ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Injectivity of ReLU Neural Networks at Initialization

Master Thesis

Daniel Paleka

$26^{\text{th}}$ November 2021

Advisor: Prof. Afonso S. Bandeira

Department of Mathematics, ETH Zürich

**Abstract**

Injectivity of ReLU neural networks plays an important role in generative models and compressed sensing. A natural question is when randomly initialized neural networks are injective.

The recent work Puthawala et al. [2020] has investigated the phase transition of injectivity probability of neural network layers with the ReLU activation. The injectivity depends on the ratio $m/n$ of the output dimension $m$ and the input dimension $n$.

We calculate an expected Euler characteristic surrogate for the injectivity probability in terms of $m$ and $n$, which undergoes a phase transition when $m/n \approx 8.34$. We conjecture that the phase transition for the surrogate is the same as for the injectivity probability.

Moreover, we improve the current upper bound on the injectivity phase transition ratio, and experimentally show the existing lower bounds are not sharp. The new bounds are consistent with the Euler characteristic phase transition matching the injectivity phase transition.

For deep networks, we give the first proof of injectivity of deep neural networks of polynomially bounded width. We additionally connect injectivity of deep networks to well-known contractive phenomena of ReLU networks at initialization.

# Contents

# Acknowledgements

I would like to thank Afonso Bandeira for being a good supervisor and a good person. We interacted significantly more than students and professors usually do. He has great taste in mathematics and a consistent positive attitude to every part of academia, both of which will leave a lasting impression on me. I am grateful for the life and career advice, and for the many future colleagues I have met through his group. I will keep recommending great people to apply to work with him as long as I am in a position to do so.

I would like to thank Charles Clum and Dustin Mixon for collaborating on the original research in this thesis, and specifically for coming up with the Euler characteristic idea in Chapter 4.

I am grateful to Boris Hanin for a great discussion on the number of activation regions in random ReLU networks. I am thankful to Hadi Daneshmand for introducing me to the question of rank collapse and helping me with research internship opportunities. I would like to thank Peter Hinz for helpful tips about the worst-case behaviour of the number of activation regions in ReLU networks, and Petar Nizić-Nikolac for reading a draft of this thesis. I would like to thank Antoine Maillard for a discussion on how to approximate the injectivity probability via the replica method.

I could never have moved to Zurich if not for the financial support of the ETH Zurich Foundation. [1] On a related note, I would like to thank Mislav and Vedran for helping me apply to ETH. I would like to thank Domagoj and Leon for listening to my early ramblings about injectivity. My parents know I'm forever thankful for their unconditional support, and for a warm home to return to for a while when life gets tough.

Finally, Paula, thank you for being here for me.

[1] I had the Master/D-ETH scholarship. As of 2021, the different scholarships have been condensed into the Excellence Scholarship & Opportunity Programme (ESOP).

Chapter 1

# Introduction

Some basic properties of neural networks are yet to be fully understood. A natural question we study in this thesis is: Given a "typical" neural network which goes from $\mathbb{R}^n$ to $\mathbb{R}^m$, is it *injective*?

It turns out this basic question is unsolved even for simple neural network architectures. If we used linear models instead of neural networks, the question would be trivial: an affine map from $\mathbb{R}^n$ is injective if and only if it has rank $n$. But as soon as very simple nonlinearities come into play, we get a hard problem with connections to various areas of mathematics.

The goals of this thesis are:

- to prove several new results about injectivity of ReLU neural networks, both deep and shallow; and
- to highlight connections with well-established mathematical areas related to polyhedral geometry and concentration of measure.

We model a typical neural network as a random function determined by independent Gaussian weights, which is how neural networks are often initialized in theory and practice.

## 1.1 Why injectivity?

As discussed in Puthawala et al. [2020], injectivity of deep neural networks is important in several applications. Invertibility properties of neural networks have been studied in relation to compressed sensing [Bruna et al., 2013, Bora et al., 2017] and generative modeling [Lei et al., 2019]. Normalizing flows [Kobyzev et al., 2021] are widely used bijective neural network models, and any hidden representation in the normalizing flow network must be an injective map from $\mathbb{R}^n$ to $\mathbb{R}^m$ for some $n \leq m$.

Local invertibility and injectivity are not equivalent. In applications it is often enough that the Jacobian is of full rank everywhere, which means the network is invertible around a point in the input space. Injectivity is a stronger property and automatically implies local invertibility: we can think of local invertibility as "injectivity in a small neighbourhood of a point".

