# Rank Lifting and Random Non-Linear Maps

**Andrea Drago**[1] **Maria Sofia Bucarelli**[1,2,3,4] **Francesco Caso**[1,5] **Marius Michetti**[1,6]

**Federico Siciliano**[1] **Fabrizio Silvestri**[1] **Luca Becchetti**[1]

[1]Sapienza University of Rome, [2]CNRS, [3]Inria, [4]i3s, [5]University of Cambridge, [6]ENS Paris-Saclay

## Abstract

Deep neural networks exhibit improved training and generalization performance as the number of parameters grows well beyond the size of the training set, contradicting classical intuitions about overfitting. In order to gain a better understanding of this "benign overparameterization", we analyze the representational capacity of a random one-hidden-layer perceptron with Gaussian weights, no bias and threshold activations. More precisely, we investigate the following question: when does a hidden layer of dimension $n$ maps $k$ input vectors with pairwise angles at least $\theta$, to a full-rank activation matrix, thus ensuring that a simple linear classifier can perfectly fit those inputs in feature space? This problem has an immediate impact on memorization capacity at initialization and we frame it as a question about hyperplane arrangements on the unit sphere, and we prove new isoperimetric-like inequalities. This allows us to derive non-trivial lower bounds on the probability that a random embedding avoids the arrangement's zero-measure regions. Our results show that once the hidden dimension exceeds a threshold (depending on $\theta$ and the input dimension), hidden representations are linearly independent with high probability. While the case we consider is challenging due to the sparsity of the solution space, this setting highlights crucial, underlying geometric problems and connections to related questions in spherical geometry and linear algebra.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. 

## 1 INTRODUCTION

Overparameterization is a hallmark of modern machine learning. Neural networks trained with more parameters than data points not only fit the training set perfectly, but often generalize surprisingly well (Zhang et al., 2021).

To better understand their effectiveness, it is useful to ask which structural properties of neural networks are already present at random initialization. In this work we focus on one such property: the rank of the hidden representation at initialization. Specifically, we ask under which conditions a single random hidden layer maps a set of $k$ well-separated input points into linearly independent feature vectors. Having linearly independent feature embeddings at initialization is a strong form of separation capacity: it guarantees that any labeling of the inputs can be interpolated by a linear classifier in feature space, hence the training set can be shattered regardless of the distribution of the labels.

This phenomenon is closely tied to the ability of overparameterized models to interpolate data (Zhang et al., 2021). Studying this property directly at random initialization is natural for several reasons. For example, recent work has shown that highly performing subnetworks ("strong lottery tickets") already exist within untrained, randomly initialized models (Frankle and Carbin, 2018; Ramanujan et al., 2020), suggesting that expressivity can be present before training. Moreover, randomly initialized, untrained neural networks have been successfully used in other two ways: first, as feature maps that enable accurate downstream linear classifiers (Rahimi and Recht, 2008; Huang et al., 2006); and second, as initial random embeddings that feeds into further architectures: this has been investigated in the context of RBF networks (Igelnik and Pao, 1995), feedforward neural networks (Schmidt et al., 1992), extreme learning machines (Huang et al., 2006), and reservoir computing

(Lukoševičius and Jaeger, 2009; Jaeger, 2001).

This highlights the importance of understanding which structural properties of randomly initialized networks already provide expressive feature representations prior to training.

The study of rank properties of randomly initialized feature maps thus offers a potentially fruitful perspective on several aspects of "benign overparameterization" observed in practice.

It is worth noticing that the wide limit of many network architectures is well understood in terms of neural tangent kernels (NTK) (Jacot et al., 2020; Domingos, 2020; Arora et al., 2019). This connection shows how training in the limit becomes a convex optimization problem, and gives some understanding on the generalization capacity. On the other hand it is still not clear what happens outside of the NTK regime (Xiao et al., 2020; Chizat et al., 2020), i.e. when, like in our case, the number of parameters is bounded from above, nor what happens when hidden parameters are initialized randomly and fixed during training. In the latter case, it is known that networks with random gaussian features converge to gaussian processes (Neal, 1996; Lee, 2013; Basteri and Trevisan, 2023), a.k.a. NNGP, as their dimension grows, and their behaviours can be studied at the asymptotic or finite scale (Bowman and Montufar, 2022; Lillo et al., 2025). Nonetheless, feature embeddings are not independently distributed, and their linear independence is equivalent to the positive definiteness of their Gram matrix. To our knowledge this is an open problem to which we answer positively in our setting. Unlike NTK/NNGP, which capture asymptotic behavior, we give finite-width guarantees showing that random hidden layers can already produce linearly independent embeddings, revealing expressivity before training or infinite limits.

**Our contributions.** We analyze shallow networks with Gaussian weights and sign activations, and establish conditions under which the hidden representation of a set of input points achieves full rank with non-negligible probability. Note that this immediately implies that with the same probability, the class of hyperplanes shatters the input points in feature space. Our results can be summarized as follows:

- For $d = 2$ and $d = 3$, we prove high-probability guarantees: the required hidden dimension grows only linearly (resp. polynomially) with the number of inputs.
- In general dimension, we reduce the problem to estimating volumes of regions in a conical tessellation of the sphere, obtaining explicit lower bounds on the probability of having a full rank embedding matrix.
- Beyond the immediate probabilistic bounds, our geometric formulation highlights structural connections between random neural networks, spherical geometry, and problems of convexity and tessellation.
- Experimental results on both real and synthetic data, while qualitatively consistent with our theory, suggest that the theoretical analysis is not tight and that a deeper understanding of the underlying mechanisms is required to close the gap.

Our results isolate a minimal geometric mechanism, the independence of pre-activation weights, that helps explain the ability of overparameterized networks to interpolate arbitrary training labels. In particular, our focus on the sign activation enables a clear connection between the problem we study and the geometry of conical tessellation in input space. At the same time, intuitively it represents a worst case scenario, given its binary codomain, as we elaborate further upon in Section 6. This perspective complements existing work on separation margins (Dirksen et al., 2022; Ghosal et al., 2022), while being strictly stronger, since full rank implies shattering of the input set. Moreover, the geometric tools we develop highlight both obvious and less obvious connections to a number of related areas, including Gaussian processes, spherical geometry and the algebra of covariance matrices. We briefly elaborate on some of these connections in Appendix D.

## 2 RELATED WORK

The approximation capabilities of neural network models have been the focus of extensive research over the past four decades, beginning with the pioneering works of (Cybenko, 1989) and Hornik (1991).

The problem we are considering is a special case of the accuracy-related problem of interpolation, i.e., the ability of (large) models, with more parameters than training points, to fit arbitrary functions of the input, in particular arbitrary dichotomies over the input points.

Baum (1988) was the first to show that $\lceil \frac{k}{d} \rceil$ hidden neurons are sufficient to interpolate $k$ points in general positions, i.e., with no more than $d$ points belonging the the same hyperplane, using simple threshold activations. Bubeck et al. (2020) extended this results to the case of ReLU activations with arbitrary real labels.

In general, it is known that, as soon as $n \geq k$, it is possible to choose the weights of a hidden layer $W$, so

that its output $Y = f(X)$ has rank $k$, where $X \in \mathbb{R}^{d\times k}$ is the matrix of $k$ input vectors in dimension $d$ (see for example (Sartori and Antsaklis, 1991; Huang and Babri, 1998; Pinkus, 1999)). This means that any downstream hyperplane-based classifier, such as (Support vector machine ) SVM or even the perceptron algorithm, will shatter the points $Y$ and thus correctly classify those in $X$, however labelled they are.

It is important to note that ability to interpolate depends on the total number of parameters rather than width of the network. In particular, a number of previous works propose deep architectures that can achieve interpolation with a total number of parameters that is linear in the number $k$ of input points using $\mathcal{O}(\sqrt{k})$ neurons (see for example (Vardi et al., 2021; Rajput et al., 2021)) when the points satisfy mild separability conditions, while $\mathcal{O}(kd)$ parameters suffice when points are assumed to be distinct (Huang and Huang, 1991) and biases are used. All these (mostly recent) contributions provide constructive proofs. On the other hand, it is well-known that $(k-1)/2$ parameters are necessary to shatter a sample of $k$ distinct points (Sontag, 1997), a result that has been refined over the recent past, to account for its dependence on the minimum separation between points of the sample (see for example (Siegel, 2023)).

One should also note that the results above simply imply that, if the number of parameters is large enough, for every training set of size $k$ with given labels, there exists a setting of the parameters that will fit the desired labelling with zero error. The question we are interested in is different: if $f$ denotes a random (non-linear) map from $\mathbb{R}^d$ to $\mathbb{R}^n$ implemented by one (or more) randomly initialized hidden layers, what is the probability that $f$ shatters a set $\mathcal{X} \subset \mathbb{R}^d$ of points in feature space?

While the idea of using randomly initialized shallow networks as pre-processing maps for downstream application of well-understood classification models in feature space is not new (Huang et al., 2006), related rigorous results are relatively recent. In Dirksen et al. (2022), the authors consider the ability of a randomly initialized, two-layer ReLU network, to separate two, possibly infinite, sets of points with a minimum margin in feature space, under the assumption that any two points from different sets are at (euclidean) distance at least $\delta$. These results have been extended in Ghosal et al. (2022) to the case of a single, randomly initialized hidden layer, improving dependence on the number of dimensions. The problem studied in Dirksen et al. (2022); Ghosal et al. (2022) is clearly related to the one considered here, albeit with some important differences. On the one hand, the problem we consider is fundamentally harder than the ones considered in Dirksen et al. (2022); Ghosal et al. (2022). In particular, achieving full rank implies shattering the input points and being able to interpolate any possible dichotomy over the dataset, something way stronger than achieving separability of a dataset according to one fixed dichotomy, which is the problem studied there. On the other hand, differently from Dirksen et al. (2022) we generally assume bias is zero, consistently with many initialization schemes used in practice, while our analysis becomes somewhat simpler if a uniform bias is introduced. Finally, the approaches are technically very different. For example, Dirksen et al. (2022) investigates the random tessellations of $S^{d-1}$ induced by the random, affine hyperplane associated with the weights and bias of every neuron, whereas we study the distribution of vectors chosen uniformly on $S^{d-1}$ on the regions defined by the tessellation of $S^{d-1}$ induced by the input points themselves, which we think is a very natural way to look at the problem we consider. Another related work is the seminal paper Cover (1965), which analyzes the *deterministic capacity of decision surfaces*, showing how many dichotomies can be realized under a *general position* assumption on inputs. Our work studies the *probabilistic capacity of random neural networks*, assuming $\theta$-separated inputs and asking when random features yield *full-rank representations*. In essence, Cover (1965) counts what can be separated in principle, while our work measures how likely random networks are to achieve separation.

## 3 PRELIMINARIES AND PROBLEM

In this section, we introduce the problem we study, some important definitions and concepts, along with the notation that will be used throughout this work.

**Problem Definition.** We investigate the rank of the matrix formed by the hidden representations produced by a randomly initialized, shallow feed-forward network.

In the remainder, vectors are columns. If $M$ is a matrix and $f$ a function, $f(M)$ means that $f$ is applied to $M$ entry-wise. We use $I_d$ to denote the identity matrix of dimension $d$.

Consider a single hidden layer with $n$ neurons and the sign activation function. We assume pre-activation is described by a matrix $W \in \mathbb{R}^{n\times d}$, where $d$ is the dimension of the input space. The hidden layer is a non-linear map

$$f : \mathbb{R}^d \to \mathbb{R}^n.$$

Assume that $W_{ij} \sim \mathcal{N}(0, 1)$ for every $i, j$ and all entries

are independent. We consider a $\theta$-separated family of $k$ samples $\mathrm{x}_1, \ldots, \mathrm{x}_k$ from some input space $\mathcal{X} \subseteq \mathbb{R}^d$. I.e., for every $i \neq j$, the angle $\theta_{ij} = \sphericalangle(\mathrm{x}_i, \mathrm{x}_j)$ satisfies $\theta_{ij} \in [\theta, \pi - \theta]$. Let $X = [\mathrm{x}_1, \ldots, \mathrm{x}_k] \in \mathbb{R}^{d \times k}$, be the matrix whose columns are the vectors $\mathrm{x}_i \in \mathbb{R}^d$. The output of the input layer applied to $X$ is Y:

$$Y = f(X) = \mathrm{sign}(WX) \in \{\pm 1\}^{n \times k}.$$

We assume $d < k < n$ to make the math relevant.

The main questions we address in this study are the following: *What is the probability that $Y$ has rank $k$? How does this probability depend on the value $n \geq k$ of the number of neurons? How does it depend on properties of the input, in particular, the dimension $d$ of the input space and the minimum angle between input points?*

We note that when points are collinear, full rank cannot be achieved, since the images $f(\mathrm{x}_1)$ and $f(\mathrm{x}_2)$ of any two collinear points are themselves collinear. We elaborate briefly on this in Section 4.3.

Our analysis turns algebraic questions about rank to geometric questions on the sphere; in the following we introduce some of the fundamental concepts we will use.

