',2)$-separated dataset $\mathcal{D} \in \mathcal{D}_{d,N,2}(\delta', 2)$ which is not (σ', k)-preservable. Therefore, by Theorem 17 we get that there isn't a neural network f with width equal to k that (σ, p)-robustly memorizes the dataset $\mathcal{D}$. Note that from Lemma 69, $\mathcal{D} \in \mathcal{D}_{d,N,2}(\delta, q)$ and so by noting that $2416 d^{\left[\frac{1}{2}-\frac{1}{q}\right]_+ + \left[\frac{1}{p}-\frac{1}{2}\right]_+} = \frac{c^+_{q,2}(d)}{c^-_{p,2}(d)} 2416$ we are done. ■

Appendix G. Lemmas Used for Network Approximations

Theorem 37 *Let $p \in [1, \infty]$, $r < \tau/2$ and let $\mathcal{D} \in \mathcal{D}_{k,N,C}$ be a dataset such that for all x_i, x_j with $y_i \neq y_j$ we have $\tau \leq \|x_i - x_j\|_p$. Denote $w = \tau - 2r$. Assume that for every $(x_i, y_i) \in \mathcal{D}$ there exists a neural network $\hat{f}_{x_i,w,p} : \mathbb{R}^k \to \mathbb{R}^{k+1}$ with width $W_{w,p}$ and depth $L_{w,p}$ such that for all $x \in \mathbb{R}^k$ we have $\hat{f}_{x_i,w,p}(x) = (f_{x_i,w,p}(x), x)$, $f_{x_i,w,p}(x) \leq y_i$ and*

$$f_{x_i,w,p}(x) = \begin{cases} y_i & \|x - x_i\|_p \leq r \\ 0 & r + w \leq \|x - x_i\|_p \end{cases},$$

Then there exists a neural network $F_{k,\tau,r,p} : \mathbb{R}^k \to \mathbb{R}$ with width $W_{w,p} + 1$ and depth $O(N L_{w,p})$ that (r, p)-robustly memorizes the dataset $\mathcal{D}$.

Proof [Proof of Theorem 37] For every data point $(x_i, y_i) \in \mathcal{D}$, we will use the (approximate) indicator network $f_{x_i,w,p}$ in order to inspect whether the input x lies inside the ball $B^k_p(x_i, r)$, and keep the answer in a neuron denoted by z_i. The construction is done in a way that ensures that if $x \in B^k_p(x_j, r)$ for some j, then $\max\{z_i \mid 1 \leq i \leq N\} = y_j$. Hence, by keeping track of the maximum value of the z_i's up to j (which we will denote by m_j) we will manage to return the desired result (by returning m_N).

Let $1 \leq i \leq N$. Using $\hat{f}_{x_i,w,p}(x) = (f_{x_i,w,p}(x), x)$ from the assumption of the theorem, we will compute it in a dedicated sub-net A_i (see Figure 4) and update the running maximum value m_i.

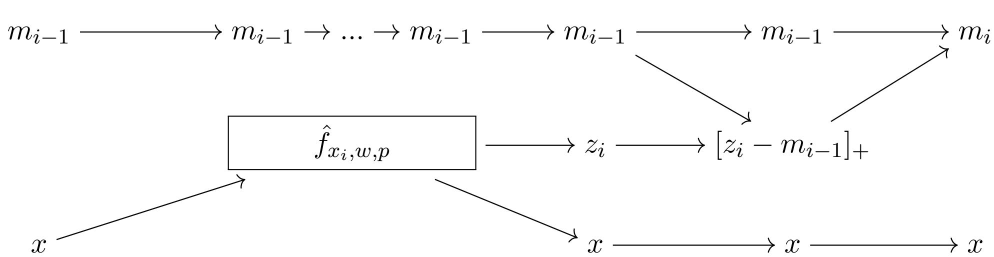


Figure 4: The architecture of A_i

A_i will preform the following:

- append a neuron with the value m_{i-1}, where $m_0 = 0$ (this is the accumulated maximum of previous indicator computations).
- compute $z_i = f_{x_i,w,p}(x)$ (current indicator computation).

- compute $[z_i - m_{i-1}]_+$.
- compute $m_i = [m_{i-1} + [z_i - m_{i-1}]_+]_+$ ($m_i = \max\{z_j \mid j \le i\}$).

Note that the width and depth of A_i satisfy $\mathcal{W}(A_i) = W_{w,p} + 1$ and $\mathcal{L}(A_i) = O(L_{w,p})$. Now, we will define $F_{k,\tau,r,p}(x)$ to simply return m_N (see Figure 5).

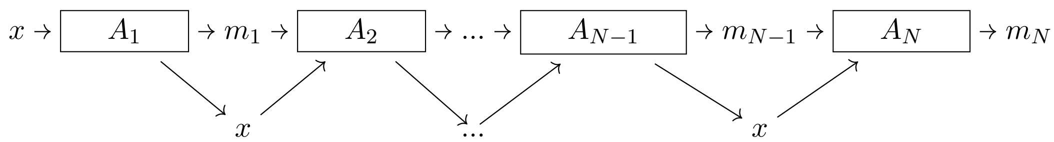


Figure 5: The architecture of $F_{k,\tau,r}$

Note that by construction the maximal width of $F_{k,\tau,r,p}$ is the maximal width of A_i so $F_{k,\tau,r,p} : \mathbb{R}^k \to \mathbb{R}$ is a neural network of width $W_{w,p} + 1$ and depth $\mathcal{L}(F_{k,\tau,r,p}) = O(NL_{w,p})$.

Let us show that $F_{k,\tau,r,p}(x)$ indeed (r,p)-robustly memorizes the dataset $\mathcal{D}$. Let $i \in [N]$ and let $x \in B_p^k(x_i, r)$. We inspect the possible values of $f_{x_j,w,p}$ for every $j \in [N]$:

- If $j = i$: In this case $\|x - x_j\|_p = \|x - x_i\|_p \le r$ so $f_{x_j,w,p}(x) = y_j = y_i$.
- If $j \ne i$:
  - If $y_i = y_j$: By definition $f_{x_j,w,p}(x) \le y_j = y_i$.
  - If $y_i \ne y_j$: Note that in this case since x is in a different class than that of x_j (and because $\|\cdot\|_p$ is a norm as $p \in [1,\infty]$) we have from the triangle inequality $\tau - r \le \|x - x_j\|_p$. Hence, in this case $r + w = \tau - r \le \|x - x_j\|_p$ so $f_{x_j,w,p}(x) = 0$.

We see that when $j \ne i$ one has $z_j = f_{x_j,w,p}(x) \le y_i$ and when $i = j$, $z_i = f_{x_i,w,p}(x) = y_i$, so $\max\{z_j \mid 1 \le j \le N\} = y_i$, but by construction we have $m_j = \max\left\{z_{j'} \mid j' \le j\right\}$ so we conclude that $F_{k,\tau,r,p}(x) = m_N = y_i$. We have thus shown that $F_{k,\tau,r,p}$ does indeed (r,p)-robustly memorizes the dataset $\mathcal{D}$. ■

The following lemma provides a construction of an approximate indicator function that returns a desired value on a fixed l_p ball.

Lemma 38 *Let $p \in [1,\infty)$, $x_0 \in \mathbb{R}^k$, $0 < y_0$ and let $0 < r$ and any $0 < w$. Denote $\tilde{\epsilon} = \frac{w^p}{4k(w+2r)^p}$. Assume that there exists a network $g_{\tilde{\epsilon},p}$ with width $\mathcal{W}(g_{\tilde{\epsilon},p})$ and depth $\mathcal{L}(g_{\tilde{\epsilon},p})$ such that $|g_{\tilde{\epsilon},p}(t) - t^p| < \tilde{\epsilon}$ for every $t \in [0,1]$. Then there exists a neural network $f_{x_0,w,p} : \mathbb{R}^k \to \mathbb{R}$ with width $k + 2 + \mathcal{W}(g_{\tilde{\epsilon},p})$ and depth $O\left(k\mathcal{L}(g_{\tilde{\epsilon},p})\right)$ such that for all $x \in \mathbb{R}^k$ we have $f_{x_0,w,p}(x) \le y_0$ and*

$$f_{x_0,w,p}(x) = \begin{cases} y_0 & , \|x - x_0\|_p \le r \\ 0 & , r + w \le \|x - x_0\|_p \end{cases},$$

Furthermore, $f_{x_0,w,p}$ can be modified to return also the input vector x without changing its width and depth.