When learning neural *representations* of data from $\mathbb{R}^n$ in $\mathbb{R}^m$ for $n \leq m$, injectivity is equivalent to the representation not losing information from the original data. Learned representations are well-known to be transferable across very different learning problems [Yosinski et al., 2014], thus not losing any information from the original data is an important theoretical property.

Finally, injectivity is a basic and interesting property of a function. Tools required for mathematical understanding of injectivity of neural networks could help prove general important properties of neural networks.

## 1.2 Why random networks?

Several phenomena relating to early training of neural networks are explained well by the properties of neural networks at initialization. As initializing with independent weights is standard, neural networks at initialization correspond to our model of random networks in Section 2.2. Notable examples include explanations to why batch normalization and residual connection help early training [Labatie, 2019, Daneshmand et al., 2020], and vanishing and exploding gradients [Hanin and Nica, 2018].

In addition, networks with randomized weights are actually an useful object on their own: untrained neural networks contain "lottery ticket" subnetworks which approximate arbitrary functions without training [Malach et al., 2020, Ramanujan et al., 2020]. Randomly initialized networks have uses in image denoising [Ulyanov et al., 2020] and compressed sensing [Heckel and Soltanolkotabi, 2020].

Although networks that interpolate random data (as opposed to networks with random parameters) have been the main random neural network object to study for many years, there is a resurgent interest in randomly initialized networks [Gallicchio and Scardapane, 2020, Schoenholz et al., 2017, Hanin and Rolnick, 2018].

The recent literature on the Neural Tangent Kernel approximation [Jacot et al., 2018] has given great incentive to study neural networks with random weights. Many phenomena about neural networks can be understood in the regime close to the initial random initialization.

Recent research has been concerned with training just the final layers of a neural network. This is motivated by the fact that, in very wide networks with a low-dimensional output, gradient descent often leaves the weights in

all layers but the final one essentially fixed [Jacot et al., 2018, Chizat et al., 2020]. Thus it makes sense to fix the randomly generated weights in the expanding layers and optimize only the final layer. This is equivalent to a *random features* model, which has known limitations [Yehudai and Shamir, 2020]. The injectivity questions in our paper are equivalent to asking when the random features representation loses information, in terms of the number of neurons in each layer.

## 1.3 Our results

### 1.3.1 Layerwise results

As Puthawala et al. [2020] have shown, the probability of injectivity of a single neural network layer with $n$ input and $m$ output neurons undergoes a "phase transition" depending on the ratio $m/n$. Thus it makes sense to consider the "injectivity threshold" of the ratio $m/n$, under which the injectivity probability goes to zero, and over which the injectivity probability goes to 1. Our Theorem 3.12 improves upon their upper bound for the injectivity threshold. In Section 3.3, we experimentally show that their lower bound is far from sharp.

Calculating the injectivity probability in terms of $n$ and $m$ is not feasible with current methods. We use tools inspired by stochastic polyhedral geometry to approximate the injectivity probability with the expectation of the Euler characteristic of the intersection of a particular random cone with the sphere. In Section 4.1.1, we argue that the approximation may preserve the phase transition. This "Euler characteristic heuristic" implies the injectivity threshold for the ratio $m/n$ could be around 8.34.

### 1.3.2 Multilayer results

In Chapter 5, we give the first proof that deep ReLU networks can be injective at initialization without the hidden layer widths expanding exponentially. To be precise, Theorem 5.1 implies a network with the first layer of width $n$ and the remaining $L$ layers of width $Cn$ is injective with high probability, when $C \gtrsim L \log L$. An intermediate result which may be of independent interest is Corollary 5.7, which gives a new characterization of injectivity in deep random ReLU networks.

Moreover, we conjecture the optimal expansivity should not depend on the depth at all. This is due to the known contractive properties of deep ReLU networks at initialization, which we call *angle convergence*. In Proposition 5.14, we prove injectivity for "uniformly contractive" networks, in the sense of deep preactivations of all possible inputs being contained in a small angle.

## 1.4 Related work

The most relevant prior work for the layerwise injectivity phase transition is Puthawala et al. [2020], which gives various motivations for the precise phrasing of the layerwise injectivity question. In particular, their Lemma 1 implies the injectivity of a layer with a bias term is not fundamentally a different problem than the injectivity of biasless layers we consider in this work. Moreover, they show that batch, weight and spectral normalizations do not influence injectivity.