**Hyperplane arrangements.** A *hyperplane arrangement in* $\mathbb{R}^m$ is a collection of hyperplanes $\mathcal{A} = \{H_i \subset \mathbb{R}^m\}$, the corresponding *regions* of $\mathcal{A}$ are the connected components of $\mathbb{R}^m \setminus \bigcup_i H_i$. If we fix a subspace $V \subset \mathbb{R}^m$, the arrangement $\mathcal{A}$ induces by intersection an arrangement $\mathcal{A} \cap V = \{H_i \cap V\}$ in $V$.
Let $\{H_{i_1}, \ldots, H_{i_l}\} \subset \mathcal{A}$ be $l$ distinct hyperplanes of the arrangement $\mathcal{A}$, and let us denote $\mathcal{A}_l \coloneqq \mathcal{A} \setminus \{H_{i_1}, \ldots, H_{i_l}\}$ and $V_l \coloneqq H_{i_1} \cap \cdots \cap H_{i_l}$. A *face* (of codimension $l$) of $\mathcal{A}$ is a region of $\mathcal{A}_l \cap V_l$.
A hyperplane arrangement is *generic* if every $l$-tuple of hyperplanes intersect at a different subspace of codimension $l$, and *central* if every hyperplane contains the origin. A central hyperplane arrangement is *generic* if every $l$-tuple of hyperplanes, with $l < m$, intersects at a different subspace of codimension $l$, and *essential* if the intersection of all hyperplanes is exactly the origin.
A hyperplane arrangement is *oriented* if it is endowed with the choice of an orientation for each hyperplane $H_i \in \mathcal{A}$, i.e. an orthonormal vector $h_i \perp H_i$. Let $H_1, H_2$ be two oriented hyperplanes with perpendicular vectors $h_i \perp H_i$, then $\sphericalangle(H_1, H_2) = \sphericalangle(\mathrm{h}_1, \mathrm{h}_2)$, so consistently we say that a hyperplane arrangement is *$\theta$-separated* if their perpendicular vectors are $\theta$-separated. The interested reader can find additional details in Appendix A.

**Conical tessellations.** Let $S^m \subset \mathbb{R}^{m+1}$ be the unit sphere of dimension $m$, with its usual Riemannian structure (cf. Appendix A.2). We denote $\omega_m \coloneqq |S^m|$ is total measure. A central hyperplane arrangement $\mathcal{A} = \{H_i \subset \mathbb{R}^{m+1}\}$ defines by intersection a family of diameters $\mathcal{D} = \{S^m \cap H_i\}$ of $S^m$, called a *conical tessellation of $S^m$*. The *regions* of a conical tessellation are the closed connected components of the complement $\mathcal{D}^c$ in $S^m$. It is easy to check that each region of a conical tessellation is a convex spherical polytope.
All the terminology for hyperplane arrangements extends to the conical tessellation in the same way: the faces of $\mathcal{D}$ are defined analogously by intersection, is defined in the same way; if $D \subset S^m$ is a diameter of $S^m$, $\mathcal{D}$ induces by intersection a conical tessellation $\mathcal{D} \cap D$ on it; and $\mathcal{D}$ is *$\theta$-separated* if the corresponding hyperplane arrangement is $\theta$-separated. In our formulation of the problem, each region will correspond to a consistent sign pattern of inner products with the input vectors. In dimension 2, for instance, $k$ input vectors define $2k$ arcs on the circle; in dimension 3, they cut the sphere into a collection of spherical polythopes. In higher dimension the picture is harder to visualize, but the key idea persists: input vectors carve the hypersphere into regions.

## 4 RANK LIFTING

As stated before, our goal is to determine when the hidden representation matrix

$$Y = \mathrm{sign}(WX) \in \{\pm 1\}^{n \times k}$$

attains full rank. If $Y$ is not full rank, we denote

$$S = \mathrm{span}(\mathrm{y}_1^T, \ldots, \mathrm{y}_n^T) \subset \mathbb{R}^k,$$

and let $\mathrm{u} \perp S$ be any nonzero vector orthogonal to this subspace.

Each input vector $\{x_i\}_{i=1}^k$ defines an oriented hyperplane $H_i = \{y \mid \langle x_i, y \rangle = 0\}$ and a $\theta$-separated family of vectors $x_i$ induces a $\theta$-separated arrangement $\mathcal{A} = \{H_i\}$. Intersecting $\mathcal{A}$ with the unit sphere $S^{d-1}$ yields a tessellation $\mathcal{C}$ where each region inherits the sign pattern. We define by z the map that makes this correspondence explicit: $\mathrm{z} : \mathbb{R}^d \to \mathbb{R}^k$ such that $\mathrm{z}(\mathrm{v})^T = \mathrm{sign}(\mathrm{v}^T X)$.

For each vector v, the output z(v) records on which side of each hyperplane $H_i$, v lies (positive side, negative side, or on the hyperplane if the product is zero).[1] z acts as a labeling function: it encodes the position of a vector v relative to the arrangement $\mathcal{A}$.

[1] We note that since we assume $\mathrm{v} \sim \mathcal{N}(\mathbf{0}, I_d)$, the event that the product is zero has probability measure zero.

Two vectors $\mathrm{v}_1, \mathrm{v}_2$, have the same value under z if and only if $\mathrm{v}_1$ and $\mathrm{v}_2$ belong to the same region of $\mathcal{C}$. This map is essentially a coordinate map that encodes which conical cell of the tessellation the vector belongs to.

**Incremental construction.** To prove Theorem 4.2, we analyze the process of increasing the rank of $Y$ incrementally, by adding new neurons and, correspondingly, new rows of the pre-activation in sequence, with the latter sampled independently from a standard normal distribution. Specifically, assume we add one more neuron with associated random weights $\mathrm{w} \sim \mathcal{N}(\mathbf{0}, I_d)$. Extending the weight matrix to $W' = [W^T|\mathrm{w}]^T$ produces an updated hidden representation

$$Y' = \mathrm{sign}(W'X),$$

whose last row is exactly $\mathrm{z}(\mathrm{w})^T$. By definition,

$$\mathrm{rank}(Y') > \mathrm{rank}(Y) \quad \iff \quad \mathrm{z}(\mathrm{w})^T \mathrm{u} \neq 0,$$

for some vector $\mathrm{u} \in \mathbb{R}^k$ orthogonal to $S$, the subspace spanned by $Y$'s rows. Thus, considered any $\mathrm{u} \perp S$, the probability that the rank increases by one is at least the probability that the new random row added to $Y$ is not orthogonal to u:

$$\mathbb{P}_{\mathrm{w}\sim\mathcal{N}^{\otimes d}}\left(\mathrm{rank}(Y') > \mathrm{rank}(Y)\right) \geq \mathbb{P}_{\mathrm{w}\sim\mathcal{N}^{\otimes d}}\left(\mathrm{z}(\mathrm{w})^T \mathrm{u} \neq 0\right).$$

The key technical result of our work is a quantitative estimate of the probability of the event above. The following lemma is the building block to prove our main theorem; it quantifies how likely rank lifting is at each incremental step.

**Lemma 4.1.** *Let $\mathrm{x}_1, \ldots, \mathrm{x}_k \in \mathbb{R}^d$ be $\theta$-separated. Let* w *be a vector with i.i.d. components $w_j \sim \mathcal{N}(0,1)$ and $\mathrm{z}(\mathrm{w}) = \mathrm{sign}(\mathrm{w}^T X)$. Then, for any vector $\mathrm{u} \perp S$ we have:*

$$\mathbb{P}\left(\mathrm{z}(\mathrm{w})^T \mathrm{u} \neq 0\right) > \frac{\theta}{2^k}\sqrt{\frac{2}{d\pi}}.$$

The previous lemma is the key technical ingredient of our work, and the proof will be given in Section 4.2. Given that, we now present and prove our main result.

**Theorem 4.2.** *Let $X = [\mathrm{x}_1, \ldots, \mathrm{x}_k] \in \mathbb{R}^{d\times k}$ and assume the angle between any two vectors $\mathrm{x}_i$ and $\mathrm{x}_j$ is at least some constant $\theta$. Let $Y = \mathrm{sign}(WX)$, with $W \in \mathbb{R}^{n\times d}$. Assume that $W_{ij} \sim \mathcal{N}(0,1)$ independently for every $i, j$. Then:*

$$\mathbb{P}\left(\mathrm{rank}(Y) = k\right) \geq 1 - e^{-(2-\ln 3)k},$$

*whenever $n \geq 3\frac{k}{\alpha}$ with $\alpha$ is the probability estimated in Lemma 4.1, i.e. $\alpha = \theta/2^k\sqrt{2/d\pi}$.*

*Proof.* We sequentially add neurons to the hidden layer (and the corresponding rows to the activation layer) until $\mathrm{rank}(Y) = k$, as described by the following algorithm:

```
Sample w ~ N(0, I_d); W = w^T;
Y = sign(WX); Z = 1;
while rank(Y) < k do
    Add new neuron;
    Sample w ~ N(0, I_d);
    W = [W^T|w]^T;
    Y = sign(WX);
    Z = Z + 1
end
```

**Algorithm 1:** Incremental Construction.

The value of $Z$ when Algorithm 1 terminates is the number of neurons we need to add to the hidden layer for $Y$ to achieve rank $k$. In the remainder of this proof, for $i = 1, \ldots, k$, we denote by $Z_i$ the number of new rows (and associated neurons) we need to sample, for the rank of $Y$ to increase from $i-1$ to $i$. Note that we have $Z_1 = 1$ deterministically and $Z = \sum_{i=1}^k Z_i$ by definition, so that $\mathbb{E}[Z] = \sum_{i=1}^k \mathbb{E}[Z_i]$. We next focus on $\mathbb{E}[Z_i]$, for $i > 1$. Assume $\mathrm{rank}(Y) = i-1$ and let $\alpha = \frac{\theta}{2^k}\sqrt{\frac{2}{d\pi}}$ as above. Then, considered any $\mathrm{u} \perp S$,[2] if we sample a new row w, we have $\mathrm{z}(\mathrm{w})^T \mathrm{u} \neq 0$ with probability at least $\alpha$ from Lemma 4.1, i.e., the matrix $Y'$ obtained by adding the new row $\mathrm{z}(\mathrm{w})^T$ to $Y$ satisfies $\mathrm{rank}(Y') = \mathrm{rank}(Y) + 1$ with probability at least $\alpha$. This immediately implies that $\mathbb{E}[Z_i] < \frac{1}{\alpha}$ and $\mathbb{E}[Z] < \frac{k}{\alpha}$.

More precisely, each $Z_i$ is stochastically dominated by a geometric random variable $G_i$ with parameter $p = \frac{1}{\alpha}$, in the sense that $\mathbb{P}(Z_i \geq x) \leq \mathbb{P}(G_i \geq x)$ for every $x$. Likewise $Z$ is stochastically dominated by $G = \sum_{i=1}^k G_i$. We therefore have, for every $\varepsilon > 0$:

$$\mathbb{P}\left(Z > 3\frac{k}{\alpha}\right) \leq \mathbb{P}\left(G > 3\frac{k}{\alpha}\right) \leq e^{-(3-\ln 2)k},$$

where the result follows since $G$ is the sum of identical, independent geometric random variables with parameter $p = \alpha$, so that its expectation is $\frac{k}{\alpha}$. We can thus apply Chernoff-like tail bounds for the sum of geometric random variables. In particular, (Janson, 2018, Thm. 2.1) yields the result above. This concludes the proof of Theorem 4.2. □

The rest of this section is devoted to the proof of Lemma 4.1. Particularly, in Section 4.1 we analyze

[2] In order to keep notation simple, we use $S$ to denote the span of $Y$'s rows during the current iteration, which is understood from context.

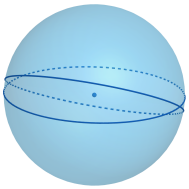

**Figure 1:** *A region has small area if an angle is small...*

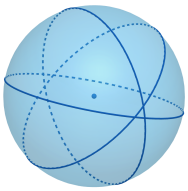

**Figure 2:** *...but a region can have small area also with big angles!*

the low dimensional case and referred the general case to Section 4.2.

### 4.1 Warm-up: Low dimensional cases

To build up the geometric intuition, in this section we show how to prove Lemma 4.1 in dimension 2 and 3. Let us denote $\mathcal{C}$ the conical tessellation associated to the samples, and $F_{\mathrm{u}} : S^d \to \mathbb{R}$ the function defined by $F_{\mathrm{u}}(\mathrm{w}) = \mathrm{z}(\mathrm{w})^T\mathrm{u}$, which is constant on the regions of $\mathcal{C}$. The main idea is that since $F_{\mathrm{u}}$ only depends on the direction of w, and w is isotropic, it suffices to bound from below the measure of the regions where $F_{\mathrm{u}} \neq 0$. *The key observation is that* $F_{\mathrm{u}}$ *attains strictly different values on adjacent regions of* $\mathcal{C}$, implying that, if $A$ and $B$ are adjacent regions in $\mathcal{C}$, then $\mathrm{z}(\mathrm{a})^T\mathrm{u} = 0$ for $\mathrm{a} \in A$ implies $\mathrm{z}(\mathrm{b})^T\mathrm{u} \neq 0$ for any $\mathrm{b} \in B$ (cf. the proof of Lemma 4.3 for a detailed discussion).

In dimension 2 the problem is straightforward. In fact, since the points $\mathrm{x}_i$ are $\theta$-separated, the associated conical tessellation of $S^1$ has exactly $2k$ regions of length at least $\theta$. Then there are at least $k$ regions, of total length at least $k\theta$, where $F_x \neq 0$. So

$$\mathrm{Prob}(F_x = 0) \leq \frac{2\pi - k\theta}{2\pi}, \quad \text{and} \quad \mathrm{Prob}(F_x \neq 0) \geq \frac{k\theta}{2\pi}.$$

In higher dimension a region can have arbitrarily small area also (cf. Fig 2), so we need a better strategy. We pick a diameter $D \in \mathcal{C}$ of the conical tessellation of $S^2$, and we look at all regions adjacent to it. The

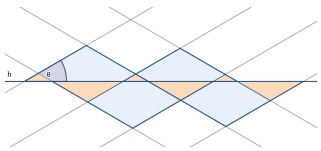

**Figure 3:** *The horizontal line depicts the diameter* $D$*. The other diameters intersect* $D$ *at angle* $\theta$*, and cut it in a finite number of regions. Some regions can be arbitrarily small, but collectively their measure is bounded from below. The same is true for their adjacent regions (in peach).*

diameter $D$ is cut into $N = 2(k-1)$ segments $s_j$ by all the other diameters of the tessellation, and each of the two regions $R_j^{\pm}$ adjacent to any segment contain a triangle $T_j^{\theta}$ with base $s_j$ and the angles between $s_j$ and the other two sides $\theta$. The area of these triangles is approximately[3] $s_j^2 \tan\theta/4$. It holds $F_x \neq 0$ at least on one region among $R_j^{\pm}$ for each $j$, so we deduce

$$\mathrm{Prob}(F_x \neq 0) \geq \min_{\sum s_j = 2\pi} \sum_j s_j^2 \frac{\tan\theta}{4} = \frac{\pi^2 \tan\theta}{2(k-1)}$$

where the last equality follows from Jensen's inequality and the fact that $D$ intersect each of the $k-1$ remaining diameters in exactly two antipodal points. In general we will follow exactly this approach, but more effort will be needed to estimate the measure of the regions adjacent to the faces $s_j$.