Similar analysis yields the following lemma for the case that $p = \infty$:

Lemma 39 *Let $x_0 \in \mathbb{R}^k$, $0 < y_0$ and let $0 < r$ and any $0 < w$. Then there exists a neural network $f_{x_0,w,\infty} : \mathbb{R}^k \to \mathbb{R}$ with width $k+3$ and depth $O(k)$ such that for all $x \in \mathbb{R}^k$ we have $f_{x_0,w,\infty}(x) \le y_0$ and*

$$f_{x_0,w,\infty}(x) = \begin{cases} y_0 & , \|x - x_0\|_\infty \le r \\ 0 & , r + w \le \|x - x_0\|_\infty \end{cases},$$

Furthermore, $f_{x_0,w,\infty}$ can be modified to return also the input vector x without changing its width and depth.

Proof [Proof of Lemma 38] For a vector v we will denote by $(v)_j$ its j-th coordinate. We will use $g_{\tilde{\epsilon},p}$ to approximate the values of $|(x)_j - (x_0)_j|^p$ for $1 \le j \le k$, and use them to approximate the p norm $\|x - x_0\|_p^p$. This, in turn will allow us to return the desired result. Let $1 \le j \le k$. We would like to compute $|(x)_j - (x_0)_j|^p$ using $g_{\tilde{\epsilon},p}$, so we have to modify the input so that it lies in the range $[0, 1]$. $|(x)_j - (x_0)_j|$ is unbounded in general so we have to normalize it carefully. Denote $\delta = (w + 2r)$, and the following functions of $x \in \mathbb{R}^k$:

$$\begin{aligned} (a)_j &= (x)_j - (x_0)_j \\ (b)_j &= \left[\left([(a)_j]_+ + [-(a)_j]_+\right)/\delta\right]_+ \\ (c)_j &= [2(b)_j - 1]_+ \end{aligned}.$$

We now show some properties of these quantities.

1. $(b)_j = \frac{|(x)_j - (x_0)_j|}{\delta}$, $(c)_j = \left[\frac{2|(x)_j - (x_0)_j|}{\delta} - 1\right]_+$:

   Follows immediately from the definition of a, b, c.

2. If $|(x)_j - (x_0)_j| \le \delta$, then $0 \le (b)_j \le 1$:

   Follows immediately from the definition of a, b.

3. If $|(x)_j - (x_0)_j| \le r$, then $(c)_j = 0$:

   We have $\frac{2|(x)_j - (x_0)_j|}{\delta} - 1 \le \frac{2r}{\delta} - 1 = \frac{2r}{2r+w} - 1 \le 0$, so $(c)_j = 0$.

4. If $\delta \le |(x)_j - (x_0)_j|$, then $(c)_j \ge 1$:

   We have $\frac{2|(x)_j - (x_0)_j|}{\delta} - 1 \ge \frac{2\delta}{\delta} - 1 = 1$.

We will now construct $f_{x_0,w,p}$ using these quantities. From the assumption of the theorem, there exists a neural network $g_{\tilde{\epsilon},p} : \mathbb{R} \to \mathbb{R}$ with width $\mathcal{W}(g_{\tilde{\epsilon},p})$ and depth $\mathcal{L}(g_{\tilde{\epsilon},p})$ such that $|g_{\tilde{\epsilon},p}(t) - t^p| < \tilde{\epsilon}$ for every $t \in [0, 1]$. For every coordinate j we apply the sub-network M_j (see Figure 6 for layout of M_j) which preforms the following:

- compute the penalty $(c)_j$ and the normalized input $(b)_j$.
- compute $(\gamma)_j = g_{\tilde{\epsilon},p}((b)_j)$ (this will approximate $(b)_j^p$ for relevant values of $(x)_j$).
- compute $(\eta)_j = [\delta^p(\gamma)_j]_+$ (this will approximate $|(x)_j - (x_0)_j|^p$ for relevant values of $(x)_j$).

- compute the neuron $(s)_j = (\eta)_j + (r^p + \frac{3}{4}w^p)(c)_j$ (current coordinate to the power p with its penalty).

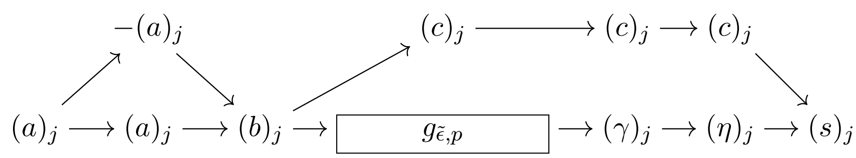


Figure 6: The architecture of M_j

We note that the width and depth of M_j satisfy $\mathcal{W}(M_j) = 1 + \mathcal{W}(g_{\tilde{\epsilon},p})$, and $\mathcal{L}(M_j) = 5 + \mathcal{L}(g_{\tilde{\epsilon},p})$. In order to compute M_j for every coordinate sequentially, we append to every layer in M_j memory neurons with x, and with the accumulated sum $\Sigma_{j-1} = \sum_{i=1}^{j-1}(s)_i$.

After computing M_j for all $1 \le j \le k$ sequentially, we obtain the neuron

$$\Sigma_k(x) = \sum_{j=1}^{k}(\eta(x))_j + (r^p + \frac{3}{4}w^p)(c(x))_j.$$

Finally the network will return as output the value

$$f_{x_0,w,p}(x) = \left[y_0\left(1 - \frac{\left[\Sigma_k(x) - r^p - \frac{w^p}{4}\right]_+}{w^p/2}\right)\right]_+ ,$$

as can be seen in the following sketch of the layout of $f_{x_0,w,p}$:

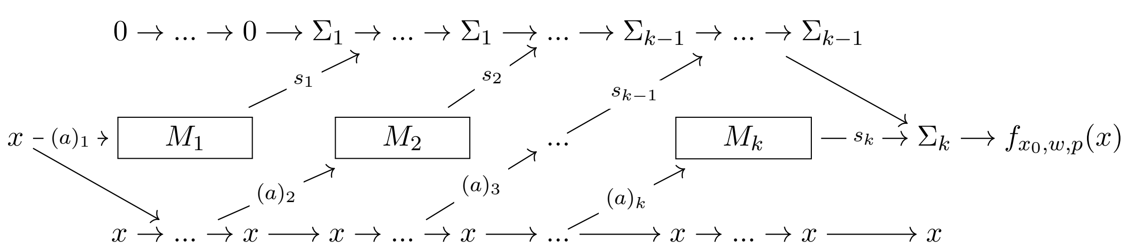


Figure 7: The architecture of $f_{x_0,w,p}$

Note that one can append to the output neuron the vector x wothout changing the width and depth of $f_{x_0,w,p}$.

Let us show that $f_{x_0,w,p}$ behaves as we desired. Let $x \in \mathbb{R}^k$.

- Note that $0 \le \frac{\left[\Sigma_k(x) - r^p - \frac{w^p}{4}\right]_+}{w^p/2}$ and so because ReLU is increasing we always have

$$f_{x_0,w,p}(x) = \left[y_0\left(1 - \frac{\left[\Sigma_k(x) - r^p - \frac{w^p}{4}\right]_+}{w^p/2}\right)\right]_+ \le y_0 .$$

- If $\|x - x_0\|_p^p \le r^p$, then for every $1 \le j \le k$ we have $|(x)_j - (x_0)_j| \le r < \delta$, so $(c)_j = 0$ and $0 \le (b)_j = \frac{|(x)_j - (x_0)_j|}{\delta} < 1$. Hence

$$
\begin{aligned}
\Sigma_k(x) &= \sum_{j=1}^{k} (\eta)_j = \sum_{j=1}^{k} [\delta^p (\gamma)_j]_+ \\
&\le \sum_{j=1}^{k} [\delta^p ((b)_j^p + \tilde{\epsilon})]_+ = \sum_{j=1}^{k} \left[\delta^p \left(\frac{|(x)_j - (x_0)_j|^p}{\delta^p} + \tilde{\epsilon} \right) \right]_+ \\
&= \sum_{j=1}^{k} [|(x)_j - (x_0)_j|^p + \delta^p \tilde{\epsilon}]_+ = \left(\sum_{j=1}^{k} |(x)_j - (x_0)_j|^p \right) + k\delta^p \tilde{\epsilon} \\
&\le r^p + k\delta^p \frac{w^p}{4k\delta^p} = r^p + \frac{w^p}{4} \ ,
\end{aligned}
$$

which yields $f_{x_0,w,p}(x) = y_0$.