The Euler characteristic surrogate for the injectivity probability in Chapter 4 is motivated by several works on intrinsic volumes and the statistical dimension [Schneider and Weil, 2008, Amelunxen et al., 2013, Amelunxen and Lotz, 2017]. In Proposition 4.14, we give an explicit formula for the intrinsic volumes of the nonconvex union of orthants in $\mathbb{R}^m$ with at most $n$ positive coordinates, which may be of independent interest for stochastic geometry research.

Regarding deep networks, Puthawala et al. [2020] uses Whitney's embedding theorem and random projections to prove there exist randomly-initialized injective deep ReLU networks with end-to-end expansivity 2. However, the given initialization is nonstandard; in particular, the weight matrices are random matrices of low rank, and therefore do not correspond to the theoretical model of random neural networks usually studied in the literature.

Chapter 2

# Notation and background

## 2.1 ReLU neural networks

Let $n$ and $m$ be positive integers. In the machine learning literature, the simplest example of a neural network from $\mathbb{R}^n$ to $\mathbb{R}$ is of the form

$$x \mapsto \sum_{i=1}^{m} a_i \, \sigma(w_i^T x), \tag{2.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an *activation function*, and $(w_i)_{1 \le i \le m} \in \mathbb{R}^n$ and $a = (a_i)_{1 \le i \le m} \in \mathbb{R}$ are *weights*. The weights $w_i$ are stacked in a *weight matrix* $W \in \mathbb{R}^{m \times n}$.

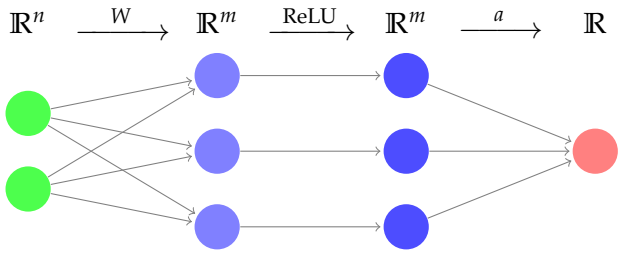
In this work, the only activation we consider is the piecewise linear $\sigma = \mathrm{ReLU}$, the *rectified linear unit*:

$$\mathrm{ReLU}(x) = \max\{x, 0\} \text{ for all } x \in \mathbb{R}. \tag{2.2}$$

It is by far the most widely used activation in both theory and practice, owing to its simple mathematical form, 1-Lipschitzness, and very fast evaluation on standard computer architectures. We use the same notation to denote the pointwise application $\mathrm{ReLU} : \mathbb{R}^m \to \mathbb{R}^m$:

$$\mathrm{ReLU}(x)_i = \max\{x_i, 0\} \text{ for all } x \in \mathbb{R}^m \text{ and } 1 \le i \le m. \tag{2.3}$$

We can draw Equation (2.1) as a composite mapping:



The vectors of dimension $m$ in the above figure are usually called *hidden preactivations* and *hidden activations*. The sets of arrows corresponding to $W$ and $a$ are called *affine layers*, and similarly the pointwise activation arrows are called the *ReLU layer*. The mapping $x \mapsto \mathrm{ReLU}(Wx)$ given by the first two arrows in the above diagram is called a *(hidden) representation*.

Deep neural networks have multiple iterations of the affine and ReLU layers, with varying dimensions of layers. Conceptually, only the output of the final ReLU layer is considered a representation. This is because the last affine layer (going to $\mathbb{R}$) can be considered as a linear classifier on a complicated feature embedding of the input. The composition of all layers but the final affine layer "represents" the input in the feature space $\mathbb{R}^m$, and the final layer gives a score based on a simple linear combination of the representation neurons.

The network in Equation (2.1) is homogeneous in $x$, thus it cannot approximate general functions when varying $W$ and $a$. This issue is often alleviated by adding *biases* $b \in \mathbb{R}^m$:

$$x \mapsto \sum_{i=1}^{m} a_i \, \sigma(w_i^T x + b_i). \tag{2.4}$$

The network in Equation (2.4) is obviously not injective, because the input dimension is higher than the output dimension. When discussing injectivity of neural networks, we only care about networks that expand the dimension of the input. Henceforth, when we talk about injectivity of neural networks, we actually want injective representations. As discussed in Section 1.1, injective representations are important because they do not lose information from the input space.