### 4.2 Arbitrary number of dimensions

To prove Lemma 4.1 in the general case, we proceed following the same geometric intuition as in the previous section, obtaining the following formulation.

**Lemma 4.3** (Geometric reformulation)**.** *Let* $\mathcal{C}$ *be the conical tessellation of* $S^{d-1}$ *associated to vectors* $\mathrm{x}_1, \dots, \mathrm{x}_k$*. There exists a conical tessellation* $\mathcal{C}'$ *associated to a subset of the vectors* $\mathrm{x}_1, \dots, \mathrm{x}_k$ *and a family* $\mathcal{F}$ *of pairwise non-adjacent regions of* $\mathcal{C}'$*, such that*

$$\mathbb{P}_{\mathrm{w}\sim\mathcal{N}^{\otimes d}}\left(\mathrm{z}(\mathrm{w})^T\mathrm{u} \neq 0\right) \geq \frac{|\mathcal{C}' \setminus \mathcal{F}|}{\omega_{d-1}} \tag{1}$$

*Proof of Lemma 4.3.* Since w is a gaussian vector, it is isotropic, i.e. $\mathrm{w}/\|\mathrm{w}\| \sim \mathrm{Unif}(S^{d-1})$. Let $F_{\mathrm{u}} : S^{d-1} \to \mathbb{R}$ be defined as $\mathrm{w} \to \mathrm{z}(\mathrm{w})^T\mathrm{u}$, and notice that by construction it is constant on each region of $\mathcal{C}$.
So it is sufficient to prove that the regions of $\mathcal{C}$ on which $F_{\mathrm{u}} = 0$ are pairwise non-adjacent.
*Step 1)* Assume that the vector u has no null components, i.e. $u_i \neq 0$ for $i = 1, \dots, k$. If two regions $A$ and $B$ of $\mathcal{C}$ are adjacent it means that they intersect on some diameter $D_i = \{\mathrm{v} \in S^{d-1} \,|\, \mathrm{v} \perp \mathrm{x}_i\}$, and if $\mathrm{a} \in A$ and $\mathrm{b} \in B$ the components of $\mathrm{z}(\mathrm{a})$ and $\mathrm{z}(\mathrm{b})$ satisfy $z(\mathrm{a})_j = z(\mathrm{b})_j$ if and only if $j \neq i$. In particular, since $z_i = \pm 1$, it holds $F_{\mathrm{u}}(\mathrm{a}) = F_{\mathrm{u}}(\mathrm{b}) \pm 2u_i$, and the result follows because $u_i \neq 0$.
*Step 2)* In general we proceed by induction on $k$. The case $k = 2$ is easy. Recall $\mathrm{u}_1 \neq 0$, and $\mathrm{u} \perp S$ means $\mathrm{u}^T\mathrm{y}_i = 0$ for $i = 0, \dots, n$. Since $\mathrm{y}_i = \mathrm{z}(\mathrm{w}_i)$ with $\mathrm{w}_i \sim \mathcal{N}^{\otimes d}(0,1)$ it holds $y_{ij} = \pm 1$ for $j = 1, 2$. Suppose by contradiction that $u_1 = 0$, then $\mathrm{u}^T\mathrm{y}_i =$

[3] We are using Euclidean trigonometry for simplicity. To be precise we should use spherical trigonometry, as we will in the actual proof (cf. the proof of Lemma 4.5 in Section B.3 of the Appendix).

$u_1 y_{i1} + u_2 y_{i2} = \pm u_2 = 0$, so $u_2 = 0$, hence $\mathrm{u} = 0$ which is a contradiction. So the case $k = 2$ follows from step 1.
We assume the lemma holds for $k-1$ and we prove it for $k$. If u has no null components the result follows from step 1. If $u_j = 0$ for some $j$ we can consider the family $\{\mathrm{x}_i\}_{i \neq j}$ of samples vectors from which we have removed the vector $\mathrm{x}_j$ (hence with cardinality $k-1$), and the associated vectors and function $\mathrm{u}'$, $\mathrm{z}'$, $F'_{\mathrm{u}'}$. By construction $F'_{\mathrm{u}'} \cong F_{\mathrm{u}}$, and the result follows by induction. □

The following theorem provides a quantitative lower bound on the measure of families of non adjacent regions in a conical tessellation.

**Theorem 4.4** (measure of non adjacent regions)**.** *Let $\mathcal{D} = \{D_i \subset S^m\}_{i=1,\dots,k}$ be a generic $\theta$-separated conical tessellation of $S^m$, with regions $\mathcal{C}$, and let $\mathcal{F} \subset \mathcal{C}$ be a family of pairwise non-adjacent regions. Then*

$$|\mathcal{C} \setminus \mathcal{F}| \geq \frac{\omega_{m-1}\theta}{mN},$$

*where $N = 2\sum_{i=0}^{m-1} \binom{k-2}{i}$ is the constant coming from Zaslavsky's Theorem A.3.*

To prove this theorem we need the following preliminary lemma, in the spirit of isoperimetric inequality.

**Lemma 4.5** (Surface-Volume inequality)**.** *Let $R$ be a region of a $\theta$-separated conical tessellation in $S^m$, with a face $\xi$ of $(m-1)$-dimensional measure $|\xi|$. Then*

$$|R| \geq \frac{2|\xi|^2\theta}{\omega_{m-1}m}.$$

Due to length constraints, the proof of Lemma 4.5—together with the geometric tools needed—is contained in Appendix B.

*Proof of Theorem 4.4.* Let $D_1 \in \mathcal{D}$ be a diameter of the tessellation, $D_1 \cong S^{m-1}$.
We proceed by finding a lower bound for the total measure of all regions adjacent to $D_1$ that are not in $\mathcal{F}$. Let $\mathcal{D}_1 \coloneqq \{\mathcal{D} \setminus D_1\} \cap D_1$ be the tessellation of $D_1$ defined by diameters $\{D_i \cap D_1\}_{i=2,\dots,k}$, and let $2t$ be the number of regions of $\mathcal{D}_1$. We denote $\xi_i$ the regions and $|\xi_i|$ their measure. Notice that $|\xi_i| \leq \omega_{m-1}/2$ because each region $|\xi_i|$ is an intersection of hemispheres in $D_1$. Each region $\xi_i$ is the common face of exactly two regions $R_i^{\pm}$ of $\mathcal{C}$, and at most one of them belongs $\mathcal{F}$, because $\mathcal{F}$ contained non-adjacent regions. Then we have

$$|\mathcal{C} \setminus \mathcal{F}| \geq \min \sum_{i=1}^{2t} \frac{2|\xi_i|^2\theta}{\omega_{m-1}m} \geq \frac{\omega_{m-1}\theta}{tm} \geq \frac{\omega_{m-1}\theta}{Nm}. \quad (2)$$

The first inequality follows from Lemma 4.5, the second follows by convexity of $x^2$ from Jensen's inequality, and the last from the fact that a generic arrangement has maximal number of regions $N$. □

We are now prepared to prove Lemma 4.1, building on the results derived above.

*Proof of Lemma 4.1.* By Lemma 4.3, and Theorem 4.4, we have

$$\mathbb{P}_{\mathrm{w}\sim\mathcal{N}^{\otimes d}}\left(\mathrm{z}(\mathrm{w})^T\mathrm{u} \neq 0\right) \geq \frac{\omega_{d-2}\theta}{\omega_{d-1}(d-1)N}.$$

By Gautschi's inequality (Elezović et al., 2000)

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (1+x)^s,$$

hence setting $s = 1/2$ and $x = d/2$ we get

$$\frac{\omega_{d-2}}{\omega_{d-1}} = \frac{(d-1)\Gamma(\frac{d}{2}+1)}{d\sqrt{\pi}\Gamma(\frac{d+1}{2})} > \frac{d-1}{\sqrt{2\pi d}},$$

from which follows

$$\mathbb{P}_{\mathrm{w}\sim\mathcal{N}^{\otimes d}}\left(\mathrm{z}(\mathrm{w})^T\mathrm{u} \neq 0\right) > \frac{\theta}{N\sqrt{2d\pi}}.$$

Now recall that $N = 2\sum_{i=0}^{d-1}\binom{k-2}{i} \leq 2^{k-1}$, so we get

$$\mathbb{P}_{\mathrm{w}\sim\mathcal{N}^{\otimes d}}\left(\mathrm{z}(\mathrm{w})^T\mathrm{u} \neq 0\right) > \frac{\theta}{2^k}\sqrt{\frac{2}{d\pi}}.$$

which concludes the proof. □

### 4.3 Collinear vectors

If the input $\{\mathrm{x}_1, \dots, \mathrm{x}_k\}$ contains collinear vectors, $Y = \mathrm{sign}(WX)$ cannot have full rank. In more detail, if two input vectors $\mathrm{x}_i$ and $\mathrm{x}_j$ are collinear, then it is immediate to see that $\mathrm{sign}(W\mathrm{x}_i) = \pm\,\mathrm{sign}(W\mathrm{x}_j)$, depending on whether $\theta_{ij} = 0$ or $\theta_{ij} = \pi$.

In this case, consider the equivalence relation $\sim$ over $\{\mathrm{x}_1, \dots, \mathrm{x}_k\}$ induced by collinearity so that, for any $i, j = 1, \dots, k$, $\mathrm{x}_i \sim \mathrm{x}_j$ if and only if $\theta_{ij} \in \{0, \pi\}$. Denote by $m$ the number of equivalence classes and consider a subset $\mathrm{x}_{i_1}, \dots, \mathrm{x}_{i_m}$, obtained by including exactly one member from each equivalence class. Let $X' \in \mathbb{R}^{d\times m}$ denote the corresponding matrix. Since vectors from different classes are not collinear, $X'$ satisfies the assumptions of Lemma 4.1 and Theorem 4.2 for some minimum angle $\theta$, so that Theorem 4.2 again holds with $m$ replacing $k$.

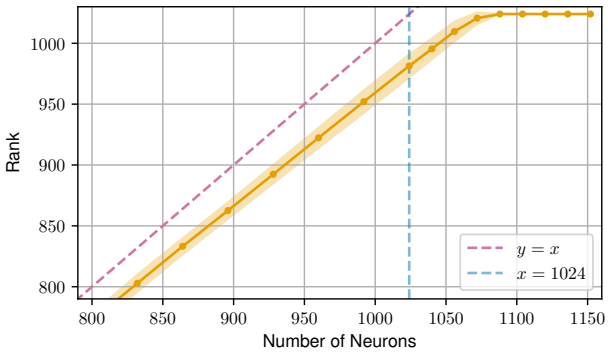


**Figure 4:** *Rank vs. number of neurons (FashionMNIST, $k = 1024$). Plotted is the numerical rank of $Y$ (vertical axis) against hidden-layer width $n$ (horizontal axis) for a fixed subset of 1024 flattened FashionMNIST images ($d = 784$). The dashed line marks the bisector ($y = x$), shaded band denotes $\pm$std over repeated trials, and a vertical marker indicates $n = 1024$*

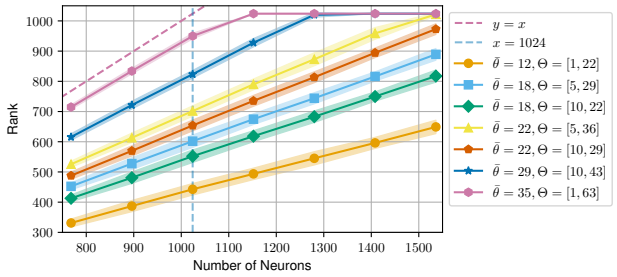


**Figure 5:** *Rank vs. number of neurons (synthetic spherical data, $k = 1024$). Plotted is the numerical rank of $Y$ (vertical axis) against hidden-layer width $n$ (horizontal axis) for multiple synthetic point clouds sampled on the unit sphere ($d = 784$). Each curve corresponds to a different average pairwise angle ($\bar{\theta}$); legend shows also minimum and maximum angles ($\Theta$). The dashed line marks the bisector ($y = x$), shaded band denotes $\pm$std over repeated trials, and a vertical marker indicates $n = 1024$.*

# 5 EXPERIMENTAL ANALYSIS

In this section, we present the results of experiments designed to validate our theoretical findings. For all experiments, we fix a subset of $k = 1024$ data points and arrange them as a data matrix $X \in \mathrm{R}^{d\times k}$. For FashionMNIST(Xiao et al., 2017), each sample is flattened to $d = 28 \times 28 = 784$; for the synthetic experiments, $d$ equals the dimension used to sample points on the unit sphere (see the Appendix C for the exact sampling procedure and the distribution of pairwise angles). To produce hidden representations, we draw a random weight matrix $W \in \mathrm{R}^{n\times d}$ with entries $W_{ij} \sim \mathcal{N}(0, 1)$. Then, we compute the corresponding feature matrix $Y = \mathrm{sign}(WX)$ and measure the numerical rank of $Y$. We repeat the procedure for a grid of different $n$ values and different initialization seeds, to observe how the rank and its variability evolve as the number of neurons increases.