- If $(r + w)^p \le \|x - x_0\|_p^p$, then one of the following occurs:

  1. There exists some $1 \le j_0 \le k$ such that $\delta \le |(x)_{j_0} - (x_0)_{j_0}|$. In this case, $(c)_{j_0} \ge 1$, and therefore:

$$
\begin{aligned}
\Sigma_k(x) &= \sum_{j=1}^{k} (\eta)_j + (r^p + \frac{3}{4} w^p)(c(x))_j \\
&\ge \sum_{j=1}^{k} (r^p + \frac{3}{4} w^p)(c(x))_j \ge (r^p + \frac{3}{4} w^p)(c(x))_{j_0} \\
&\ge r^p + \frac{3}{4} w^p \ .
\end{aligned}
$$

2. For every $1 \le j \le k$ we have $|(x)_j - (x_0)_j| < \delta$, so $0 \le (b)_j = \frac{|(x)_j-(x_0)_j|}{\delta} < 1$. Hence

$$
\begin{aligned}
\Sigma_k(x) &= \sum_{j=1}^{k}(\eta)_j + (r^p + \frac{3}{4}w^p)(c(x))_j \\
&\ge \sum_{j=1}^{k}(\eta)_j = \sum_{j=1}^{k}[\delta^p(\gamma)_j]_+ \ge \sum_{j=1}^{k}[\delta^p((b)_j^p - \tilde{\epsilon})]_+ \\
&= \sum_{j=1}^{k}\left[\delta^p\left(\frac{|(x)_j-(x_0)_j|^p}{\delta^p} - \tilde{\epsilon}\right)\right]_+ \\
&= \sum_{j=1}^{k}\left[|(x)_j-(x_0)_j|^p - \delta^p\tilde{\epsilon}\right]_+ \\
&\ge \sum_{j=1}^{k}\left(|(x)_j-(x_0)_j|^p - \delta^p\tilde{\epsilon}\right) \\
&\ge (r+w)^p - k\delta^p\tilde{\epsilon} = (r+w)^p - \frac{w^p}{4} \\
&\ge r^p + w^p - \frac{w^p}{4} = r^p + \frac{3}{4}w^p\,. && \text{(since } 1 \le p\text{)}
\end{aligned}
$$

In any case we have $\Sigma_k(x) \ge r^p + \frac{3}{4}w^p$, which yields

$$
f_{x_0,w,p}(x) \le \left[y_0\left(1 - \frac{[r^p + \frac{3w^p}{4} - r^p - \frac{w^p}{4}]_+}{w^p/2}\right)\right]_+ = 0
$$

.

From all of the above combined we get that $f_{x_0,w,p}$ behaves as desired. Note that by construction, the maximal width of $f_{x_0,w,p}$ is the maximal width of M_j (which is $1 + \mathcal{W}(g_{\tilde{\epsilon},p})$) plus one memory neuron for the accumulated sum Σ_j, and k neurons to carry x. Thus we have a total of $k + 1 + \mathcal{W}(g_{\tilde{\epsilon},p}) + 1$, so $f_{x_0,w,p} : \mathbb{R}^k \to \mathbb{R}$ is a neural network of width $k + 2 + \mathcal{W}(g_{\tilde{\epsilon},p})$. Additionally, $\mathcal{L}(f_{x_0,w,p}) = O(k\mathcal{L}(M_j))$ so $f_{x_0,w,p}$ has depth $O\left(k\mathcal{L}(g_{\tilde{\epsilon},p})\right)$.

■

Proof [Proof of Lemma 39] Denote the following functions of $x \in \mathbb{R}^k$:

$$
\begin{aligned}
(a)_j &= (x)_j - (x_0)_j \\
(b)_j &= \left[\left([(a)_j]_+ + [-(a)_j]_+\right)\right]_+ \quad .
\end{aligned}
$$

Note that for all j we have $(b)_j = |(x)_j - (x_0)_j|$. We will now construct $f_{x_0,w,\infty}$ using these quantities. For every coordinate j we apply the sub-network M_j (see Figure 8 for layout of M_j) which updates the running maximum m_j (initialize $m_0 = 0$).

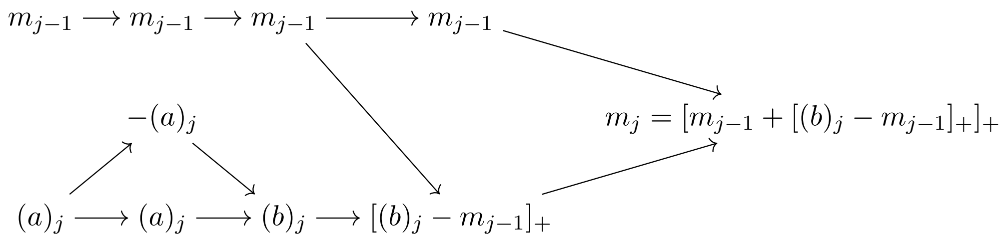


Figure 8: The architecture of M_j

We note that the width and depth of M_j satisfy $\mathcal{W}(M_j) = 3$, and $\mathcal{L}(M_j) = 5$. In order to compute M_j for every coordinate sequentially, we append to every layer in M_j memory neurons with x.

After computing M_j for all $1 \leq j \leq k$ sequentially, we obtain the neuron

$$m_k(x) = \max\{(b)_j \mid 1 \leq j \leq k\} = \max\{|(x)_j - (x_0)_j| \mid 1 \leq j \leq k\} = \|x - x_0\|_\infty.$$

Finally the network will return as output the value

$$f_{x_0,w,\infty}(x) = \left[y_0 \left(1 - \frac{[m_k(x) - r]_+}{w}\right)\right]_+ ,$$

as can be seen in the following sketch of the layout of $f_{x_0,w,\infty}$:

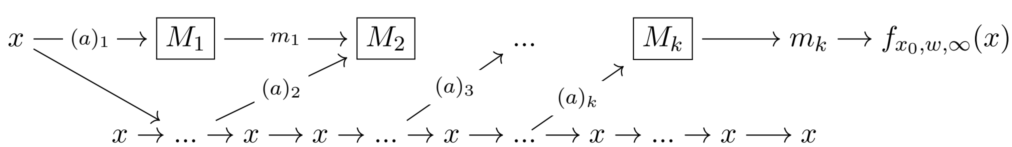


Figure 9: The architecture of $f_{x_0,w,\infty}$

Note that one can append to the output neuron the vector x wothout changing the width and depth of $f_{x_0,w,\infty}$.

Let us show that $f_{x_0,w,\infty}$ behaves as we desired. Let $x \in \mathbb{R}^k$.

- Note that $0 \leq \frac{[m_k(x)-r]_+}{w}$ and so because ReLU is increasing we always have $f_{x_0,w,\infty}(x) \leq y_0$.
- If $\|x - x_0\|_\infty \leq r$, then $m_k(x) - r \leq 0$ and so $f_{x_0,w,\infty}(x) = y_0$.
- If $r + w \leq \|x - x_0\|_\infty$, then $\frac{[m_k(x)-r]_+}{w} = \frac{\|x-x_0\|_\infty - r}{w} \geq \frac{w}{w} = 1$ and so $f_{x_0,w,\infty}(x) = 0$.

From all of the above combined we get that $f_{x_0,w,\infty}$ behaves as desired. Note that by construction, the maximal width of $f_{x_0,w,\infty}$ is the maximal width of M_j (which is 3) plus k neurons to carry x. Thus we have a total of $k+3$, so $f_{x_0,w,\infty} : \mathbb{R}^k \to \mathbb{R}$ is a neural network of width $k+3$. Additionally, $\mathcal{L}(f_{x_0,w,\infty}) = O(k\mathcal{L}(M_j))$ so $f_{x_0,w,\infty}$ has depth $O(k)$. ■

Remark 40 (Constructions equivalent to Theorems 38 and 39) *In Lemma 38, if we construct $f_{x_0,w,p}$ the same way where instead of computing the M_j components sequentially we stack them and perform the computation in parallel, we would obtain a network $f_{x_0,w,p}$ that behaves exactly the same and has width $k \cdot \mathcal{W}(M_j) = k(1 + \mathcal{W}(g_{\tilde{\epsilon},p}))$ and depth $O(\mathcal{L}(M_j)) = O(\mathcal{L}(g_{\tilde{\epsilon},p}))$. Similarly, in Lemma 39 we would get a network with width $2k$ and depth $O(1)$. Returning also the input vector would increase the width by additional k neurons in both cases.*

The following lemma is used to approximate the square of a given number.

Lemma 41 *(Elbrächter et al., 2019, Proposition III.2) Let $0 < \epsilon < \frac{1}{2}$. Then there exists a neural network $g_\epsilon : \mathbb{R} \to \mathbb{R}$ with width 3 and depth $O\left(\log_2(\epsilon^{-1})\right)$ such that $|g_\epsilon(\alpha) - \alpha^2| \leq \epsilon$ for every $\alpha \in [0, 1]$.*

The following lemma is used to approximate the p power of a given number.