## 2.2 Deep random neural networks

Let $L \geq 1$ be the number of layers, or *depth* of the network. Let $d_0, \ldots, d_L$ be positive integers denoting the layer widths. For all $1 \leq \ell \leq L$ define the affine layers $B^{(\ell)}(x) = W^{(\ell)}x + b^{(\ell)}$ as affine maps from $\mathbb{R}^{d_{\ell-1}}$ to $\mathbb{R}^{d_\ell}$. Then a neural network $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$ is given by

$$f = \mathrm{ReLU} \circ B^{(L)} \circ \ldots \circ \mathrm{ReLU} \circ B^{(1)}. \tag{2.5}$$

If $L = 1$ and $b^{(1)} = 0$, then $f$ takes the simple form $x \mapsto \mathrm{ReLU}(Wx)$.

We assume that the entries of the matrices $W^{(\ell)}$ are independent Gaussians with mean 0 and variance 1. This is an often-used assumption in the literature, and is standard in injectivity-related work [Puthawala et al., 2020, Bruna et al., 2013]. A different often-used approach is to initialize weights with variance depending on the layer width [He et al., 2015]. In the case of biasless networks with $b^{(\ell)} = 0$, this has no impact on injectivity.

As discussed above, we omit the conventional last affine map in the definition of a neural network. This omission does not matter as long as we are interested in high-dimensional representations, because injectivity is unaffected by a random affine embedding. All networks we consider in this work end with a ReLU layer.

## 2.3 Orthants

An *orthant* in $\mathbb{R}^d$ is a generalization of the concept of a quadrant in $\mathbb{R}^2$, or a ray from the origin in $\mathbb{R}$. In our proofs, we will use two subtly different concepts of an orthant.

**Definition 2.1** Given a set $S \subseteq \{1, 2, \ldots, d\}$, a *half-open orthant* $O_S^d$ is the set

$$O_S^d \stackrel{\text{def}}{=} \{(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d : x_i > 0 \text{ for all } i \in S;\ x_i \leq 0 \text{ for all } i \notin S\}. \tag{2.6}$$

The half-open orthants in $\mathbb{R}$ are the intervals $\langle -\infty, 0]$ and $\langle 0, \infty\rangle$. The full space $\mathbb{R}^d$ is the disjoint union of $2^d$ half-open orthants, indexed by all $2^d$ subsets of $\{1, 2, \ldots, d\}$.

We define half-open orthants this way because of the following property:

**Proposition 2.2** For a half-open orthant $O_S^m$, if $x \in O_S^m$, then $\mathrm{ReLU}(x) \in O_S^m$.

A closed orthant is the topological closure of a half-open orthant:

**Definition 2.3** Given a set $S \subseteq \{1, 2, \ldots, d\}$, a *closed orthant* $\mathcal{O}_S^d$ is the set

$$\mathcal{O}_S^d \stackrel{\text{def}}{=} \{(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d : x_i \geq 0 \text{ for all } i \in S;\ x_i \leq 0 \text{ for all } i \notin S\}. \tag{2.7}$$

We will use the word "orthant" for both concepts interchangeably if it is clear from the context. The notations $O$ and $\mathcal{O}$ are always used for half-open and closed orthants, respectively.

**Definition 2.4** Given an orthant $O_S^m$ or $\mathcal{O}_S^m$, we say it *has k pluses* if $|S| = k$.

## 2.4 Miscellaneous notation

For a vector $z \in \mathbb{R}^d$, we define $z_{\text{pos}} = \mathrm{ReLU}(z)$ and $z_{\text{neg}} = z - z_{\text{pos}}$, the positive and negative parts of $z$.

For nonnegative integers $a$ and $b$, we use the notation $T(a, b)$ for the *prefix sum of binomial coefficients*:

$$T(a, b) = \sum_{i=0}^{b} \binom{a}{i}. \tag{2.8}$$

Our convention is that $\binom{a}{b} = 0$ whenever $b \notin \{0, 1, \ldots, a\}$.

The *binary entropy function* $H : [0,1] \to \mathbb{R}$ is defined as

$$H(p) = -p \log_2(p) - (1-p) \log_2(1-p). \tag{2.9}$$

In combinatorial calculations in Section 4.3, we use the machinery of generating functions, which are formal power series in a single variable. We never evaluate any generating function at a point, so all generating functions are just sequences of their coefficients.