## 5.1 FashionMNIST

In Fig. 4, we show the measured rank of $Y$ for FashionMNIST (Xiao et al., 2017) in relation to the number of neurons ($n$). Two regimes are visible. In the neuron-limited regime ($k < n$), the rank is essentially upper-bounded by $n$ (shown as the bisector line) and grows approximately linearly with it. Once $n$ passes $k$ ($n \geq k$), the rank continues to grow with $n$ and the curve slowly approaches full interpolation. In our runs, the representation achieved rank $= k$ after around 100 additional neurons were added beyond $k$.

## 5.2 Synthetic Dataset

We show in Fig. 5 the rank curves for various synthetic point clouds that have been sampled on the unit sphere. Each curve corresponds to a different angular regime. The minimum and maximum angles between any two lines are also reported in the legend. All curves display an initial, approximately linear increase in rank with $n$, similar to the FashionMNIST case. Differently from it, the number of neurons required to reach interpolation depends strongly on the angular distribution of the points. Point clouds with larger average pairwise angles reach interpolation with fewer neurons. As the average angle decreases, the curves shift downwards, meaning more neurons are required to attain the same numerical rank. In fact, quasi-collinearity slows rank growth so much that, for average angles below 20°, we do not observe interpolation, even when $n$ is as large as $1.5k$. Moreover, we deliberately included cases with the same average angle but different angle ranges. This is to show that, although there is a strong correlation with the average angle, there is also a dependence on the distribution of angles. For cases with angles of 18° and 22°, it can be seen that it is easier to interpolate points with a wider range than with a smaller one.

The synthetic curve with an average angle of around 35° (roughly the same as our subset of FashionMNIST) closely resembles the Fig. 4's curve obtained above. This suggests that the average pairwise angle is a strong predictor of the rank trend, but it is not the full picture: the full distribution of angles (variance and tails) also affects how quickly rank increases with $n$, as the four plots corresponding to angles 18° and 22° suggest.

# 6 DISCUSSION AND OUTLOOK

We studied the *rank of the embedding matrix produced by a perceptron under random initialization* using sign activations, as a minimal geometric mechanism for shattering in overparameterized networks. Our analysis establishes explicit probabilistic bounds in terms of input separation, number of dimensions and hidden width. The model we studied is simple yet rich enough to clearly elucidate key geometric mechanisms underpinning interpolation and shattering. Our results suggest that these phenomena are typical in overparameterized regimes even under elementary models that induce discrete feature spaces.

To many readers, our contribution might at first appear to be a theoretical exercise of little practical import. While this may sound a reasonable criticism, we argue below that there actually is more to it than meets the eye. A first aspect is our focus on sign activations. The main reason behind our choice is that sign activations enable a clean and in our opinion elucidating connection between rank in feature space and the properties of conical tessellations produced by a set of vectors in input space. At the same time, far from being an easier case to study, the sign activation intuitively represents a worst case scenario, given its binary codomain, as opposed to continuous codomains of activations typically used in practice. As a first, anecdotal example, two collinear vectors with opposite directions will be mapped to orthogonal vectors in feature space using ReLU, while collinear vectors are always mapped to collinear images under sign activations, something we discuss and analyze in detail in Section 4.3. Further empirical evidence from a focused experiment comparing the behaviours of sign activations versus ReLU on small, "hard" instances characterized by small average pairwise angles, supports this intuition and is reported in Appendix C.3. Moreover, the absence of a bias, another seemingly marginal difference with respect to previous work on related topics (notably, (Dirksen et al., 2022; Ghosal et al., 2022)), actually seems to make the problem significantly harder, resulting in a one-to-one correspondence with the problem of estimating the probability mass associated to different regions of conical tessellations of the sphere, itself an important problem in high-dimensional linear algebra and spherical geometry. Finally and in a more practical perspective, a number of commonly used activation functions are continuously differentiable approximations of non linear step functions. For example, just like the sigmoid can be seen as a continuously differentiable approximation for the Heaviside function, so is the hyperbolic tangent for the sign.

In a more general perspective, the study of randomly initialized neural layers has a practical bearing in several areas, from properties of the initialization that may affect subsequent training to properties of models that are explicitly used for classification. For example, even the case of the sign function (or Heaviside) can be of interest for models that are only partially trained, such as extreme learning machines, where the ability to achieve full rank is a feature that is considered in some approaches to these models (Huang et al., 2006). Moreover, relatively recent work has made the case for the existence of strong lottery tickets, whereby randomly initialized, overparameterized neural networks contain subnetworks that may well classify unknown inputs. These are clearly properties of the representations resulting from randomly initialized layers and intuition at least suggests that they should be related to their ability to interpolate arbitrary inputs. Finally, as we remarked before, initialization plays a crucial role in the training of deep models in general, and understanding its theoretical aspects may have repercussions on future work and practical developments that might be hard to anticipate.

While our results shed light on one mechanism whereby overparameterization can enable perfect fitting, gaps remain between our theory and empirical observations. Our experiments indicate that rank behavior is influenced not only by minimum separation but by the full distribution of pairwise angles among inputs. In practice, we observe that mean angle serves as a reliable proxy for predicting when full rank is likely to occur. Moreover, experiments suggest that the true probability of increasing rank is substantially larger than our analysis predicts and is hardly consistent with the dependence on the number of points and dimensions resulting from Lemma 4.1.

A first direction of further research clearly concerns closing the aforementioned gap, something that could shed light on finer mechanisms of rank growth. A number of potential approaches are possible. For example, increasing rank of the feature matrix should be easier at the beginning of the idealized construction described in Algorithm 1, something that is not adequately captured in our analysis. Further, potential approaches are purely geometric and exploit properties of the conical tessellations that could be investigated and modeled with finer detail.

A further, broader research direction concerns connecting this type of results to generalization. Full rank guarantees shattering, but not generalization per se. Understanding how the rank profile of random features impacts margins, implicit bias, or stability could help connect our results to the "benign overfitting" regime of modern practice.

**Acknowledgments**

Maria Sofia Bucarelli acknowledges partial support from the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001. Francesco Caso acknowledges support from Renaissance Philanthropy through the AI for Math Fund for research conducted at the Department of Computer Science and Technology, University of Cambridge. This work acknowledges support from projects FAIR (PE0000013), under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, and project NEREO (Neural Reasoning over Open Data), funded by the Italian Ministry of Education and Research (PRIN) Grant no. 2022AEFHAZ.

**References**

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019). On exact computation with an infinitely wide neural net.

Basteri, A. and Trevisan, D. (2023). Quantitative gaussian approximation of randomly initialized deep neural networks.

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

Bilyk, D., Dai, F., and Matzke, R. (2018). The stolarsky principle and energy optimization on the sphere. *Constructive Approximation*, 48(1):31–60.

Bilyk, D. and Matzke, R. (2019). On the fejes tóth problem about the sum of angles between lines. *Proceedings of the American Mathematical Society*, 147(1):51–59.

Bowman, B. and Montufar, G. (2022). Spectral bias outside the training set for deep networks in the kernel regime.

Bubeck, S., Eldan, R., Lee, Y. T., and Mikulincer, D. (2020). Network size and size of the weights in memorization with two-layers neural networks. *Advances in Neural Information Processing Systems*, 33:4977–4986.

Chan, T. M. and Har-Peled, S. (2011). Approximation algorithms for maximum independent set of pseudodisks.

Chizat, L., Oyallon, E., and Bach, F. (2020). On lazy training in differentiable programming.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334.

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

Dirksen, S., Genzel, M., Jacques, L., and Stollenwerk, A. (2022). The separation capacity of random neural networks. *Journal of Machine Learning Research*, 23(309):1–47.

Domingos, P. (2020). Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*.

Elezović, N., Giordano, C., and Pečarić, J. (2000). The best bounds in gautschi's inequality. *Mathematical Inequalities & Applications*, 3(2):239–252.

Flum, J. and Grohe, M. (2006). *Parameterized Complexity Theory*. Springer Berlin, Heidelberg.

Frankle, J. and Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Ghosal, P., Mahankali, S., and Sun, Y. (2022). Randomly initialized one-layer neural networks make data linearly separable. *arXiv preprint arXiv:2205.11716*.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Huang, G.-B. and Babri, H. A. (1998). 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.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501.

Huang, S.-C. and Huang, Y.-F. (1991). Bounds on the number of hidden neurons in multilayer perceptrons. *IEEE Transactions on Neural Networks*, 2(1):47–55.

Igelnik, B. and Pao, Y.-H. (1995). Stochastic choice of basis functions in adaptive function approximation and the functional-link net. *IEEE Transactions on Neural Networks*, 6(6):1320–1329.

Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural tangent kernel: Convergence and generalization in neural networks.

Jaeger, H. (2001). The"echo state"approach to analysing and training recurrent neural networks.

Janson, S. (2018). Tail bounds for sums of geometric and exponential variables. *Statistics & Probability Letters*, 135:1–6.

Lee, J. M. (2013). *Introduction to Smooth Manifolds*. Springer New York, NY.

Lillo, S. D., Marinucci, D., Salvi, M., and Vigogna, S. (2025). Spectral complexity of deep neural networks.

Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Neal, R. M. (1996). *Priors for Infinite Networks*, pages 29–53. Springer New York, New York, NY.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pinkus, A. (1999). Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195.

Python Software Foundation (2024). Python: A programming language. `https://www.python.org/`.

Rahimi, A. and Recht, B. (2008). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21.

Rajput, S., Sreenivasan, K., Papailiopoulos, D., and Karbasi, A. (2021). An exponential improvement on the memorization capacity of deep threshold networks. *Advances in Neural Information Processing Systems*, 34:12674–12685.

Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., and Rastegari, M. (2020). What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11893–11902.

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

Schmidt, W., Kraaijveld, M., and Duin, R. (1992). Feedforward neural networks with random weights. In *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems*, pages 1–4.

Schoenberg, I. J. (1935). Remarks to maurice frechet's article"sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de hilbert. *Annals of Mathematics*, pages 724–732.

Siegel, J. W. (2023). Sharp lower bounds on interpolation by deep relu neural networks at irregularly spaced data. *arXiv preprint arXiv:2302.00834*.

Sontag, E. D. (1997). Shattering all sets of 'k'points in "general position" requires (k—1)/2 parameters. *Neural Computation*, 9(2):337–348.

Steinerberger, S. (2023). The first eigenvector of a distance matrix is nearly constant. *Discrete Mathematics*, 346(4):113291.

Thomas Erlebach, K. J. and Seidel, E. (2005). Polynomial-time approximation scheme for geometric intersection graphs.

Todhunter, I. (1886). Spherical trigonometry.

Vardi, G., Yehudai, G., and Shamir, O. (2021). On the optimal memorization power of relu neural networks. In *International Conference on Learning Representations*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xiao, L., Pennington, J., and Schoenholz, S. S. (2020). Disentangling trainability and generalization in deep neural networks.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64:107 – 115.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] The setting, problem definition, and relevant preliminary concepts can be found in Sections 3 and 4 and Appendix A. However, we do not introduce new algorithm or models: the contribution of this paper lies in providing a proof of the probability of being full rank. For the experimental part, the algorithm used is solely for obtaining the results.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We do not introduce new algorithms this paper. We only used shallow network, the number that we do not train. The number of neurons is cited. The only algorithm presented is Algorithm 1, which serves as an incremental construction to guide the proof of the main theorem.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The source code is included in the supplementary materials.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] The assumptions, problem definition, and relevant concepts are presented in Sections 3 and 4 and Appendix A.
   (b) Complete proofs of all theoretical results. [Yes] Some proofs are included in Section 4; the remainder are provided in the supplementary materials, specifically in Appendix B.
   (c) Clear explanations of any assumptions. [Yes] See Sections 1 and 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The code is included in the supplementary materials.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] We do not train models, as our focus is on initialization. However, the provided code and Appendix C provide complete information for running the experiments and reproducing the plots, including synthetic dataset generation and the random seeds used to initialize the networks.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Error bars are included in our plots, representing the standard deviation of the rank, as described in Section 5.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] The computing infrastructure is described in Appendix C.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] We cite the datasets and libraries used in our experiments.
   (b) The license information of the assets, if applicable. [Yes] The license are reported in Appendix C.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] We introduce synthetic datasets in our experiments; details are included in the supplementary materials (code and Appendix C).
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable] We did not used crowdsourcing or conducted research with human subjects.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# A BACKGROUND ON HIGH-DIMENSIONAL AND SPHERICAL GEOMETRY

In this section, we go more in detail with respect to the geometric background introduced in 3. Sections A.1 and A.2 introduce key definitions and notions in high-dimensional and spherical geometry that will be needed in the technical sections. A reader already familiar with these topics can skip this section.

## A.1 Hyperplane arrangements

**Remark A.1.** *An oriented hyperplane arrangement is generic if and only if its perpendicular vectors are in general position. This is a stronger condition than being $\theta$-separated, in fact there are $\theta$-separated families of vectors that are not in general position for any $\theta$. However:*

1. *For any $\theta$-separated family of vectors $\{\mathrm{v}_i\}$ and any $\epsilon > 0$, there exist a $(\theta - \epsilon)$-separated family of vectors $\mathrm{v}'_i$ in general position with $d(\mathrm{v}_i, \mathrm{v}'_i) < \epsilon$ for all $i$.*
2. *If the vectors are sampled uniformly on the sphere, then they are in general position with probability 1.*
3. *In general, we will be interested in estimating from below geometric quantities that decrease as the number of regions of the arrangement increase, and the number of regions is maximal for generic arrangements. So if we assume that the vectors are in general position we will find valid (although possibly weaker) estimates.*

**Example A.2.** *A generic central hyperplane arrangement in $\mathbb{R}^m$ with at least $m$ elements is essential.*

The *adjacency graph* of $\mathcal{A}$ has the regions of $\mathcal{A}$ as vertices, and two vertices are connected by an edge if and only if the corresponding regions share a face of codimension 1.