Lemma 42 *Let $p \in \mathbb{N}_{\geq 2}$, $0 < \epsilon < 1/2$. Then there exists a neural network $g_{\epsilon,p} : \mathbb{R} \to \mathbb{R}$ with width 6 and depth $O\left(p \log_2(p\epsilon^{-1})\right)$ such that $|g_{\epsilon,p}(\alpha) - \alpha^p| < \epsilon$ for every $\alpha \in [0, 1]$.*

In the proof we will use the following lemma to compute multiplication:

Lemma 43 *(Elbrächter et al., 2019, Proposition III.3) Let $0 < \epsilon < \frac{1}{2}$. Then there exists a neural network $h_\epsilon : \mathbb{R} \to \mathbb{R}$ with width 5 and depth $O\left(\log_2(\epsilon^{-1})\right)$ such that $|h_\epsilon(\alpha, \beta) - \alpha\beta| \leq \epsilon$ for every $\alpha, \beta \in [0, 1]$.*

Proof [Proof of Lemma 42] Denote $\epsilon_2 = \frac{\epsilon}{p-1}$ then from Lemma 43 there exists a neural network $h_{\epsilon_2} : \mathbb{R} \to \mathbb{R}$ with width 5 and depth $O\left(\log_2(\epsilon_2^{-1})\right)$ such that $|h_{\epsilon_2}(\alpha, \beta) - \alpha\beta| \leq \epsilon_2$ for every $\alpha, \beta \in [0, 1]$. We will apply h_{ϵ_2} repeatedly p times as in the following figure:

$$\begin{tikzcd}
 & \alpha \arrow[r] & \boxed{h_{\epsilon_2}} \arrow[r] & \boxed{h_{\epsilon_2}} \arrow[r] & \cdots \arrow[r] & \boxed{h_{\epsilon_2}} \arrow[r] & \boxed{h_{\epsilon_2}} \arrow[r] & g_{\epsilon,p}(\alpha) \\
\alpha \arrow[r] \arrow[ur] & \alpha \arrow[r] \arrow[ur] & \alpha \arrow[r] \arrow[ur] & \alpha \arrow[r] & \cdots \arrow[r] & \alpha \arrow[ur] & &
\end{tikzcd}$$

Figure 10: The architecture of $g_{\epsilon,p}$

Now from the definition we have that for all $\alpha \in [0, 1]$:

$$|g_{\epsilon,p}(\alpha) - \alpha^p| \leq \epsilon_2 \sum_{j=0}^{p-2} \alpha^j \leq \epsilon_2(p - 1) = \epsilon$$

Note that $g_{\epsilon,p}$ has width $5 + 1 = 6$ and depth $O\left(p \log_2(\epsilon_2^{-1})\right) = O\left(p \log_2(p\epsilon^{-1})\right)$ and we are done. ■

For $p \notin \mathbb{N}$ we have the following generalization:

Lemma 44 *Let $p \in (1,\infty) \setminus \mathbb{N}$, $0 < \epsilon < 1$. Then there exists a neural network $g_{\epsilon,p} : \mathbb{R} \to \mathbb{R}$ with width 9 and depth $O\left(p\epsilon^{-\frac{1}{p}}\left(\log_2(p\epsilon^{-1}) + p\right)\right)$ such that $|g_{\epsilon,p}(\alpha) - \alpha^p| < \epsilon$ for every $\alpha \in [0,1]$.*

Proof [Proof of Lemma 44] Define $D = \lceil p(\pi\epsilon/2)^{-1/\lfloor p \rfloor}\rceil + 2$, and the polynomial $P_{\epsilon/2,p} : \mathbb{R} \to \mathbb{R}$ by $P_{\epsilon/2,p}(x) = \sum_{i=0}^{D} \binom{p}{i}(x-1)^i$. Then from Lemma 64 we have $|P_{\epsilon/2,p}(x) - x^p| < \epsilon/2$ for every $x \in [0,1]$. Define $Q_{\epsilon/2,p}(x) = P_{\epsilon/2,p}(x+1) = \sum_{i=0}^{D} \binom{p}{i} x^i$, and note that the coefficients satisfy $B := \max_{0\le i\le D} \left|\binom{p}{i}\right| \le 2^p$. Now, from Lemma 45 there exists a neural network $\Phi_{\epsilon/2,D,B} : \mathbb{R} \to \mathbb{R}$ with width 9 and depth $O\left(D\left(\log_2(\frac{2}{\epsilon}) + \log_2(D) + \log_2(B)\right)\right)$ such that $|\Phi_{\epsilon/2,D,B}(x) - Q_{\epsilon/2,p}(x)| < \epsilon/2$ for every $x \in [-1,1]$. Define the network $g_{\epsilon,p}(x) = \Phi_{\epsilon/2,D,B}(x-1)$. Let $\alpha \in [0,1]$, then $\alpha - 1 \in [-1,0]$ and so

$$\begin{aligned}
|g_{\epsilon,p}(\alpha) - \alpha^p| &= |\Phi_{\epsilon/2,D,B}(\alpha-1) - \alpha^p| \\
&\le |\Phi_{\epsilon/2,D,B}(\alpha-1) - Q_{\epsilon/2,p}(\alpha-1)| + |Q_{\epsilon/2,p}(\alpha-1) - \alpha^p| \\
&\le |\Phi_{\epsilon/2,D,B}(\alpha-1) - Q_{\epsilon/2,p}(\alpha-1)| + |P_{\epsilon/2,p}(\alpha) - \alpha^p| \\
&< \epsilon/2 + \epsilon/2 = \epsilon\,.
\end{aligned}$$

The width of $g_{\epsilon,p}$ is 9, and plugging D, B we get that its depth is $O\left(p\epsilon^{-\frac{1}{p}}\left(\log_2(p\epsilon^{-1}) + p\right)\right)$. ■

Lemma 45 *(Elbrächter et al., 2019, Proposition III.5) Let $0 < \epsilon < \frac{1}{2}$, $D \in \mathbb{N}$, $0 < B$, and $P : \mathbb{R} \to \mathbb{R}$ a polynomial given by $P(\alpha) = \sum_{i=0}^{D} c_i\alpha^i$ where $\max_{0\le i\le D}|c_i| \le B$. Then there exists a neural network $\Phi_{\epsilon,D,B} : \mathbb{R} \to \mathbb{R}$ with width 9 and depth $O\left(D\left(\log_2(\frac{1}{\epsilon}) + \log_2(D) + \log_2(B)\right)\right)$ such that $|\Phi_{\epsilon,D,B}(\alpha) - P(\alpha)| < \epsilon$ for every $\alpha \in [-1,1]$.*

Proof [Proof sketch for Proposition 34] Since the balls $B_q^d(x_i, c_{p,q}^+(d)\sigma)$ are disjoint and $\delta \le \|x_i - x_j\|_q$, there exists a Lipschitz continuous function $g : \mathbb{R}^d \to \mathbb{R}$ with the following properties:

- g has Lipschitz constant $\frac{2C}{\delta - 2c_{p,q}^+(d)\sigma}$.
- g has compact support $K_{\mathcal{D}}$.
- The diameter of $K_{\mathcal{D}}$ w.r.t the l_2 norm satisfies $\text{diam}(K_{\mathcal{D}}) \le 2\text{diam}(\mathcal{D})$.
- for every $1 \le i \le N$ we have $g(B_q^d(x_i, c_{p,q}^+(d)\sigma)) = y_i$.

Denote by ω_g the modulus of continuity of g, and $\omega_g^{-1}(\epsilon) = \sup\{\alpha \mid \omega_g(\alpha) \le \epsilon\}$ then by (Hanin and Sellke, 2017, Theorem 1) there exists a neural network $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}$ with width $d+1$ and depth $O\left(\frac{\text{diam}(\mathcal{D})}{\omega_g^{-1}(1/4)}\right)^{d+1}$ such that

$$\sup_{x \in K_{\mathcal{D}}} |g(x) - \mathcal{N}(x)| \le 1/4\,.$$

For every α we have $\omega_g(\alpha) = \frac{2C}{\delta - 2c^+_{p,q}(d)\sigma}\alpha$ so $\frac{\delta - 2c^+_{p,q}(d)\sigma}{8C} \leq \omega_g^{-1}(1/4)$ and hence $\mathcal{N}$ has depth $O\left(\frac{C\text{diam}(\mathcal{D})}{\delta - 2c^+_{p,q}(d)\sigma}\right)^{d+1}$. From Lemma 46 there exists a neural network $\psi : \mathbb{R} \to \mathbb{R}$ of width 2 and depth $O(C)$ such that for every $1 \leq m \leq C$ and every $t \in [-1/4 + m, m + 1/4]$ we have $\psi(t) = m$. Define $f = \psi \circ \mathcal{N}$ then f has width $d+1$ and depth $O\left(\frac{C\text{diam}(\mathcal{D})}{\delta - 2c^+_{p,q}(d)\sigma}\right)^{d+1}$. Let $1 \leq i \leq N$ and $x \in B^d_p(x_i, \sigma))$. We have

$$|y_i - \mathcal{N}(x)| = |g(x) - \mathcal{N}(x)| \leq 1/4 \, .$$

Therefore, $\mathcal{N}(x) \in [-1/4 + y_i, y_i + 1/4]$ and so $f(x) = \psi(\mathcal{N}(x)) = y_i$. Now $d + 1 \leq k$ and so by padding each hidden layer of f with $k - (d+1)$ neurons we obtain f with width k and depth

$$O\left(\frac{C\text{diam}(\mathcal{D})}{\delta - 2c^+_{p,q}(d)\sigma}\right)^{d+1}$$

that (σ, p)-robustly memorizes the dataset $\mathcal{D}$. ■

Lemma 46 *Let $C \in \mathbb{N}_{\geq 2}$. There exists a neural network $\psi : \mathbb{R} \to \mathbb{R}$ of width 2 and depth $O(C)$ such that for every $1 \leq m \leq C$ and every $t \in [-1/4 + m, m + 1/4]$ we have $\psi(t) = m$.*