For a generating function

$$A(x) = \sum_{k=-\infty}^{\infty} a_k x^k, \tag{2.10}$$

we denote "taking its $k$-th coefficient" by $\left[x^k\right] A(x) \stackrel{\text{def}}{=} a_k$. All standard arithmetic operations on rational functions translate analogously to the algebra of formal power series. We interpret the derivative of a power series in the formal sense:

$$A'(x) = \sum_{k=-\infty}^{\infty} k a_k x^{k-1}. \tag{2.11}$$

Chapter 3

# The layerwise injectivity threshold

## 3.1 Characterizing injectivity

Puthawala et al. [2020] introduce the notion of a *directed spanning set* of a $m \times n$ matrix with respect to a vector $x \in \mathbb{R}^n$. [1]

**Definition 3.1** For $m \geq n$, let $W \in \mathbb{R}^{m \times n}$. We say that $W$ has a *directed spanning set* with respect to $x \in \mathbb{R}^n$ if $W$ has $n$ linearly independent rows having nonnegative inner product with $x$.

Using directed spanning sets, they characterize injectivity of fully connected layers with arbitrary weights:

**Theorem 3.2 (Puthawala et al. [2020])** For $m \geq n$, let $W \in \mathbb{R}^{m \times n}$ be a matrix. The function $x \mapsto \mathrm{ReLU}(Wx)$ is injective if and only if $W$ has a directed spanning set with respect to all $x \in \mathbb{R}^n$.

As all matrices we consider in this work have independent Gaussian entries, we may simplify Theorem 3.2. We first note that the image of such a matrix is a random subspace.

**Proposition 3.3** Let $m \geq n$. If the rows of $W \in \mathbb{R}^{m \times n}$ are independent standard normal vectors, then the subspace $W\mathbb{R}^n \subseteq \mathbb{R}^m$ is a uniformly random $n$-dimensional subspace [2] of $\mathbb{R}^m$.

**Proof** The $n$ basis vectors $(e_i)_{1 \leq i \leq n}$ in $\mathbb{R}^n$ are mapped to $n$ independent standard normal vectors in $\mathbb{R}^m$. Those are linearly independent almost surely, and rotationally invariant, hence they form a basis of a uniformly random subspace. □

[1] Their Definition 1 has a typo: $n \times m$ should be $m \times n$.

[2] Formally, it follows the uniform Haar measure on the Grassmanian $\mathrm{Gr}_{n,m}$. This produces the same distribution as as any intuitive way to randomly sample a subspace.

Then Theorem 3.2 reduces to the following statement, which connects random polyhedral geometry with injectivity.

**Theorem 3.4** For $m \geq n$, let $W \in \mathbb{R}^{m \times n}$ be a matrix with independent $N(0,1)$ entries. The function $x \mapsto \mathrm{ReLU}(Wx)$ is injective if and only if the subspace $W\mathbb{R}^n$ does not contain a vector with less than $n$ nonnegative coordinates.

We are interested in the probability that the function $x \mapsto \mathrm{ReLU}(Wx)$ is injective. As the distribution over subspaces is atomless, we can use "nonnegative coordinates" and "positive coordinates" interchangeably, almost surely.

**Definition 3.5** Let $C_{m,n}$ be the set of vectors in $\mathbb{R}^m$ with strictly less than $n$ positive coordinates. Thus, $C_{m,n}$ is the union of $T(m, n-1) = \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{n-1}$ orthants in $\mathbb{R}^m$.

**Theorem 3.6** Let $V$ be a random $n$-dimensional subspace of $\mathbb{R}^m$. The probability of $x \mapsto \mathrm{ReLU}(Wx)$ being injective is equal to

$$p_{m,n} \stackrel{\text{def}}{=} \mathbb{P}\left[V \cap C_{m,n} = \{0\}\right]. \tag{3.1}$$

**Proof** Note that the event "the subspace $V$ has some coordinate constantly zero" is an event of measure zero. Then the statement follows directly from Theorem 3.4 and Proposition 3.3. □

**Remark 3.7** The probability $p_{m,n}$ is equal to $\mathcal{I}(m,n)$ from Puthawala et al. [2020]. We use a different notation because we focus on the random polyhedral geometry characterization of this probability.

As proved in Puthawala et al. [2020], the probability of injectivity when $n$ and $m$ go to infinity is governed by the ratio $m/n$. In fact, there is a phase transition happening: as we increase the ratio $m/n$, the random neural network layer goes from w.h.p. not injective to w.h.p. injective.