To any hyperplane arrangement $\mathcal{A}$ corresponds also a ranked poset $\mathcal{P}(\mathcal{A})$ whose elements are non-empty intersections of sets of hyperplanes, ordered by reverse inclusion, and with rank function $\mathrm{rk}(f) = \mathrm{codim}(f)$.
The *Möbius function* $\mu$ is defined inductively on $\mathcal{P}(\mathcal{A})$ by setting

$$\mu(\hat{0}, \hat{0}) = 1 \quad \text{and} \quad \mu(\hat{0}, x) = -\sum_{y<x} \mu(\hat{0}, y),$$

and the characteristic polinomial of $\mathcal{A}$ is defined as

$$\chi_{\mathcal{A}}(t) = \sum_{x \in \mathcal{P}(\mathcal{A})} \mu(\hat{0}, x) t^{\mathrm{rk}(x)}.$$

**Theorem A.3** (Zaslavsky)**.** *The number $\#R$ of regions of a hyperplane arrangement is equal to sum of the absolute values of the coefficients of $\chi_{\mathcal{A}}(t)$.*

**Corollary A.4.** *Let $\mathcal{A} = \{H_i\}_{i=1\dots,k}$ (resp. $\mathcal{A}_c$) be a generic, non-central (resp. central) arrangement of $k$ hyperplanes in $\mathbb{R}^m$, then*

$$\#R(\mathcal{A}) = \sum_{i=0}^{m} \binom{k}{i}, \qquad \text{and} \qquad \#R(\mathcal{A}_c) = 2 \sum_{i=0}^{m-1} \binom{k-1}{i}.$$

*Proof.* For a generic non-central hyperplane arrangement $\mathcal{A}$ in $\mathbb{R}^m$ with $k$ hyperplanes one can check that $\mu(\hat{0}, x) = -1^{\mathrm{rk}(x)}$[4], from which follows $\#R = \sum_{i=0}^{m} \binom{k}{i}$. For a generic central arrangement $\#R = 2\sum_{i=0}^{m-1} \binom{k-1}{i}$ follows easily by choosing $V \in \mathcal{A}$, considering two hyperplanes $V^{\pm}$ parallel to $V$ and on the two opposite sides of $V$, and noticing that $\#R_{\mathcal{A}} = \#R_{\{\mathcal{A}\setminus V\}\cap V^+} + \#R_{\{\mathcal{A}\setminus V\}\cap V^-}$.
Another way to see this is by computing the characteristic polynomial of a generic central hyperplane arrangement, which is

$$\chi(t) = \sum_{i=0}^{m-1} (-1)^i \binom{k}{i} t^i + (-1)^m \binom{k-1}{d-1} t^m.$$

□

[4] In fact this property holds for the Möbius function of *any* hyperplane arrangement.

### A.2 Spherical geometry

We denote $S^m = \{\mathrm{x} \in \mathbb{R}^{m+1} \mid \|\mathrm{x}\| = 1\}$ the $m$-dimensional unit sphere, which is a Riemannian manifold with the *geodesic distance*

$$d_{S^m}(\mathrm{x},\mathrm{y}) = \sphericalangle(\mathrm{x},\mathrm{y}) = \arccos\frac{\langle \mathrm{x},\mathrm{y}\rangle}{\|\mathrm{x}\|\,\|\mathrm{y}\|} = \arccos(\mathrm{x}^T\mathrm{y}).$$

A *diameter* $D$ of $S^m$ is a submanifold of codimension 1 given by the intersection of $S^m$ with a hyperplane $H \subset \mathbb{R}^{m+1}$, and clearly $D \cong S^{m-1}$. We say that a diameter $D = S^m \cap H$ is *orthogonal* to a point $\mathrm{x} \in S^m$ if $\mathrm{x} \perp H$.
We denote $B_{S^m}(\mathrm{x}, r) = \{y \in S^m \,|\, d_{S^m}(\mathrm{x},\mathrm{y}) \le r\}$ the closed ball in $S^m$ of radius $r$ and centered in x. To any point $\mathrm{x} \in S^m$ is associated the *hemisphere* $E_\mathrm{x} = B_{S^m}(\mathrm{x}, \pi/2) = \{\mathrm{y} \in S^m | \langle \mathrm{x},\mathrm{y}\rangle \ge 0\}$, whose boundary $D_\mathrm{x} = \partial E_\mathrm{x} \cong S^{m-1}$ is the diameter of $S^m$ orthogonal to x.
The distance $d_{S^m}$ induces a distance $d^H_{S^m}$ on closed subsets of $S^m$, called the *Hausdorff distance*, namely

$$d^H_{S^m}(A,B) = \max\{\min\{r \ge 0 \,|\, B \subset A_r\}, \min\{r' \ge 0 \,|\, A \subset B_{r'}\}\},$$

where $A_r = \cup_{\mathrm{x}\in A} B_{S^m}(\mathrm{x}, r)$ denote the set of points at distance $\le r$ from $A$.
Let $\mathrm{x},\mathrm{y} \in S^m$ be two points with $\sphericalangle(\mathrm{x},\mathrm{y}) < \pi/2$. Then for two diameters $D_x, D_y$ we have $d^H_{S^m}(D_\mathrm{x}, D_\mathrm{y}) = d_{S^m}(\mathrm{x},\mathrm{y}) = \sphericalangle(\mathrm{x},\mathrm{y})$, i.e. the Hausdorff distance of $D_\mathrm{x}$ and $D_\mathrm{y}$ is equal to the geodesic distance of x and y, which is also equal to the angle between the two diameters. Hence, if $d(\mathrm{x},\mathrm{y}) \ge \theta$ we extend the terminology by saying that $E_\mathrm{x}$ and $E_\mathrm{y}$ are *$\theta$-separated*.

A geodesic triangle $\Delta(A,B,C)$ is the union of three geodesic segments connecting the points $A, B, C$. Angles between the geodesic segments are defined via the tangent space.
The following is a classical result of spherical trigonometry.

**Lemma A.5.** *[(Todhunter, 1886) Art. 62] Let $\Delta(A,B,C)$ be a right spherical triangle with height $h = \bar{BC}$, base $b = \bar{AB}$ and angles $\sphericalangle_B(C,A) = \pi/2$ and $\sphericalangle_A(B,C) = \theta$. Then*

$$\tan h = \tan\theta \sin b.$$

**Spherical cones** A *cone $C(A,p)$ from a point $p$ over a set $A$* is defined as the geodesic convex hull of $A \cap \{p\}$. The subset $A$ and the point $p$ are called respectively *base* and *apex* of $C$. Notice that the notion of convexity here depends on the godesic structure of the space, not on its embedding into some ambient space.
In $\mathbb{R}^n$, if $A$ is a compact subset contained in some hyperplane $V$ and $p$ is contained in some parallel hyperplane $W$ with $d^H(V,W) = h$, then $C(A,p)$ could be called a cone (or pyramid) of base $A$ and height $h$.
For our purpose we define a *spherical cone* to be a cone in $S^m$, with base $A$ contained in some totally geodesic hypersphere $S^{m-1} \subset S^m$ and apex $p$ whose projection on $A$ is contained in the interior of $A$, i.e. $\pi_A(p) \in A^\circ = A \setminus \partial A$. Then the *height* of $C$ is simply the distance $d(p,A)$.
For brevity we will not specify that a cone is *spherical* when it is clear from the context.

**Spherical polytopes** A family of hemispheres $\{E_i\}_{i\in I}$ defines by intersection a region $C = \cap_{i\in I} E_i$, and we call it *minimal* if there is no $j \in I$ such that $\cap_{i\in I\setminus\{j\}} E_i = C$, i.e. if every hemisphere $E_i$ is necessary to define the region $C$.
A *spherical convex polytope* $C \subset S^m$ is the non-empty intersection of a finite minimal family of hemispheres $\{E_i\}$. A *side of codimension $l$* of $C$ is a non-empty intersection $C \cap D_{i_1} \cap \cdots \cap D_{i_l}$, where $D_{i_l} = \partial E_{i_l}$. Notice that the minimality condition on the family of hemispheres is important to properly define the sides of $C$. A polytope with exactly two sides of codimension 1 is called a *ditope*.
The *inscribed radius* of a convex polytope $C$ is the maximum radius of a ball contained in $C$, namely

$$r_{in}(C) = \max\{r \ge 0 \,|\, B_{S^m}(r,\mathrm{x}) \subset C, \mathrm{x} \in C\},$$

and any ball $B_{in} \subset C$ of radius $r_{in}(C)$ contained in $C$ is called *inscribed ball of $C$*.

If $A \subset \mathbb{R}^m$ is a subset of (Hausdorff) dimension $l$ we will denote by $|A| = |A|_l$ the *($l$-dimensional) Hausdorff measure* of $A$. To simplify the notation we will avoid to specify the dimension of the measure, and will always write $|A|$ instead of $|A|_l$, since there is only one choice of $l$ that makes the measure positive and finite, namely

when $l$ matches the (Hausdorff) dimension of $A$.
We denote

$$\omega_{m-1} = |S^{m-1}| = \frac{m\pi^{m/2}}{\Gamma(\frac{m}{2}+1)}$$

the measure of the $(m-1)$-sphere of radius 1.

# B PROOF OF THEOREM 4.4

This appendix collects the technical results that support the main theorems in Section 4. In particular, we develop the geometric machinery needed to analyze measures of regions in conical tessellations. Section B.1 establishes a cone volume comparison theorem, providing explicit lower bounds for spherical cones by comparison with Euclidean cones. Section B.2 introduces an "isoinradii inequality" for spherical polytopes, showing that ditopes maximize volume under fixed inradius, and develops auxiliary lemmas (covering and symmetrization arguments) that are needed to prove the isoinradii inequality. Section B.3 put these results together finally proving the lemmas needed to provide the quantitative bounds on non-adjacent regions that Theorem 4.4 uses.

## B.1 Cone volume comparison

Comparing geometric properties of objects in curved and Euclidean spaces is a fruitful and well established research area in geometry, and the Bishop-Gromov comparison theorem is a cornerstone in this field, comparing the volume of balls under lower Ricci curvature bounds.
As we have seen in the low dimensional examples, a central part of our approach is estimating the measure of $\theta$-separated convex polytopes given the measure of one of their sides. To do so we constructed triangles with sides that form an angle $\theta$ with the face and measured their area. In higher dimension the analogue of this construction involves the construction of cones, but the measure of harder to compute. In order to appreciate a readable lower bound in our final result, we prove here the following explicit estimate of the measure of a spherical cone, which is reminiscent of Bishop-Gromov's comparison theorem, by comparing it with the measure of an appropriate Euclidean comparison cone.

**Theorem B.1** (Cone volume comparison)**.** *Let $C = C(A,p) \subset S^m$ be a cone with convex base $A \subset S^{m-1}$ of measure $|A|$ and apex $p$ of height $h = d(A,p) \in (0,\pi/2]$. Let $\hat{C} = \hat{C}(|A|,h)$ be a comparison euclidean cone, i.e. a cone in $\mathbb{R}^n$ with base of measure $|A|$ and height $h$. Then $|C| \geq |\hat{C}|$.*

As one may notice, in our theorem the inequality is reversed with respect to the classical Bishop-Gromov's comparison theorem. While balls can indeed be seen as cones (from any interior point over their boundary), our setting has a couple of slight differences. First of all, we assume the base to be totally geodesic, which is not the case for the boundary sphere of general balls. Secondly, and most importantly, we are fixing the measure of the base, i.e. the perimeter of the ball, rather than its radius. Since both the perimeter and the area of the balls scale with the radius, we see how our estimates is independent on the classical one.
Before proving the theorem let us recall the following

**Proposition B.2** (coarea formula)**.** *Let $(M,g)$ be a ($m$-dimensional) Riemannian manifold, $A \subset M$ a Borel subset, and $f : M \to \mathbb{R}$ a 1-lipschitz map. Then*

$$|A|_m \geq \int_{x\in\mathbb{R}} |f^{-1}(x) \cap A|_{m-1} dx$$

*Proof of Theorem B.1.* Let us fix notation. We call $D_0 = q^\perp$ the diameter containing $A$, $A \subset D_0 = q$. The function $f : S^m \to \mathbb{R}$ defined as $f(x) = d_{S^m}(x, D_0)$ is clearly 1-lipschitz, and we denote $H_s := f^{-1}(s)$. If we represent $S^m = \{\mathrm{x} \in \mathbb{R}^{m+1} \mid \|\mathrm{x}\| = 1\}$ so that $q = e_0$, then $H_s$ is simply the intersection of the sphere with the hyperplane orthogonal to $q$ at distance $\sin(s)$ from the origin, $H_s = S^m \cap \{x_0 = \sin(s)\}$. In particular $H_s$ is an $(m-1)$-dimensional sphere of radius $\cos(s)$.
The proof will proceed as follows. By the coarea formula

$$|C| \geq \int_0^h |A_s| ds = \int_0^h \sigma(s)^{m-1} |A| ds,$$

where in the last passage we expressed $|A_s|$ proportionally w.r.t. $|A|$.[5]
Furthermore

$$|\hat{C}| = \int_h \left(1 - \frac{s}{h}\right)^{m-1} |A| ds.$$

To conclude the proof we proceed by computing the proportional factor $\sigma(s)$ and showing that $\sigma(s) \geq (1 - s/h)$.