Proof [Proof of Lemma 46] Define the following functions for every $x \in \mathbb{R}$ and every $0 \leq l \leq C - 1$:

- $\psi_{3l}(x) = \left[2x - \frac{1}{2}(2(C-l) - 1) - \psi_{3l-1}(x)\right]_+$ (where $\psi_{-1}(x) = 0$).
- $\psi_{3l+1}(x) = [l + 1 - \psi_{3l}(x)]_+$.
- $\psi_{3l+2}(x) = [l + 1 - \psi_{3l+1}(x)]_+$,

and define the network $\psi : \mathbb{R} \to \mathbb{R}$ by $\psi = \psi_{3C-1}$. See Figure 11 bellow:

$$\begin{tikzcd}
x \arrow[r] \arrow[dr] & \psi_0(x) \arrow[r] & \psi_1(x) \arrow[r] & \psi_2(x) \arrow[r] & \psi_3(x) \arrow[r] & \cdots \arrow[r] & \psi_{3C-3}(x) \arrow[r] & \cdots \arrow[r] & \psi(x) \\
 & x \arrow[r] & x \arrow[r] & x \arrow[r] \arrow[ur] & x \arrow[r] & \cdots \arrow[ur] & & &
\end{tikzcd}$$

Figure 11: The architecture of ψ

Then ψ has width 2 and depth $O(C)$ and a computation using the definitions yields that for all $1 \leq m \leq C$ and all $t \in [m - 1/4, m + 1/4]$ we have $\psi(t) = m$. ■

Appendix H. Background Material

H.1. Invariant Measures on $O(d)$

We recall the following facts (see e.g (Mattila, 1999, Chapter 3)):

Lemma 47 *The group $O\left(d\right)$ is compact (as a topological group with Borel topology induced by operator norm).*

Lemma 48 *The real Grassmannian $Gr_{d,k}$ is a compact Hausdorff space with respect to the topology induced by the operator norm.*

Lemma 49 *The real Grassmannian $Gr_{d,k}$ is a transitive $O\left(d\right)$-space with respect to the action $(g, V) \mapsto gV$ where for $V = Span\left\{v_1, ..., v_k\right\}$, $gV = Span\left\{gv_1, ..., gv_k\right\}$.*

Lemma 50 *The sphere $\mathbb{S}^{d-1}$ is a compact Hausdorff space with respect to the topology induced by the geodesic metric.*

Lemma 51 *The sphere $\mathbb{S}^{d-1}$ is a transitive $O\left(d\right)$-space with respect to the action $(g, x) \mapsto g^{-1}x$.*

From Lemma 54, Lemma 47 the group $O\left(d\right)$ has a unique Haar probability measure denoted here by ν_d. From Lemma 58, Lemma 48, Lemma 49 we get that $\mathrm{Gr}_{d,k}$ has a unique $O\left(d\right)$-invariant probability measure which we will denote by $\gamma_{d,k}$. Similarly from Lemma 58, Lemma 50, Lemma 51 we get that $\mathbb{S}^{d-1}$ has a unique $O\left(d\right)$-invariant probability measure which we will denote by μ_{d-1}. Furthermore these unique measures are simply the pushforward of the measure ν_d as summarized by the following:

Lemma 52 *For any $V \in Gr_{d,k}$ and any measurable $E \subseteq Gr_{d,k}$,*

$$\gamma_{d,k}\left(E\right) = \nu_d\left(\left\{g \in O\left(d\right) \mid gV \in E\right\}\right) \ .$$

Similarly, for any $x \in \mathbb{S}^{d-1}$ and any measurable $E \subseteq \mathbb{S}^{d-1}$,

$$\mu_{d-1}\left(E\right) = \nu_d\left(\left\{g \in O\left(d\right) \mid g^{-1}x \in E\right\}\right) \ .$$

Proof [Proof of Lemma 52] The claim follows from Lemma 58 and from the fact that the push-forward measure doesn't depend on the choice of the fixed point - indeed this measure is unique and is identical for any choice of the fixed point, and in particular for the choice of V and x. ■

The following presents the relation between the measure of complement spaces

Lemma 53 *Let $E \subseteq Gr_{d,k}$ measurable then*

$$\gamma_{d,k}\left(E\right) = \gamma_{d,d-k}\left(\left\{V^{\perp} \in Gr_{d,d-k} \mid V \in E\right\}\right) \ .$$

Proof [Proof of Lemma 53] Define a new measure on $\mathrm{Gr}_{d,k}$ by

$$\gamma'_{d,k}\left(E\right) = \gamma_{d,d-k}\left(\left\{V^{\perp} \in \mathrm{Gr}_{d,d-k} \mid V \in E\right\}\right)$$

. It is a probability measure and one can check that it is also $O\left(d\right)$-invariant and hence the claim follows from uniqueness. ■

H.2. Background in G-spaces

Lemma 54 *Let G be a compact group. Then, there exists a unique Haar probability measure (which is both left and right invariant) on its Borel sigma algebra.*

Proof [Proof of Lemma 54] By (Folland, 2015, Theorem 2.10) G has a left Haar measure λ. A Haar measure is a Radon measure and so by definition it is finite on compact sets, so λ is a finite positive measure and so by normalizing, λ is a left Haar probability measure. Let μ be a left Haar probability measure on G. By (Folland, 2015, Theorem 2.20) $\exists c \in (0, \infty)$ such that $\mu = c\lambda$, but $1 = \mu(G) = c\lambda(G) = c \cdot 1$ so $c = 1$ and $\mu = \lambda$. Finally, because G is compact it is unimodular and hence the obtained unique left Haar measure is also right Haar measure. ■

Lemma 55 *Let G be a compact group and denote by ν its unique probability Haar measure, then for all measurable sets $E \subseteq G$,*

$$\nu(E) = \nu\left(E^{-1}\right) \ .$$

Proof [Proof of Lemma 55] G is compact and hence unimodular and so the claim follows from (Folland, 2015, Proposition 2.31). ■

Lemma 56 *Let G be a compact group, $H \leq G$ a closed subgroup then there exists a unique G-invariant probability measure on G/H.*

Proof [Proof of Lemma 56] By (Folland, 2015, Theorem 2.51) and compactness of G, there is a G-invariant Radon measure μ on G/H which is unique up to a constant factor. Now, G/H is compact so μ is finite so we can normalize it to be a probability measure, and by the uniqueness up to a factor we get that this probability measure is unique. ■

Lemma 57 *Let G be a compact group, X a locally compact Hausdorff, transitive G-space with the action map $\alpha : G \times X \longrightarrow X$, and $x_0 \in X$ some fixed point. Define $f : G \longrightarrow X$ by $f(g) = \alpha(g, x_0)$, $H = Stab_{G,\alpha}(x_0)$, $q : G \longrightarrow G/H$ the natural quotient map. Then, H is a closed subgroup and $F = f \circ q^{-1} : G/H \longrightarrow X$ is a homeomorphism.*

Proof [Proof of Lemma 57] See beginning of (Folland, 2015, Chapter 2.6) and (Folland, 2015, Proposition 2.46). ■

Lemma 58 *Let G be a compact group with its unique probability Haar measure ν, X a locally compact Hausdorff, transitive G-space with the action map $\alpha : G \times X \longrightarrow X$, and $x_0 \in X$ some fixed point. Define $f : G \longrightarrow X$ by $f(g) = \alpha(g, x_0)$, then X has a unique G-invariant probability radon measure which is given by $f_*\nu$, where $f_*\nu$ denotes the pushforward measure of ν under f.*