There is a trivial lower bound on $m/n$ in order for the map $x \mapsto \mathrm{ReLU}(Wx)$ to be injective:

**Proposition 3.8** When $m < 2n$, $p_{m,n} = 0$.

**Proof** For any $x \in \mathbb{R}^n$, the negative coordinates of $x$ are positive coordinates of $-x$. Hence when $m < 2n$, at least one of $x$ and $-x$ is in the set $C_{m,n}$. □

**Remark 3.9** Some machine learning intuition: if $m < 2n$, taking a random input and weights, the number of "alive" ReLU activations will be less than $n$ on average. This means that there is a large chance of the Jacobian being of rank less than $n$ on a random input, hence the layer is not locally invertible.

The interesting regime is when $m = \Theta(n)$, and there are explicit bounds on the ratio $m/n$ with regards to injectivity:

**Theorem 3.10** [Puthawala et al. [2020]] If $m \geq 10.5n$, then $p_{m,n} \to 1$ as $n$ goes to $\infty$.

**Theorem 3.11** [Puthawala et al. [2020]] If $m \leq 3.4n$, then $p_{m,n} \to 0$ as $n$ goes to $\infty$. In fact, with high probability there is an explicit "noninjectivity certificate": $x \mapsto \mathrm{ReLU}(Wx)$ doesn't have full rank at $x = -\sum_{i=1}^{m} w_j$, where $(w_j)_{1 \leq j \leq m}$ are the rows of the matrix $W$.

We may think of an "injectivity threshold" for $m/n$, under which $p_{m,n}$ goes to zero, and over which $p_{m,n}$ goes to 1. In Chapter 4, we will use advanced random polyhedral geometry tools in a heuristic derivation of what the injectivity threshold should be, via the characterization in Theorem 3.4. The conjectured threshold is close to $m = 8.34n$.

We do not make any formal claims about the sharpness of the phase transition. Although it would be unusual if there was a constant-sized region of the ratio $m/n$ where $p_{m,n}$ was neither very large nor very small, we do not yet have the tools to prove this rigorously.

## 3.2 Upper bound on the injectivity threshold

In this section, we improve the bound in Theorem 3.10 via a new proof, using the characterization from Theorem 3.4.

**Theorem 3.12** If $m \geq 9.09n$, then $p_{m,n} \to 1$ as $n$ goes to $\infty$.

**Proof** For a fixed orthant $\mathcal{O} \subseteq \mathbb{R}^m$, the probability of a random subspace of dimension $n$ intersecting $\mathcal{O}$ nontrivially is exactly

$$\xi(m,n) \stackrel{\text{def}}{=} \frac{1}{2^{m-1}} \sum_{i=0}^{n-1} \binom{m-1}{i} = \frac{1}{2^{m-1}} T(m-1, n-1); \tag{3.2}$$

see Lemma 6.6.

We union bound over all orthants that $C_{m,n}$ is made of:

$$\begin{aligned}
\mathbb{P}\left[V \cap C_{m,n} \neq \{0\}\right] &\leq \sum_{k=0}^{n-1} \binom{m}{k} \xi(m,n) && (3.3)\\
&= \frac{1}{2^{m-1}} T(m-1, n-1) T(m, n-1) && (3.4)\\
&< \frac{1}{2^m} T(m,n)^2 && (3.5)\\
&\leq \frac{1}{2^m} 2^{2mH(n/m)} = 2^{-n(c-2cH(1/c))}, && (3.6)
\end{aligned}$$

where we used the well-known bounds on $T(a,b)$ via the binary entropy function $H$; see Appendix A.1 for the details. For $c \geq 9.09$, the probability of noninjectivity decays exponentially in $n$. □

## 3.3 Lower bound on the injectivity threshold

Puthawala et al. [2020] prove a lower bound on the injectivity threshold by exhibiting an explicit $x \in \mathbb{R}^n$ where the network is not locally injective with high probability.

**Theorem 3.11** [Puthawala et al. [2020]] If $m \leq 3.4n$, then $p_{m,n} \to 0$ as $n$ goes to $\infty$. In fact, with high probability there is an explicit "noninjectivity certificate": $x \mapsto \mathrm{ReLU}(Wx)$ doesn't have full rank at $x = -\sum_{i=1}^{m} w_j$, where $(w_j)_{1 \leq j \leq m}$ are the rows of the matrix $W$.

We present the basic idea behind the proof.

**Proof (Outline)** Consider a single dot product $w_i^T(-x)$ for $1 \leq i \leq m$. It is distributed as

$$-w_i^T x = \|w_i\|_2^2 + w_i^T \sum_{j \neq i} w_j = \|w_i\|_2^2 + \|w_i\|_2 Y, \tag{3.7}$$

where $Y \sim N(0, m-1)$, because the sum $\sum_{j \neq i} w_j$ is a sum of $m-1$ Gaussian vectors independent of $w_i$.