Let $b = b_0 \in A$ be the projection of the apex $p$ to the base, and let $b_s$ be the point along the geodesic segment $\overline{pb}$ at distance $s$ from $b$. We can write the measure of $A_s$ as an integral in exponential coordinates of $\cos(s)S^{m-1}$ centered in $b_s$. For $s = 0$ we have

$$|A| = \int_{S^{m-2}} \int_0^{\rho_{\max}(\omega)} \sin^{m-2}(\rho) d\rho d\omega$$

where $\rho_{\max} : S^{m-2} \to [0, \pi/2]$ is the profile function describing the distance of $b$ from $\partial A$ in direction $\omega$. For $s > 0$ we have

$$|A_s| = \int_{\cos(s)S^{m-2}} \int_0^{\rho^s_{\max}(\omega)} \sin^{m-2}(\rho) d\rho d\omega$$

Let $a$ be the point on $\partial A$ in direction $\omega$ from $b$, and let us denote $\alpha = \alpha(\omega)$ the angle $\sphericalangle_a(p, b)$ be the angle in $a$ formed by the apex and its projection $b$.
On $A_s$ we denote $r_s$ the point on $\partial A_s$ in the same radial direction $\omega$ from $b_s$. The point $r_s$ project to some point $r \in \overline{ab}$ and all these points lie in the same totally geodesic 2-sphere on which lie $a, b, p$.
For $s > 0$ the section $A_s$ is not totally geodesic, so in particular the lenght of the radial segment $\overline{b_s r_s}$ is longer than the spherical distance of its endpoints, i.e. $\rho^s_{\max}(\omega) > d_{S^{m-1}}(b_s, r_s)$. Let us denote $|\overline{xy}|$ the length of the segment $\overline{xy}$. We now proceed to determine the lenght of the segment $\overline{b_s r_s}$. Recall that by spherical trigonometry we have

$$\sin|\overline{ab}| = \tan(\alpha - \pi)\tan(h) \qquad \text{and} \qquad \sin|\overline{ar}| = \tan(\alpha - \pi)\tan(s),$$

from which follows

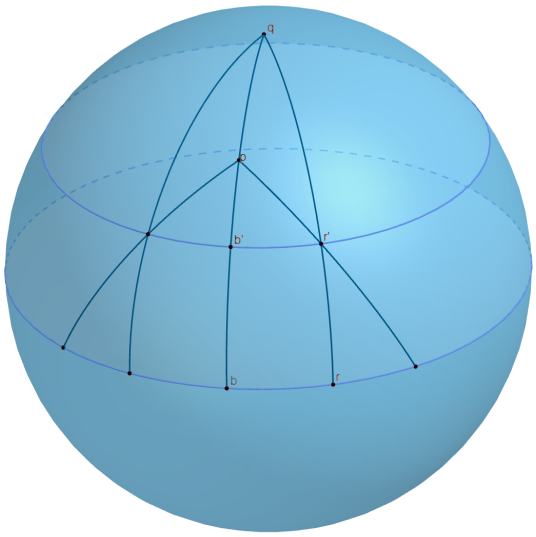


**Figure 6:** *We want to integrate the $m-1$-measure of the horizontal slices along the height.*

$$\frac{\sin(|\overline{ab}|)}{\tan(h)} = \frac{\sin(|\overline{ar}|)}{\tan(s)}$$

Furthermore $|\overline{r_s b_s}| = \cos(s)|\overline{rb}|$ and $|\overline{br}| = |\overline{ab}| - |\overline{ar}|$, so putting all of this together we get

$$|\overline{b_s r_s}| = \cos(s)\left(|\overline{ab}| - \arcsin\left(\frac{\tan(s)}{\tan(h)}\sin(|\overline{ab}|)\right)\right)$$

[5] e.g. if $h = \pi/2$ its easy to see that $\sigma(s) = \cos(s)$

Recall that by definition the lengths of the segments $\overline{ab}$ and $\overline{b_s r_s}$ are respectively $\rho_{\max}$ and $\rho^s_{\max}$, so we found

$$\rho^s_{\max} = \cos(s)\left(\rho_{\max} - \arcsin\left(\frac{\tan(s)}{\tan(h)}\sin(\rho_{\max})\right)\right)$$

which gives the following rescaling factor

$$\rho^s_{\max} = \rho_{\max}\cos(s)\left(1 - \frac{\arcsin\left(\frac{\tan(s)}{\tan(h)}\sin(\rho_{\max})\right)}{\rho_{\max}}\right) = \rho_{\max}\sigma(s).$$

Then we can write

$$|A_s| = \int_{\cos(s)S^{m-2}}\int_0^{\rho_{\max}(\omega)\sigma(s)} \sin(\rho)^{m-2}d\rho d\omega \geq \tag{3}$$

$$\geq \sigma(s)^{m-1}\int_{S^{m-2}}\int_0^{\rho_{\max}(\omega)} \sin(\rho)^{m-2}d\rho d\omega = \sigma(s)^{m-1}|A|. \tag{4}$$

The inequality follows by convexity of $\sin(x)$ which implies

$$\int_0^{\sigma R}\sin^{m-2}(\rho)d\rho = \sigma\int_0^R \sin^{m-2}(\sigma\rho)d\rho \geq \sigma^{m-1}\int_0^R \sin^{m-2}(\rho)d\rho.$$

To conclude the proof we claim $\sigma(s) \geq (1 - s/h)$.

For $t \leq 1$ we have $\arcsin(tx) < t\arcsin(x)$ by convexity, so it follows

$$\sigma(s) = \cos(s)\left(1 - \arcsin\left(\frac{\tan(s)}{\tan(h)}\sin(\rho_{\max})\right)\frac{1}{\rho_{\max}}\right) \geq \cos(s) - \frac{\sin(s)}{\tan(h)}$$

$$= \frac{\cos(s)\sin(h) - \cos(h)\sin(s)}{\sin(h)} = \frac{\sin(h-s)}{\sin(h)}$$

By convexity of $\sin(x)$, we have $\sin(h-s) \geq (1 - \frac{s}{h})\sin(h)$, so we find

$$\sigma(s) \geq \frac{\sin(h-s)}{\sin(h)} \geq \frac{(1-\frac{s}{h})\sin(h)}{\sin(h)} = (1 - \frac{s}{h})$$

which concludes the proof. □

### B.2 Isoinradii inequality and extremal spherical polytopes

Another ingredient we need in our proof to estimate the measure of a region, will be to know the minimal height of a $\theta$-separated cone that we can construct on their faces. This is equivalent to asking which is the minimal inscribed radius of a face with fixed measure. Or equivalently, which are the convex spherical polytopes that maximise the ratio $|C|/r_{in}(C)$. Clearly it is the same to either fix the measure and find the polytope of minimal inscribed radius, or to fix the inscribed radius and maximise the measure. We shall do the latter, and we conjecture that the extremal polytopes are ditopes associated to $2r_{in}(C)$-separated diameters.

**Theorem B.3** (Extremal polytopes)**.** *Let $C$ be any convex polytope in $S^m$. If $D$ is a convex ditope in $S^m$ with $r_{in}(C) = r_{in}(D)$, then $|D| \geq |C|$.*

From now on we denote $C \subset S^m$ a convex polytope, $\mathcal{B}_{in}$ be a inscribed ball of $C$, and $\mathcal{S}_{in} = \partial\mathcal{B}_{in}$ its boundary sphere. $r_{in}(C)$ is the inscribed radius of $C$, so is also the radius of $\mathcal{B}_{in}$, and we assume that $r_{in}(C) < \pi/2$.

**Lemma B.4.** *Let $\{t_i\} = \partial C \cap \mathcal{B}_{in}$ the points where $\mathcal{S}_{in}$ is tangent to $\partial C$. Then $\mathcal{S}_{in}$ is covered by hemispheres (i.e. balls of radius $\pi r_{in}(C)/2$) centered in $t_i$:*

$$\mathcal{S}_{in} \subset \bigcup_i B^{\mathcal{S}_{in}}\left(\frac{\pi r_{in}(C)}{2}, t_i\right)$$

*Proof.* Suppose by contradiction that there exist a point

$$y \in \mathcal{S}_{in} \setminus \bigcup_i B^{\mathcal{S}_{in}}\left(\frac{\pi r_{in}(C)}{2}, t_i\right).$$

Then $B_y \coloneqq B^{\mathcal{S}_{in}}\left(\frac{\pi r_{in}(C)}{2}, y\right)$ contains no $t_i$, hence we can translate $\mathcal{B}_{in}$ by some $\epsilon > 0$ in direction $y$ and obtain a translated ball $\mathcal{B}'_{in}$ which is contained in $C$ but does not intersect $\partial C$, which contradicts the hypotesis that $\mathcal{B}_{in}$ has maximal radius among balls in $C$. □

Before proving our main result we also need the following symmetrization lemma (many thanks to G. for helping with the proof).
For a function $f : S^m \to \mathbb{R}$ we say that $f$ is *p-radially symmetric* if $f$ is invariant under the action of the $p$-stabilizer subgroup of isometires of $S^m$, i.e. if $f = f \circ \sigma$ for any isometry $\sigma$ that fixes $p$.
Recall that the group of isometries of the sphere is the group of orthogonal matrices $\mathrm{Isom}(S^m) = O(m+1)$ and the stabilizer of a point $p$ is the normal subgroup of isometries that fixes $p$, $\mathrm{Stab}(p) \coloneqq \{g \in \mathrm{Isom}(S^m) \,|\, g.p = p\} \cong O(m)$ (cf. (Lee, 2013), Chapter 7, for a detailed treatment of the subject).
If $f$ is $p$-radially symmetric, the quotient of $S^m \setminus \{\pm p\}$ by $\mathrm{Stab(p)} \subset \mathrm{Isom(S^m)}$ it induces a map $\bar{f} : (0, \pi) \to \mathbb{R}$, such that $\bar{f}(r) = f(q)$ for any $q$ with $d(p, q) = r$. We say that $f$ is *p-radially non-decreasing* if $X[f] \geq 0$, where $X[f]$ denotes the derivative of $f$ with respect to the vector field $X = \nabla_x d(x, p)$. Clearly, for a $p$-radially symmetric function $f$ it is equivalent to saying that the radial component $\bar{f}$ is non-decreasing.
Finally, a region $\mathcal{R}$ that contains a point $p$ is *star-shaped with respect to p* if, for any point $q \in \mathcal{R}$ there exists a unique geodesic from $p$ to $q$ contained in $\mathcal{R}$. In particular a convex region that contains $p$ is star-shaped with respect to $p$.
If $q$ is a point in a region $\mathcal{R}$ star-shaped with respect to $p$, and $t \in [0, 1]$, we denote $tq \coloneqq \gamma_q(t)$, where $\gamma_q$ is the unique geodesic contained in $\mathcal{R}$ such that $\gamma(0) = p$ and $\gamma(1) = q$.

**Lemma B.5** (Symmetrization)**.** *Let $\mathcal{B} = B^{\mathbb{S}^m}(p, \pi/2)$ denote the hemisphere of $S^m$ centered at $p$.*
*Let $f : \mathcal{B} \to \mathbb{R}^+$ a positive, continuous, p-radially symmetric, p-radially non-decreasing function on $\mathcal{B}$, and let $R_i \subseteq \mathcal{C}$ be star-shaped regions with respect to $p$ such that $\sum_i |R_i| = |\mathcal{B}|$. Then*

$$\sum_i \int_{R_i} f(x)\, dx \leq \int_{\mathcal{B}} f(x)\, dx.$$

*Proof.* By the intermediate value theorem there exists a radius $r_m \in [0, 1]$ such that $\int_{\mathcal{B}} f = |\mathcal{B}|\bar{f}(r_m)$. For the same reason there exist radii $r_i$ such that for each region $\int_{R_i} f = |R_i|\bar{f}(r_i)$. It is sufficient to show that $r_i \leq r_m$ for each $i$. In fact it would imply

$$\sum_i \int_{R_i} f = \sum_i |R_i|\bar{f}(r_i) \leq \sum_i |R_i|\bar{f}(r_m) = |\mathcal{B}|\bar{f}(r_m) = \int_{\mathcal{B}} f$$

Let $C \subset \mathcal{B}$ be a star-shaped region with respect to $p$. Consider the regions $A = C \cap \{x \in \mathcal{B} \,|\, d(x, p) \geq r_m\}$ and $G = \{tx \in \mathcal{B} \,|\, x \in A, t \in [0, 1]\}$. Finally let $Q = \{x \in \mathcal{B} \,|\, \exists t \in [0, 1] \text{ s.t. } tx \in A\}$ be the sector of $\mathcal{B}$ generated by $A$. Notice that $G \subseteq Q$, and also that $G \subseteq C$ since $C$ is star-shaped with respect to $p$.
By construction we have

$$d(x, p) \geq r_m,\ \forall x \in Q \setminus G \qquad \text{and} \qquad d(x, p) \leq r_m,\ \forall x \in C \setminus G,$$

hence it clearly holds

$$\int_{Q \setminus G} f \geq |Q \setminus G| f(r_m) \qquad \text{and} \qquad \int_{C \setminus G} f \leq |C \setminus G| f(r_m). \tag{5}$$

Furthermore, since $Q$ is a conic sector and $f$ is radially simmetric, we have $\int_Q f = |Q| f(r_m)$. Then we can write

$$|C| f(r_c) = \int_C f = \int_Q f - \int_{Q \setminus G} f + \int_{C \setminus G} f \leq$$

$$\leq |Q| f(r_m) - |Q \setminus G| f(r_m) + |C \setminus G| f(r_m) = |C| f(r_m)$$

Since $f$ is radially non-decreasing we get $r_c \leq r_m$, which concludes the proof. □

*Proof of Theorem B.3.* Let $o$ be the center of $\mathcal{B}_{in} = B_{S^m}(o, r_{in}(C))$, the inscribed sphere of $C$. We begin expressing the measure of a polytope by integration of radii along its inscribed sphere.
For a tangent point $t_i \in \partial C \cap \mathcal{S}_{in}$ consider the diameter $T_i \subset S^m$ tangent to $\mathcal{S}_{in}$ in $t_i$. Equivalently $T_i$ is the diameter which contains the face $f_i$ of $C$ that intersect $\mathcal{S}_{in}$ in $t_i$. For a point $y \in \mathcal{S}_{in}$, the *visual projection* $\sigma_i(y)$ of $y$ to $T_i$ is the first intersection of the geodesic ray from $o$ in direction $y$ with $T_i$ (in total there are two such intersections: $\sigma_i(y)$ and its antipode). We define the *visual distance from $y$ to $T_i$* as $\rho_i(y) = d_{S^m}(o, \sigma_i(y))$, the geodesic distance between $o$ and $\sigma_i(y)$.