Proof [Proof of Lemma 58] By Lemma 56, Lemma 57 X has a unique G-invariant probability radon measure. f is a continuous map and hence a measurable map and hence $f_*\nu$ is a measure on X. Now $f_*\nu(X) = \nu\left(f^{-1}(X)\right) = \nu(\{g \in G \mid \alpha(g, x_0) \in X\}) = \nu(G) = 1$ so $f_*\nu$ is a Radon probability measure on X. Let $g \in G$, $A \subset X$ measurable then $f_*\nu(\alpha(g, A)) =$

$\nu\left(f^{-1}\left(\alpha\left(g,A\right)\right)\right)=\nu\left(\left\{g'\in G\mid\alpha\left(g',x_0\right)\in\alpha\left(g,A\right)\right\}\right)=\nu\left(\left\{g'\in G\mid\alpha\left(g^{-1}g',x_0\right)\in A\right\}\right)=\nu\left(\left\{g\hat{g}\in G\mid\alpha\left(\hat{g},x_0\right)\in A\right\}\right)=\nu\left(g\left\{\hat{g}\in G\mid\alpha\left(\hat{g},x_0\right)\in A\right\}\right)=\nu\left(\left\{\hat{g}\in G\mid\alpha\left(\hat{g},x_0\right)\in A\right\}\right)=\nu\left(f^{-1}\left(A\right)\right)=f_*\nu\left(A\right)$. We conclude that $f_*\nu$ is a G-invariant probability Radon measure on X, and from uniqueness it is the only one. ∎

H.3. Translative Coverings

We follow Naszódi (2016) and define the following

Definition 59 *Let X be a set, $A\subseteq X$, $\mathcal{F}$ a collection of subsets of X i.e $\mathcal{F}\subseteq\mathcal{P}(X)$. A covering of A by $\mathcal{F}$ is a subset of $\mathcal{F}$ whose union contains A. We define the covering number of A by $\mathcal{F}$ to be the minimal cardinality of its coverings by $\mathcal{F}$:*

$$\tau(A,\mathcal{F})=\min\left\{|\mathcal{F}'|\mid\mathcal{F}'\subseteq\mathcal{F},A\subseteq\bigcup_{F\in\mathcal{F}'}F\right\}\,.$$

If X is a transitive G-space, $A,B\subseteq X$ some sets, we can look at the covering of A by translations of B, i.e by the collection $\mathcal{G}_GB=\{g.B\mid g\in G\}$. The covering number of A by translations of B is therefore denoted by $\tau(A,\mathcal{G}_GB)$.

Definition 60 *Let X be a transitive G-space with a G-invariant measure μ. Let $A,B\subseteq X$, we define the covering density of A by (translations of) B to be*

$$\vartheta(A,\mathcal{G}_GB)=\mu(B)\tau(A,\mathcal{G}_GB)\,.$$

Theorem 61 *(Rolfes and Vallentin, 2018, Corollary 3.1) For every $k\in\mathbb{N}$, $0<\varphi<\pi/2$*

$$\vartheta\left(\mathbb{S}^k,\mathcal{G}_{O(k+1)}B^k_{arc}(\varphi)\right)\leq\inf_{1<\alpha}\left(1+\frac{1}{\alpha-1}\right)(k\ln(\alpha k)+1)\,.$$

Lemma 62 *For every $k\in\mathbb{N}$ and every $0<\varphi<\frac{\pi}{2}$*

$$\vartheta\left(\mathbb{S}^k,\mathcal{G}_{O(k+1)}B^k_{arc}\left(\varphi\right)\right)\leq 5k\ln\left(k+1\right)\,.$$

Proof [Proof of Lemma 62] One has $\inf_{1<\alpha}\left(1+\frac{1}{\alpha-1}\right)\left(k\ln(\alpha k)+1\right)\leq 5k\ln\left(k+1\right)$ and so the claim follows from Theorem 61. ∎

Lemma 63 *(Böröczky and Wintsche, 2003, Corollary 3.2(i)) For every $k\in\mathbb{N}$ and every $0<\varphi<\frac{\pi}{2}$*

$$\mu_k\left(B^k_{arc}\left(\varphi\right)\right)\geq\frac{\sin^k\varphi}{\sqrt{2\pi\left(k+1\right)}}\,.$$

Proof [Proof of Lemma 63] Denote by $|\cdot|$ the surface area, and by V_l the volume of $\mathbb{S}^{l-1}$ for any $l \in \mathbb{N}$. We have from (Böröczky and Wintsche, 2003, Lemma 3.1(i))

$$|B^k_{\text{arc}}(\varphi)| \geq V_k \sin^k \varphi \ ,$$

and so

$$\begin{aligned}
\mu_k\left(B^k_{\text{arc}}(\varphi)\right) &= \frac{|B^k_{\text{arc}}(\varphi)|}{|\mathbb{S}^{(k+1)-1}|} = \frac{|B^k_{\text{arc}}(\varphi)|}{(k+1)V_{k+1}} \\
&\geq \frac{V_k \sin^k \varphi}{(k+1)V_{k+1}} \\
&= \frac{\pi^{\frac{k}{2}} \sin^k \varphi \Gamma\left(\frac{k+1}{2}+1\right)}{(k+1)\pi^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}+1\right)} = \frac{\sin^k \varphi \Gamma\left(\frac{k+1}{2}+1\right)}{(k+1)\pi^{\frac{1}{2}}\Gamma\left(\frac{k+1}{2}+\frac{1}{2}\right)} \\
&> \frac{\sin^k \varphi}{(k+1)\pi^{\frac{1}{2}}}\sqrt{\frac{k+1}{2}} && \text{(Gautschi's inequality)} \\
&= \frac{\sin^k \varphi}{\sqrt{2\pi(k+1)}} \ .
\end{aligned}$$

■

H.4. Additional Lemmas

Lemma 64 *Let $p \in (1,\infty) \setminus \mathbb{N}$, $0 < \epsilon < 1$, and $\psi_p : \mathbb{R} \to \mathbb{R}$ defined by $\psi_p(x) = x^p$. Define $D = \lceil p(\pi\epsilon)^{-1/\lfloor p \rfloor} \rceil + 2$, and $P_{\epsilon,p} : \mathbb{R} \to \mathbb{R}$ by $P_{\epsilon,p}(x) = \sum_{i=0}^{D} \binom{p}{i}(x-1)^i$. Then $|P_{\epsilon,p}(x) - \psi_p(x)| < \epsilon$ for every $x \in [0,1]$.*

Proof [Proof of Lemma 64] If $x = 0$ then

$$\begin{aligned}
|P_{\epsilon,p}(x) - \psi_p(x)| = |P_{\epsilon,p}(0) - 0^p| = |P_{\epsilon,p}(0)| &= \left|\sum_{i=0}^{D}\binom{p}{i}(-1)^i\right| \\
&= \left|(-1)^D\binom{p-1}{D}\right| = \left|\binom{p-1}{D}\right| \\
&\leq \frac{1}{\pi}\left(\frac{p}{D}\right)^{\lfloor p \rfloor} < \frac{1}{\pi}\left(\frac{p}{p(\pi\epsilon)^{-1/\lfloor p \rfloor}}\right)^{\lfloor p \rfloor} && \text{(Lemma 67)} \\
&= \frac{1}{\pi}\left((\pi\epsilon)^{1/\lfloor p \rfloor}\right)^{\lfloor p \rfloor} = \epsilon
\end{aligned}$$

Let $x \in (0,1]$. From Lemma 65, we have $\psi_p(x) = \sum_{i=0}^{\infty}\binom{p}{i}(x-1)^i$ with a remainder $R_D(x) = \psi_p(x) - P_{\epsilon,p}(x)$. Now, ψ_p is $(D+1)$-differentiable on $(\frac{x}{2}, x)$ and $\psi_p^{(D+1)}$ is continuous on $[\frac{x}{2}, x]$

and so by the mean-value form of the remainder, there exists some $c \in [\frac{x}{2}, x]$ such that $R_D(x) = \frac{\psi_p^{(D+1)}(c)}{(D+1)!}(x - x/2)^{D+1}$. Therefore

$$\begin{aligned}
|P_{\epsilon,p}(x) - \psi_p(x)| = |R_D(x)| &= \left| \frac{\psi_p^{(D+1)}(c)}{(D+1)!}(x - x/2)^{D+1} \right| \\
&= \left| \frac{p(p-1)\cdots(p-D)}{(D+1)!} c^{p-D-1}(x/2)^{D+1} \right| \\
&= \left| \binom{p}{D+1} c^{p-D-1}(x/2)^{D+1} \right| \\
&= \left| \binom{p}{D+1} \right| c^{p-D-1}(x/2)^{D+1}
\end{aligned}$$