It is standard to prove $\|w_i\|_2$ concentrates well around $\sqrt{n}$, hence the above expression is positive approximately whenever some $N(0,m)$ variable is smaller than $-\sqrt{n}$.

On average, we thus expect around $\Phi(-\sqrt{n/m})$ fraction of rows to have a positive dot product with $x$, where $\Phi$ is the cumulative distribution function of a $N(0,1)$ variable. We have noninjectivity whenever this fraction is smaller than $n/m$, and the solution for $1/c = \Phi(-\sqrt{1/c})$ is around $c = 3.4$.

After filling in some technical details [3] and standard computations, for $m \leq 3.4n$ we get $p_{m,n} \to 0$ as $n$ goes to $\infty$. □

The intuition behind their noninjectivity certificate is the following: we expect $x = -\sum_{i=1}^{m} w_i$ to correlate negatively with more than half of the rows $w_i$, given that they are independent standard normal vectors. It turns out that for $m \leq 3.4n$, we hit less than $n$ positive inner products.

The proof of Theorem 3.11 has a simple main idea: construct $x$ such that the dot products $w_i^T x$ provably have negative mean, and then use concentration inequalities. One may think of a better expression for the noninjectivity certificate. For example, we can try $x = -\sum_i \frac{w_i}{\|w_i\|}$, which has similar concentration properties, and produces certificates up to $m = 3.5n$ for $n \leq 100$.

[3] One needs careful tail bounds because we want noninjectivity with high probability, and the dot products above are not independent. The main ingredients are concentration of the dot product of two independent standard normal vectors, and the union bound.

We give experimental evidence that this is not the right approach. More precisely, we solve a mixed-integer linear program for the optimal $x$, using the Python API of the state-of-the art solver Gurobi [Gurobi Optimization, 2021]. Here, the optimal $x$ is a vector which has positive inner product with most rows of the matrix $W$.

Consider the following problem:

$$\begin{aligned} \max_{x\in\mathbb{R}^n, a\in\mathbb{R}^m} \quad & \sum_{i=1}^{m} a_i && \text{(3.8a)} \\ \text{subject to} \quad & \|x\|_1 = n && \text{(3.8b)} \\ & -U(1-a_i) \le w_i^T x \le U a_i \text{ for } 1 \le i \le m && \text{(3.8c)} \\ & a_i \in \{0,1\} && \text{(3.8d)} \end{aligned}$$

for some large $U > 0$. It is easy to see the solution $x$ to the above program maximizes the number of positive inner products with the rows of $W$. If the objective is at least $m - n + 1$, the function $x \mapsto \mathrm{ReLU}(Wx)$ is locally noninjective at $-x$.