Given $y \in \mathcal{S}_{in}$ we can also consider $\sigma_C(y)$ to be the visual projection of $y$ to the polytope boundary $\partial C$, which is the first intersection of the geodesic ray from $o$ in direction $y$ with $\partial C$ (again, by convexity of $C$, there are exactly two such points). Again, we define the visual distance $\rho_C : \mathcal{S}_{in} \to \mathbb{R}^+$ as $\rho_C(y) = d_{S^m}(o, \sigma_C(y))$. Then the measure of the polytope can be expressed as

$$|C| = \int_{\mathcal{S}_{in}} \rho_C \leq \int_{\mathcal{S}_{in}} \min_i \rho_i$$

where the equality is achieved exactly when all sides of $C$ are tangent to $\mathcal{S}_{in}$, since $i$ is counting the sides of $C$ that are tangent to the inscribed sphere $\mathcal{S}_{in}$, as depicted in Fig. 7.
We can express the measure of a ditope $L$ with inscribed sphere $\mathcal{S}_{in}$ in the same way. Let $\pm p \in \mathcal{S}_{in}$ be two

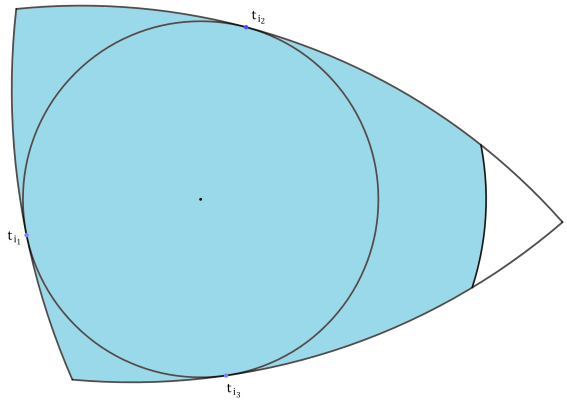


**Figure 7:** *The blue area is the true measure of the polytope, which is smaller than the area of the polytope obtained by only considering* tangent *sides.*

antipodal points, $T_p$ the diameter of $S^m$ tangent to $\mathcal{S}_{in}$ at $p$, and $\rho_p$ the visual distance to $T_p$, then the measure of $L$ is

$$|L| = \int_{\mathcal{S}_{in}} \rho_L = \int_{\mathcal{S}_{in}} \min\{\rho_p, \rho_{-p}\} = 2 \int_{E_p} \rho_p.$$

Now we can compare the measures $|L|$ and $|C|$ of the two polytopes.
The points $t_i$ in $\mathcal{S}_{in}$ define a Voronoi partition of $\mathcal{S}_{in} = \bigcup_i R_i$ into convex subsets $R_i = \{x \in \mathcal{S}_{in} \,|\, d(x, x_i) \leq d(x, x_j)\, \forall j \neq i\}$ with pairwise null-measure intersection. So the upper bound for $|C|$ can be decomposed as

$$|C| \leq \int_{\mathcal{S}_{in}} \min_i \rho_i = \sum_i \int_{R_i} \rho_i$$

By Lemma B.4 each of this regions is contained in a hemisphere, so they can be translated with a spherical isometry so that $t_i \cong p$. We still call $R_i$ these translated regions. Clearly it holds $\sum_i |R_i| = 2|E_p|$ and it is sufficient to prove that

$$\sum \int_{R_i} \rho_p \leq 2 \int_{E_p} \rho_p.$$

Notice that by Bolzano's theorem we can split every region $R_i$ along a hyperplane through $p$ into $R_i^+$ and $R_i^-$ in a way that $|R_i^+| = |R_i^-|$. Then by Lemma B.5 we get

$$\sum \int_{R_i^\pm} \rho_p \leq \int_{E_p} \rho_p$$

for each of the two families, which concludes the proof. □

**On the role of isoperimetric inequality and concentration of measure** As a final note, let us concede ourselves the following evocative remark, concerning the role of concentration of measure and isoperimetric inequalities in this work. The question of randomly embedding some vectors into a high dimensional vector space might be reminescent of an inverse version of the Johnson-Lindenstrauss lemma. Indeed, in our proof we look at the total measure of a family of regions along a diameter $D$, so arguments involving concentration of measure on the sphere are in the air. But the family we consider in our proof does not contain any neighbourhood $D^\epsilon$ of $D$, so we cannot use arguments involving concentration of measure. On the other hand, the two volume inequalities for convex polytopes on the sphere that we use were strongly inspired by isoperimetric inequalities, and their proof does in fact use a symmetrization argument. In the light of Lévy's first proof of concentration of measure using isoperimetric inequality on the sphere, the presence of these two concepts in the inception of this proof does not seem completely random. While we are not able to express a formal nor organised thought about their relation in this context, we believe it is worth to mention it and invite the reader to bring us their point of view on the matter.

### B.3 Maximal measure of independent regions

We are now ready to prove our main result, which has been reduced to the problem of finding the maximal measure of a family of pairwise non-adjacent regions in a conical tessellation. This problem can be also understood as finding a *weighted maximum independent set* (WMIS) in the adjacency graph of the conical tessellation, weighted by the measure of each region. The WMIS problem is known to be NP-hard for many classes of graphs, including for geometric intersection graphs (Chan and Har-Peled, 2011),(Thomas Erlebach and Seidel, 2005). The fact that this problem is a hard one, makes it more acceptable to find a lower bound that is not optimal. To follow the idea of proof sketched in Subsection 4.1 we need to be able to bound from below the $d$-measure of a region knowing the $d-1$-measure of a face, and the fact that the tessellation is $\theta$-separated, i.e. its faces form angles $\geq \theta$; this is the content of Lemma 4.5. The central ingredient in the proof is a quantitative estimate of the fatness of the face in terms of its measure, which is the content of Lemma B.6.

**Lemma B.6** (Volume-Radius inequality)**.** *Let $R$ be a region of a conical tessellation in $S^m$ with measure $V$. Then its inscribed radius is bounded from below by $r_{in}(R) \geq \pi V/\omega_m$.*

*Proof.* By Theorem B.3 a ditope with measure $V$ has minimal inscribed radius among all spherical convex polytopes of equal measure. So $r_{in}(R) \geq r_{in}(D)$, where $D$ is a ditope with measure $V$. Then the result follows from the fact that a ditope with inscribed radius $r$ has measure $V_r = \omega_m r/\pi$. □

**Lemma B.7** (Surface-Volume inequality)**.** *Let $R$ be a region of a $\theta$-separated conical tessellation in $S^m$, with a face $f$ of $(m-1)$-dimensional measure $|f|$. Then*

$$|R| \geq \frac{|f|h(f)}{m} \geq \frac{2|f|^2\theta}{\omega_{m-1}m},$$

*where $h(f) = \arctan\big(\sin r_{in}(f)\tan\theta\big) \geq \arctan\big(\sin(\frac{\pi|f|}{\omega_{m-1}})\tan\theta\big)$.*
*Notice that the right hand side only depends on the dimension $m$, the angle $\theta$, and the measure $|f|$ of $f$.*

*Proof.* By Lemma B.6 the inscribed radius $r_{in}(f)$ of $f$ is at least $|f|\pi/\omega_{m-1}$. The angle between any two faces of $R$ is equal to the angle between the two corresponding diameters, and since the tessellation is $\theta$-separated, the angle is $\geq \theta$. Then $R$ must contain the cone $C(f,\theta)$ with base the inscribed sphere of $f$ and angle $\theta$.
By A.5, if we denote by $c$ the center of the inscribed sphere and by $v$ the apex of the cone $C(f,\theta)$ its height is given by

$$h(f) \coloneqq d(c,v) = \arctan\big(\sin r_{in}(f)\tan\theta\big).$$

In fact, by convexity, $R$ contains the cone $C = C(f, v)$ from $v$ over $f$ (i.e. the convex hull of $f \cup \{v\}$).
By Theorem B.1 we obtain

$$|R| \geq |C| > |\tilde{C}| = \frac{|f|h(f)}{m}.$$

Finally we derive the last estimate, which has better convexity properties. Notice that by convexity it follows

$$h(f) = \arctan\left(\sin(\frac{\pi|f|}{\omega_{m-1}})\tan\theta\right) \geq \arctan\left(\frac{2|f|}{\omega_{m-1}}\tan\theta\right) \geq \frac{2|f|\theta}{\omega_{m-1}},$$

which concludes the proof. □

**Remark B.8.** *If $k > 2d$ we can obtain a much better bound than the one in Theorem 4.4 by using the fact that*

$$\sum_{i=0}^{d-1}\binom{k-2}{i} \leq 2^{(k-2)H(\frac{d-1}{k-2})},$$

*where $H(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ is the binary entropy function (c.f. (Flum and Grohe, 2006), Lemma 16.19). In particular $H(x) \leq 1$, and we can write*

$$\mathbb{P}_{\mathrm{w}\sim\mathcal{N}^{\otimes d}}\left(\mathrm{rank}(Y') > \mathrm{rank}(Y)\right) > \frac{\theta}{2^{kH(\frac{d-1}{k-2})}}\sqrt{\frac{2}{d\pi}}.$$

# C EXPERIMENTAL SETTING AND FURTHER RESULTS ON RELU VS SIGN ACTIVATION

## C.1 Synthetic dataset generation

We generate point clouds on the unit sphere using two control parameters: minimum angle $\theta_{\min}$ and maximum angle $\theta_{\max}$. Angles are defined between *lines* (i.e., $\theta \in [0°, 90°]$, computed from the absolute value of the cosine), not oriented vectors. Here's the complete sampling procedure:

1. Given target sample size $n$ and space dimension $d$, draw $n$ vectors $\mathrm{x}_i \sim \mathcal{N}(0,1)^d$ and normalize each to unit norm.

2. Select the first vector as the anchor u (angle 0° with itself).

3. Compute the angles $\theta_i = arccos(|\mathrm{x}_i^T\mathrm{u}|)$ between u and all other vectors, then linearly rescale the set $\theta_i$ so that its minimum becomes $\theta_{\min}$ and its maximum becomes $\theta_{\max}/2$. $\theta_{\max}$ is divided by two so that points that have an angle $\theta_{\max}/2$ with u, will have at most an angle of $\theta_{\max}$ between themselves.

4. Reorder the vectors in ascending order of angle from u; u is first. We have the list $V = [\mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3, \ldots, \mathrm{v}_n]$, with $\mathrm{v}_1 = \mathrm{u}$.

5. For each vector $\mathrm{v}_i$, compute its angles $\alpha_{ij}$ with all previously fixed vectors $v_{j,j<i}$. Iterating over $j$, if $\alpha_{ij} \leq \theta_{\min}$, the constraint is respected with $\mathrm{v}_j$. Otherwise, incrementally increase the angle between $\mathrm{v}_i$ and u by applying small orthogonal perturbations and renormalising, and recheck the constraints. Repeat until the constraint is satisfied or $\theta_i$, the angle between $\mathrm{v}_i$ and u surpasses $\theta_{\max}/2$; in this case, drop v.

Note that, because feasible placements can become impossible when $d$ is small, $n$ is large and/or the angular interval is highly restrictive (e.g. when $\theta_{\min}$ is high and $\theta_{\max}$ is low), the algorithm may produce fewer than $n$ points even if a tessellation exists mathematically. In our experiments, we used $d = 128$ and did not encounter any dropped points.

**Code and computing infrastructure** The implementation, parameter choices and code are provided in the referenced repository. All experiments were implemented in Python(Python Software Foundation, 2024) using PyTorch(Paszke et al., 2019). Reproducibility was ensured by fixing random seeds. The code and exact scripts, including seed values, are available at `https://anonymous.4open.science/r/rank_lifting-4524/`.

The experiments were run on a Linux machine (Ubuntu 24.04.2 LTS, kernel 6.14.0-27-generic) with an AMD Ryzen 9 7950X 16-Core Processor, 128 GB of RAM, and two NVIDIA GeForce RTX 4090 GPUs.

Here is the license information of the assets used in this work: Python (Python Software Foundation License, PSF-2.0), PyTorch (BSD 3-Clause License), Fashion-MNIST (MIT License, Zalando SE) (Xiao et al., 2017).

### C.2 FashionMNIST

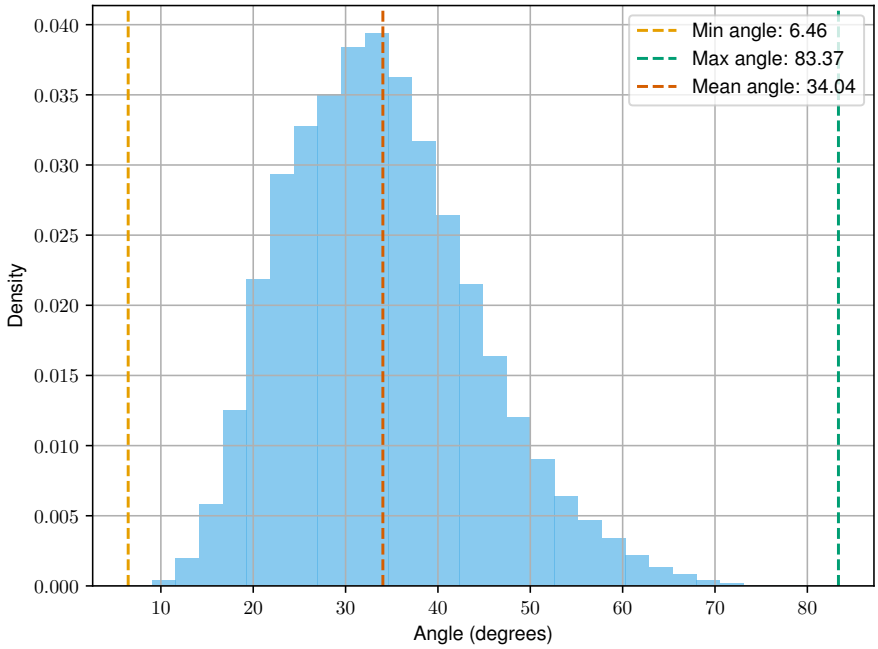


**Figure 8:** *Density of pairwise angles for the sampled subset of* 1024 *FashionMNIST images (flattened). The vertical lines indicate the sample mean, minimum and maximum angles.*

Fig. 8 shows the density of pairwise angles for the 1024 sampled images from FashionMNIST, flattened into a 784-sized vector. The vertical lines in the plot show the sample mean (approximately 35°), and the observed minimum and maximum angles (6.5° and 83.37°, respectively). The distribution is slightly left-skewed, with higher mass at smaller angles.