Note that $p < D$ and so $c \mapsto c^{p-D-1}$ is decreasing on $[x/2, x]$ and so

$$\begin{aligned}
\left| \binom{p}{D+1} \right| c^{p-D-1}(x/2)^{D+1} &\le \left| \binom{p}{D+1} \right| (x/2)^{p-D-1}(x/2)^{D+1} \\
&= \left| \binom{p}{D+1} \right| (x/2)^{p} \le \left| \binom{p}{D+1} \right| \\
&\le \frac{1}{\pi} \left(\frac{p}{D} \right)^{\lfloor p \rfloor} < \frac{1}{\pi} \left(\frac{p}{p(\pi\epsilon)^{-1/\lfloor p \rfloor}} \right)^{\lfloor p \rfloor} \qquad \text{(Lemma 67)} \\
&= \frac{1}{\pi} \left((\pi\epsilon)^{1/\lfloor p \rfloor} \right)^{\lfloor p \rfloor} = \epsilon
\end{aligned}$$

■

Lemma 65 *Let $p \in (1, \infty)$, $x \in [0, 2]$ then the binomial series $\sum_{i=0}^{\infty} \binom{p}{i}(x-1)^i$ converges absolutely to x^p.*

Proof [Proof of Lemma 65] Let $x \in [0, 2]$ and denote $u = x - 1$, then $u \in [-1, 1]$ and so by Lemma 66 the series $\sum_{i=0}^{\infty} \binom{p}{i} u^i$ converges absolutely to $(1+u)^p$. We conclude that series $\sum_{i=0}^{\infty} \binom{p}{i}(x-1)^i$ converges absolutely to x^p. ■

Lemma 66 *(Knopp, 1964, Theorem 247(a)) Let $p \in (1, \infty)$, $|x| \le 1$ then the binomial series $\sum_{i=0}^{\infty} \binom{p}{i} x^i$ converges absolutely to $(1 + x)^p$.*

Lemma 67 *Let $D \in \mathbb{N}$ and any real $1 < p < D$ such that $p \notin \mathbb{N}$. Then*

$$\left| \binom{p}{D+1} \right| \le \left| \binom{p-1}{D} \right| \le \frac{1}{\pi} \left(\frac{p}{D} \right)^{\lfloor p \rfloor}.$$

Proof [Proof of Lemma 67] For the first inequality note that because $p < D$ we have

$$\left|\binom{p}{D+1}\right| = \left|\frac{\Gamma(p+1)}{\Gamma(D+2)\Gamma(p-D)}\right| = \left|\frac{p\Gamma(p)}{(D+1)\Gamma(D+1)\Gamma(p-D)}\right|$$
$$= \frac{p}{D+1}\left|\binom{p-1}{D}\right| \leq \left|\binom{p-1}{D}\right| .$$

Now $p - D \notin \mathbb{Z}$ so

$$\left|\frac{1}{\Gamma(p-D)}\right| = \left|\frac{\sin(\pi(p-D))\Gamma(1+D-p)}{\pi}\right| \leq \frac{1}{\pi}\left|\Gamma(1+D-p)\right| ,$$

and therefore

$$\left|\binom{p-1}{D}\right| = \left|\frac{\Gamma(p)}{\Gamma(D+1)\Gamma(p-D)}\right| \leq \frac{1}{\pi}\left|\Gamma(p)\frac{\Gamma(D-p+1)}{\Gamma(D+1)}\right| .$$

Note that Γ is increasing on $[1.5, \infty)$ and we have $D - \lfloor p \rfloor + 1 \geq 2$, $D - \lfloor p \rfloor + 1 \geq D - p + 1 \geq 1$, $\Gamma(1) = \Gamma(2) = 1$ and so $\Gamma(D - \lfloor p \rfloor + 1) \geq \Gamma(D - p + 1)$. Hence

$$\Gamma(D+1) = \prod_{i=0}^{\lfloor p \rfloor - 1}(D-i) \cdot \Gamma(D - \lfloor p \rfloor + 1) \geq \prod_{i=0}^{\lfloor p \rfloor - 1}(D-i) \cdot \Gamma(D-p+1) .$$

After rearranging we get $\frac{\Gamma(1+D-p)}{\Gamma(D+1)} \leq \frac{1}{\prod_{i=0}^{\lfloor p \rfloor - 1}(D-i)}$. Furthermore, since $0 < \Gamma(p - \lfloor p \rfloor + 1) \leq 1$ we have

$$\Gamma(p) = \prod_{i=1}^{\lfloor p \rfloor - 1}(p-i) \cdot \Gamma(p - \lfloor p \rfloor + 1) \leq \prod_{i=1}^{\lfloor p \rfloor - 1}(p-i) \leq \prod_{i=0}^{\lfloor p \rfloor - 1}(p-i) .$$

Together we obtain

$$\frac{1}{\pi}\left|\Gamma(p)\frac{\Gamma(D-p+1)}{\Gamma(D+1)}\right| \leq \frac{1}{\pi}\prod_{i=0}^{\lfloor p \rfloor - 1}\frac{p-i}{D-i} \leq \frac{1}{\pi}\prod_{i=0}^{\lfloor p \rfloor - 1}\frac{p}{D} = \frac{1}{\pi}\left(\frac{p}{D}\right)^{\lfloor p \rfloor} .$$

■

Lemma 68 *For every* $0 < \alpha < \beta \leq \infty$ *and every* $v \in \mathbb{R}^d$ *one has*

$$\|v\|_\beta \leq \|v\|_\alpha \leq d^{\frac{1}{\alpha} - \frac{1}{\beta}}\|v\|_\beta ,$$

where we define $\frac{1}{\infty} = 0$.

Proof [Proof of Lemma 68] We prove the two inequalities separately.

- $\|v\|_\beta \leq \|v\|_\alpha$:

– If $\beta = \infty$, then

$$\|v\|_\beta = |v_{\max}| = (|v_{\max}|^\alpha)^{\frac{1}{\alpha}} \le \left(\sum_{i=1}^{d} |v_i|^\alpha\right)^{\frac{1}{\alpha}} \le \|v\|_\alpha$$

– If $\beta < \infty$, then for every $1 \le i \le d$ we have $\frac{|v_i|}{\|v\|_\beta} \le 1$ and so $\left(\frac{|v_i|}{\|v\|_\beta}\right)^\beta \le \left(\frac{|v_i|}{\|v\|_\beta}\right)^\alpha$. Therefore,

$$\frac{\|v\|_\alpha}{\|v\|_\beta} = \left(\sum_{i=1}^{d} \left(\frac{|v_i|}{\|v\|_\beta}\right)^\alpha\right)^{\frac{1}{\alpha}} \ge \left(\sum_{i=1}^{d} \left(\frac{|v_i|}{\|v\|_\beta}\right)^\beta\right)^{\frac{1}{\alpha}} = \frac{\|v\|_\beta^{\frac{\beta}{\alpha}}}{\|v\|_\beta^{\frac{\beta}{\alpha}}} = 1 \, .$$

- $\|v\|_\alpha \le d^{\frac{1}{\alpha} - \frac{1}{\beta}} \|v\|_\beta$:
  – If $\beta = \infty$, then

$$\|v\|_\alpha = \left(\sum_{i=1}^{d} |v_i|^\alpha\right)^{\frac{1}{\alpha}} \le \left(\sum_{i=1}^{d} |v_{\max}|^\alpha\right)^{\frac{1}{\alpha}} = d^{\frac{1}{\alpha}} \|v\|_\infty = d^{\frac{1}{\alpha} - \frac{1}{\beta}} \|v\|_\beta \, .$$

  – If $\beta < \infty$, denote $r_1 = \frac{\beta}{\alpha} > 1$ and $r_2 = \frac{r_1}{r_1 - 1}$ then $1 < r_1, r_2$ and $\frac{1}{r_1} + \frac{1}{r_2} = 1$ hence by Holder's inequality

$$\sum_{i=1}^{d} (|v_i|^\alpha) \cdot 1 \le \left(\sum_{i=1}^{d} (|v_i|^\alpha)^{r_1}\right)^{\frac{1}{r_1}} \left(\sum_{i=1}^{d} (1)^{r_2}\right)^{\frac{1}{r_2}} = \left(\sum_{i=1}^{d} |v_i|^\beta\right)^{\frac{\alpha}{\beta}} d^{1 - \frac{\alpha}{\beta}} .$$

  Therefore, $\|v\|_\alpha \le \left(\sum_{i=1}^{d} |v_i|^\beta\right)^{\frac{1}{\beta}} d^{\frac{1}{\alpha}\left(1 - \frac{\alpha}{\beta}\right)} = d^{\frac{1}{\alpha} - \frac{1}{\beta}} \|v\|_\beta$

■

Lemma 69 *For every $p \in (0, \infty]$, $q \in [1, \infty]$ we denote $c^+_{p,q}(d) = d^{\left[\frac{1}{q} - \frac{1}{p}\right]_+}$, $c^-_{p,q}(d) = d^{\left[\frac{1}{q} - \frac{1}{p}\right]_-}$. Then for every $v \in \mathbb{R}^d$ one has*

$$c^-_{p,q}(d) \|v\|_p \le \|v\|_q \le c^+_{p,q}(d) \|v\|_p \, ,$$

where we define $\frac{1}{\infty} = 0$. Furthermore, when $q < p$ the upper bound is an equality for $v = (d^{-1/p}, \cdots, d^{-1/p})$ and the lower bound is an equality for $v = e_1$. When $p < q$ the upper bound is an equality for $v = e_1$ and the lower bound is an equality for $v = (d^{-1/p}, \cdots, d^{-1/p})$.

Proof [Proof of Lemma 69] We consider the cases $p < q$ and $q < p$ (the case $p = q$ is trivial).

- If $q < p$, then by the definition of $c^-_{p,q}(d), c^+_{p,q}(d)$ and from Lemma 68 we have

$$c^-_{p,q}(d) \|v\|_p = \|v\|_p \le \|v\|_q \le d^{\frac{1}{q} - \frac{1}{p}} \|v\|_p = c^+_{p,q}(d) \|v\|_p \, .$$

  For $v = (d^{-1/p}, \cdots, d^{-1/p})$ we get $\|v\|_q = c^+_{p,q}(d) \|v\|_p$, and for $v = e_1$ we get $c^-_{p,q}(d) \|v\|_p = \|v\|_q$

- If $p < q$, then by the definition of $c^-_{p,q}(d), c^+_{p,q}(d)$ and the right inequality in Lemma 68 we have $\|v\|_p \leq d^{\frac{1}{p}-\frac{1}{q}}\|v\|_q$ so $c^-_{p,q}(d)\|v\|_p \leq \|v\|_q$, and from the left inequality in Lemma 68 we have $\|v\|_q \leq \|v\|_p = c^+_{p,q}(d)\|v\|_p$.

  For $v = e_1$ we get $\|v\|_q = c^+_{p,q}(d)\|v\|_p$, and for $v = (d^{-1/p}, \cdots, d^{-1/p})$ we get $c^-_{p,q}(d)\|v\|_p = \|v\|_q$

■

Lemma 70 *For any $d \in \mathbb{N}$, $0 < r$ and $p \in (0, \infty]$, $q \in [1, \infty]$ we have $B^d_p(r) \subseteq B^d_q(c^+_{p,q}(d)r)$, and $B^d_q(c^-_{p,q}(d)r) \subseteq B^d_p(r)$. Furthermore, any $\alpha < c^+_{p,q}(d)r$ does not satisfy the first inclusion and any $c^-_{p,q}(d)r < \beta$ does not satisfy the second inclusion.*

Proof [Proof of Lemma 70] Follows immediately from Lemma 69. ■

Lemma 71 *Let $P \in Gr_{d,k}$ be an orthogonal projection, $x \in \mathbb{R}^d$, $0 \leq r$ and $a' \in B^k_2(Px, r)$ (where B^k_2 here is a ball in $ImP \cong \mathbb{R}^k$, $ImP \subset \mathbb{R}^d$), then there exists some $a \in B^d_2(x, r)$ such that $Pa = a'$.*

Proof [Proof of Lemma 71] Define $a = a' + x - Px$, then $\|a - x\|_2 = \|a' + x - Px - x\|_2 = \|a' - Px\|_2 \leq r$ and so $a \in B^d_2(x, r)$. Finally, $P(a) = P(a' + x - Px) = Pa' + Px - PPx = a'$. ■

Lemma 72 *$\forall g \in O(d), x \in \mathbb{R}^d$, and every subspace $W \in Gr_{d,k}$ we have $\|P_{gW}(x)\|_2 = \|P_W(g^{-1}x)\|_2$.*

Proof [Proof of 72] By definition, if $W = Sp\{w_1, ..., w_k\}$ for some orthonormal basis, then $P_W = A_W A_W^\top$ where A_W has $w_1, ..., w_k$ as column vectors, and $gW = Sp\{gw_1, ..., gw_k\}$ and so $A_{gW} = gA_W$ which means that $P_{gW} = gP_W g^\top = gP_W g^{-1}$. Therefore $\|P_{gW}(x)\|_2 = \|gP_W g^{-1}(x)\|_2 = \|gP_W(g^{-1}x)\|_2 = \left\|g\left(P_W(g^{-1}x)\right)\right\|_2 = \|P_W(g^{-1}x)\|_2$, where the last equality follows from the fact that the orthogonal group consists of endomorphisms that preserve the Euclidean norm. ■

Lemma 73 *Let X be a G-space, and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be a cover of X. Then for every $g \in G$, $g\mathcal{U} = \{gU_\alpha\}_{\alpha \in A}$ is a cover of X.*

Proof [Proof of Lemma 73] Let $g \in G$, and let $x \in X$. Denote $x' = g^{-1}x$, then there exists some $\alpha \in A$ such that $x' \in U_\alpha$, so $x = gx' \in gU_\alpha$ and we are done. ■

Lemma 74 *Let $A \subseteq \mathbb{R}^d$ be some subset of $\mathbb{R}^d$ such that $\mathbb{R}A \in Gr_{d,l_1}$ for some $1 \leq l_1 \leq d$. Let $V \in Gr_{d,l_2}$ with $l_2 \geq d - l_1 + 1$, then there exists some $0 \neq x \in V \cap \mathbb{R}A$.*

Proof [Proof of Lemma 74] Note that $\mathbb{R}A$ is a vector subspace, hence by Grassmann's Identity we have

$$dim\,(V \cap \mathbb{R}A) = dimV + dim\mathbb{R}A - dim\,(V + \mathbb{R}A) \geq$$
$$dimV + dim\mathbb{R}A - d = l_2 + l_1 - d \geq d - l_1 + 1 + l_1 - d = 1\,.$$

■

Lemma 75 *For every $k \in \mathbb{N}$*

$$\left(5k \ln{(k+1)} \sqrt{2\pi\,(k+1)}\right)^{\frac{1}{k}} \leq 10\sqrt{\pi}\ln 2\,.$$

Proof [Proof of Lemma 75] Denote $c = \ln(5) + \frac{\ln(2\pi)}{2}$ then $\frac{c}{k}$ and $\frac{\ln(k+1)}{2k}$ both decrease for every $1 \leq k$, and $\frac{\ln k}{k}$ decreases for every $e \leq k$. Furthermore, $\frac{\ln(\ln(k+1))}{k}$ decreases for every $4.14 \leq k$, and so we conclude that $\frac{1}{k}\left(c + \ln k + \ln(\ln(k+1)) + \frac{\ln(k+1)}{2}\right)$ decreases for every $4.14 \leq k$. Computing for $k = 1, 2, 3, 4, 5$ we get that $\frac{1}{k}\left(c + \ln k + \ln(\ln(k+1)) + \frac{\ln(k+1)}{2}\right)$ decreases for every $k \in \mathbb{N}$. Hence, for every $k \in \mathbb{N}$ we have:

$$\begin{aligned}\frac{1}{k}\left(c + \ln k + \ln(\ln(k+1)) + \frac{\ln(k+1)}{2}\right) &\leq \frac{1}{1}\left(c + \ln 1 + \ln(\ln(1+1)) + \frac{\ln(1+1)}{2}\right)\\ &= \ln 5 + \frac{1}{2}\ln(2\pi) + \ln\ln 2 + \frac{1}{2}\ln 2\,.\end{aligned}$$

Since $x \mapsto \exp(x)$ increases monotonically we get for all $k \in \mathbb{N}$:

$$\begin{aligned}\left(5k\ln{(k+1)}\sqrt{2\pi\,(k+1)}\right)^{\frac{1}{k}} &= \exp\left[\ln\left(5k\ln{(k+1)}\sqrt{2\pi\,(k+1)}\right)^{\frac{1}{k}}\right]\\ &= \exp\left[\frac{1}{k}\left(c + \ln k + \ln(\ln(k+1)) + \frac{\ln(k+1)}{2}\right)\right]\\ &\leq \exp\left[\ln 5 + \frac{1}{2}\ln(2\pi) + \ln\ln 2 + \frac{1}{2}\ln 2\right]\\ &= 5\sqrt{2\pi}\ln 2 \cdot \sqrt{2} = 10\sqrt{\pi}\ln 2\,.\end{aligned}$$

■