### C.3 ReLU vs Sign Activation on Hard Instances

Fig. 9 reports the interpolation behavior of networks with Sign activation on hard spherical instances. As the number of points increases, the required width grows rapidly, especially for configurations characterized by smaller minimum pairwise angles. The dependence on angular separation is pronounced, with narrower angular distributions requiring substantially larger widths to achieve perfect interpolation.

Fig. 10 shows the corresponding experiment for ReLU networks. While the qualitative dependence on the number of points and angular separation is similar, the required width is consistently lower across all regimes. In particular, ReLU exhibits a more favorable scaling with respect to both sample size and decreasing angular separation. Please note that the set of angles $[0.1, 0.25]$ is not displayed in Fig. 9 for the sake of graph readability, since it is characterized by a very pronounced growth.

Overall, the comparison highlights a clear expressive advantage of ReLU over Sign activation on hard instances: for the same geometric configuration of the data, ReLU achieves interpolation with significantly fewer neurons.

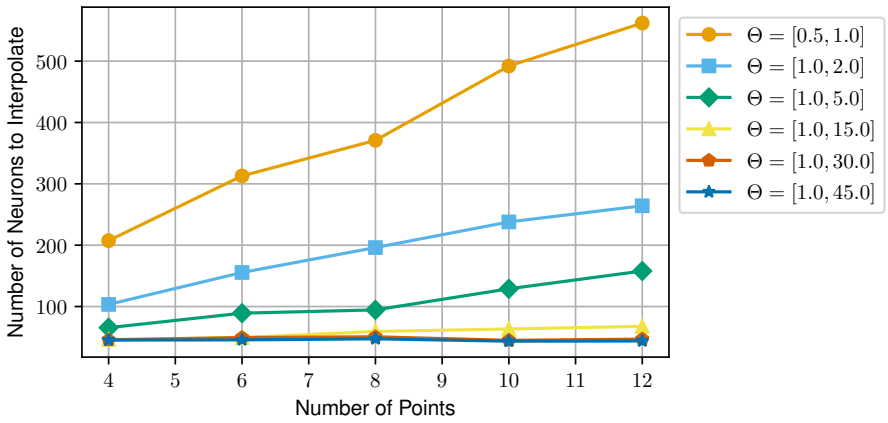


**Figure 9:** *Number of data points vs. required width (synthetic spherical data, $k = 1024$, Sign activation). Plotted is the average number of hidden neurons required to perfectly interpolate the dataset (vertical axis) against the number of data points (horizontal axis), for synthetic point clouds sampled on the unit sphere ($d = 128$). Each curve represents a minimum and maximum angles ($\Theta$).*

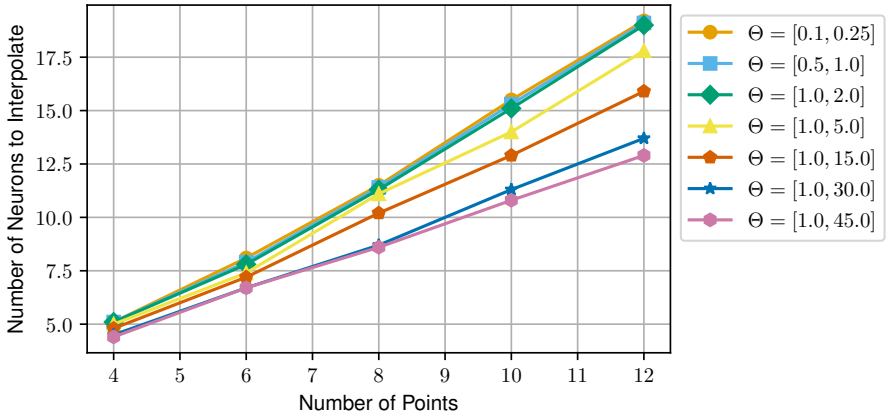


**Figure 10:** *Number of data points vs. required width (synthetic spherical data, $k = 1024$, ReLU activation). Plotted is the average number of hidden neurons required to perfectly interpolate the dataset (vertical axis) against the number of data points (horizontal axis), for synthetic point clouds sampled on the unit sphere ($d = 128$). Each curve represents a minimum and maximum angles ($\Theta$).*

# D RELATED PROBLEMS IN HIGH-DIMENSIONAL PROBABILITY

In this section, we briefly touch on connections to related problems, highlighting some consequences of our results.

## D.1 Covariance matrices

Consider again the representation of input points in feature space: $Y = \text{sign}(WX)$, where $X \in \mathbb{R}^{d\times k}$ is the matrix representation of the input points and $W \in \mathbb{R}^{n\times d}$ is the weight matrix, so that $Y \in \{-1, 1\}^{n\times k}$. In this study, we were interested in $\mathbb{P}\left(\text{rank}(Y) = k\right) = \mathbb{P}\left(\text{span}(Y^T) = \mathbb{R}^k\right)$, where the equality follows since we assume $n \geq k$. Our key technical result to prove this is Lemma 4.1. While the claim of this lemma is formulated consistently with its use in the proof of Theorem 4.2, its proof does not require u's orthogonality to any given subspace. In fact, Lemma 4.1 can be restated in this more general form as follows:

**Lemma D.1.** *Let $\mathrm{x}_1, \ldots, \mathrm{x}_k \in \mathbb{R}^d$ be $\theta$-separated. Let $\mathrm{w}$ be a vector with i.i.d. components $w_j \sim \mathcal{N}(0, 1)$ and let $\mathrm{y} = \mathrm{z}(\mathrm{w}) = \text{sign}(\mathrm{w}^T X)^T$. Then, however one chooses $\mathrm{u} \in S^{k-1}$ we have:*

$$\mathbb{P}\left(\mathrm{y}^T \mathrm{u} \neq 0\right) > \alpha,$$

*where* $\alpha = \frac{\theta}{2^k}\sqrt{\frac{2}{d\pi}}$.

This result immediately implies that the covariance matrix of the generic row of $Y$ is positive definite.

We outline the proof of this obvious fact for the sake of completeness. Consider the generic row of $Y$, i.e., $y = \text{sign}(w^T X)^T$, where $w \sim \mathcal{N}(0, I_d)$. The associated covariance matrix is $A = \mathbb{E}\left[yy^T\right] - \mathbb{E}\left[y\right]\mathbb{E}\left[y\right]^T = \mathbb{E}\left[yy^T\right]$, since the expected value of each entry of y is $\pm 1$ with probability $1/2$. Assume $\mathbb{P}\left(y^T u \neq 0\right) = \mathbb{P}\left((y^T u)^2 > 0\right) = \alpha > 0$ for every $u \in S^{k-1}$. Denote by $S(u)$ the hyperplane orthogonal to u in the remainder of this section and let $m(u) = \min_{y \in \{-1,1\}^k \setminus S(u)} x^T yy^T u$. Finally, given a square matrix $M$, denote by $\lambda_{\min}(M)$ its smallest eigenvalue. Then, for every u we have:

$$u^T A u = u^T \mathbb{E}\left[yy^T\right] u = \sum_{y \in \{-1,1\}^k \setminus S(u)} u^T yy^T u \mathbb{P}\left(y\right)$$
$$\geq \sum_{y \in \{-1,1\}^k \setminus S(u)} m(u)\mathbb{P}\left(y\right) = m(y)\mathbb{P}\left(u^T yy^T u > 0\right) = \alpha m(u),$$

where $m(u) > 0$ by definition. This immediately shows that $\mathbb{E}\left[ww^T\right]$ is positive definite, i.e., $\lambda_{\min}(A) > 0$.

On the other hand, a simple application of Markov's inequality to the random variable $\|u\|^2 - uyy^T u = 1 - uyy^T u$ implies $\mathbb{P}\left((y^T u)^2 = 0\right) \leq 1 - \alpha$, whenever $\lambda_{\min}(A) = \lambda_{\min}(yy^T) \geq \alpha$. Overall, the following holds:

**Fact D.2.** $\mathbb{P}\left(w^T x \neq 0\right) > 0$ *for every* $x \in \mathbb{R}^k \iff \lambda_{\min}(A) > 0$,

### D.2 Covariance matrix and geodesic distance

We next take a closer look at $A = \mathbb{E}\left[yy^T\right]$. The generic entry of $A$ is easy to compute:

$$A_{ij} = \mathbb{E}\left[y_i y_j\right] = \mathbb{E}\left[\text{sign}(w^T x_i) \cdot \text{sign}(w^T x_j)\right] = 1 - 2\frac{\theta_{ij}}{\pi},$$

where the last equality is Grothendieck's equality (Vershynin, 2018, Lemma 3.6.6). It may be useful to highlight two distinct components of $A$:

$$A = J - \frac{2}{\pi}\Theta,$$

where $J = 11^T$ and $\Theta = \{\theta_{ij}\}_{i,j=1}^k$. Here, $J$ has rank one and (largest) eigenvalue $\lambda_{\max}(J) = k$, while $\Theta$ is a *distance* matrix (Steinerberger, 2023), namely, its entries are the pairwise spherical distances between the points $x_1, \ldots, x_k$.

The structure of $\Theta$ provides further insights into the geometric and algebraic aspects of the problem we study in this work. In particular, since $\Theta$ is a distance matrix, it has properties that are not necessarily shared by all symmetric matrices with non-negative entries. For example, the main eigenvector 1 of $J$ is provably close (i.e., it forms a small angle) to the main eigenvector of any distance matrix (Steinerberger, 2023), something that might possibly be leveraged in a spectral approach to this problem.[6] As for $\Theta$ we have:

$$1^T \Theta 1 = \sum_{i,j=1}^k \theta_{ij}.$$

The sum above is maximized when the vectors are distributed as evenly as possible among the coordinate axes of an orthonormal basis in $\mathbb{R}^k$ (Bilyk et al., 2018). In this case, the largest possible value for the sum above is (Bilyk et al., 2018, Theorem 3.1):[7]

$$\sum_{i,j=1}^k \theta_{ij} = \frac{k^2\pi}{2}, \ k \text{ even}, \tag{6}$$

$$\sum_{i,j=1}^k \theta_{ij} = \frac{k^2\pi}{2}\left(1 - \frac{1}{k^2}\right), \ k \text{ odd}. \tag{7}$$

[6] Though outside the scope of this work, it is worth noting that many classes of distance matrices (including geodesic distance matrices) have long been completely characterized in terms of PSD quadratic forms (Schoenberg, 1935).

[7] It should be noted that the authors of Bilyk et al. (2018) use a normalized geodesic distance, with normalization factor $1/\pi$.

The case in which the "true" angles between the vectors are replaced by the acute angles between the corresponding directions is known as the Féjes - Tòth conjecture and is still an open problem. In this case, the corresponding upper bound is conjectured to be $\frac{k^2\pi}{2} \cdot \frac{d}{d+1}$ Bilyk and Matzke (2019) and it has only been proved for specific values of $k$ (e.g., $k = 1, 2$).

It should be noted that the upper bound in (Bilyk et al., 2018, Theorem 3.2) is achieved when the $\mathrm{x}_i$'s form a *centrally symmetric set*, implying the presence of collinear vector pairs, something we exclude. A natural question is therefore:

**Question D.3.** *How do the results from (Bilyk et al., 2018, Theorem 3.2) generalize when vectors share a minimum angle?*

Fact D.2 allows us to prove the following:

**Lemma D.4.** *Assume* $\mathrm{x}_1, \ldots, \mathrm{x}_k \in \mathbb{R}^k$ *are* $\theta$*-separated. Then*

$$\sum_{i,j=1}^{k} \theta_{ij} \leq (k^2 - k\alpha)\frac{\pi}{2},$$

*for some* $\alpha > 0$ *that depends on* $\mathrm{x}_1, \ldots, \mathrm{x}_k$.

*Proof.* The separation condition on the vectors $\mathrm{x}_1, \ldots, \mathrm{x}_k$ implies the results from Section 4 and hence positivity of $A = J - \frac{2}{\pi}\Theta$. Hence, assume that $\lambda_{\min}(A) = \alpha > 0$. In this case we have the following:

$$1^T A 1 = 1^T \left(J - \frac{2}{\pi}\Theta\right) 1 = k^2 - \frac{2}{\pi} \sum_{i,j=1}^{k} \theta_{ij}.$$

But since $\lambda_{\min}(A) = \alpha$ we also have $1^T A 1 \geq k\lambda_{\min}(A) = k\alpha$. Together, these inequalities imply

$$k^2 - \frac{2}{\pi} \sum_{i,j=1}^{k} \theta_{ij} \geq k\alpha,$$

which in turn yields

$$\sum_{i,j=1}^{k} \theta_{ij} \leq (k^2 - k\alpha)\frac{\pi}{2},$$

□

Lemma D.4 thus generalizes (Bilyk et al., 2018, Theorem 3.2) when vectors may not be collinear and (hence) the matrix $A$ is positive definite.