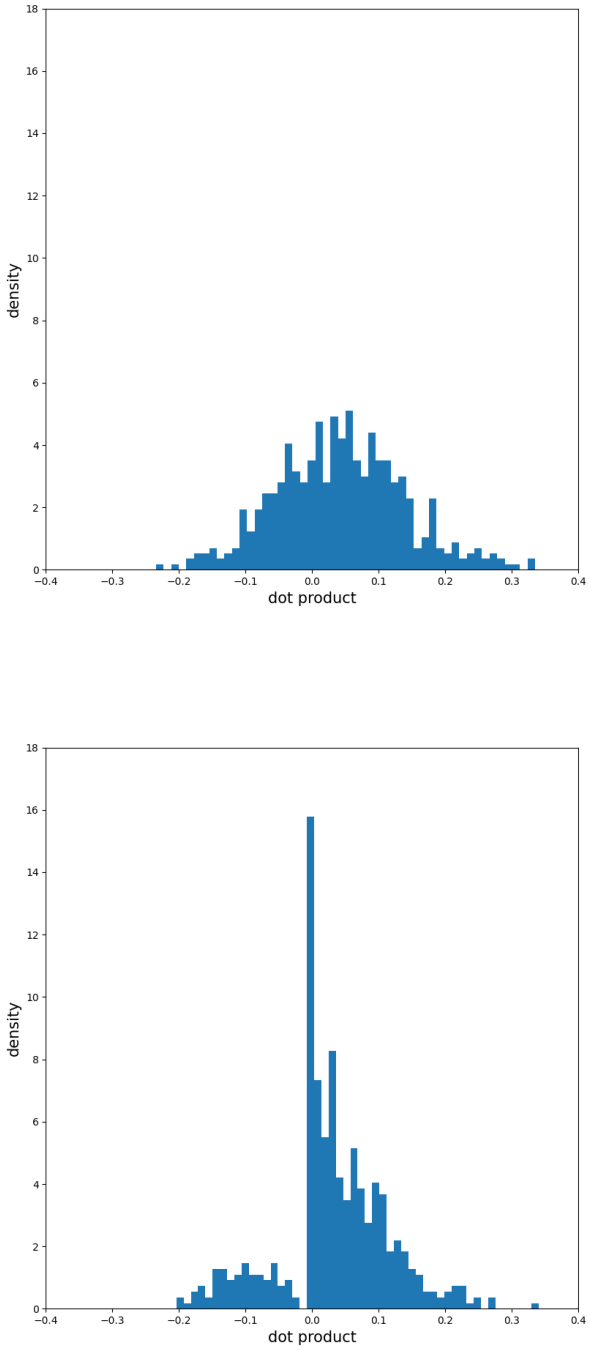
As mixed-integer programs cannot generally be solved in polynomial time, finding the optimal $x$ for $n \ge 20$ is too slow. But, as Gurobi returns *incumbent solutions* as it finds them in the branch-and-bound algorithm, we can still produce noninjectivity certificates. We empirically find that one-layer random networks $\mathbb{R}^n \to \mathbb{R}^m$ with $n = 100$ and $m = 500$ are not injective. Gurobi finds noninjectivity certificates in under a minute on an Intel Core i7 processor. This strongly suggests the injectivity threshold is greater than 5.

Moreover, the near-optimal $x$ are very different than the sum-of-rows certificate: see Figure 3.1. With the sum-of-rows certificate, the dot products follow a unimodal distribution with nice tails. With the near-optimal certificate, the inner product distribution is bimodal, with a discontinuity below zero.

It is easy to see finding the optimal injectivity certificate $x$ for a fixed matrix $W$ is equivalent to the following problem: *Given $m$ points on $\mathbb{S}^{n-1}$, find the halfsphere with the least number of points.* The question of the injectivity threshold is then equivalent to:

**Problem 3.13** Sample $m = cn$ points independently from the uniform measure on $\mathbb{S}^{n-1}$. What is the threshold for $c$ such that there is a halfsphere with less than $n$ points with high probability?

In this formulation, it is easier to believe Figure 3.1. Indeed, we are in a high-dimensional optimization regime, and random things do not behave as in the low-dimensional regime [Wainwright, 2019]. The number of points is linear in the dimension, thus concentration of measure is likely not an important factor in the optimal solution.




a) The sum-of-rows certificate.
b) The certificate from the mixed-integer program.

**Figure 3.1:** Comparing different certificates via the distribution of inner products with the row vectors. Here $n = 100$ and $m = 500$.

Chapter 4

# The Euler characteristic heuristic for the injectivity threshold

Parts of this chapter are joint work with Charles Clum. The unpublished notes [Clum, 2021a] serve as the basis for Section 4.2.

Parts of this chapter will feature in the PhD thesis [Clum, 2021b].

## 4.1 Intrinsic volumes and the Crofton formula

Recall a definition and a theorem from Chapter 3:

**Definition 3.5** Let $C_{m,n}$ be the set of vectors in $\mathbb{R}^m$ with strictly less than $n$ positive coordinates. Thus, $C_{m,n}$ is the union of $T(m, n-1) = \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{n-1}$ orthants in $\mathbb{R}^m$.

**Theorem 3.6** Let $V$ be a random $n$-dimensional subspace of $\mathbb{R}^m$. The probability of $x \mapsto \mathrm{ReLU}(Wx)$ being injective is equal to

$$p_{m,n} \stackrel{\text{def}}{=} \mathbb{P}\left[V \cap C_{m,n} = \{0\}\right]. \tag{3.1}$$

To calculate the probability in Equation (3.1), we will need tools applicable to intersections of subspaces and general cones.

**Definition 4.1** A (closed) *cone* $C \subset \mathbb{R}^d$ is a nonempty (closed) set with the following property: if $x \in C$, then $\lambda x \in C$ for all $\lambda \geq 0$.

For example, a quadrant in $\mathbb{R}^2$ is a cone; same with orthants in $\mathbb{R}^d$. Any subspace of $\mathbb{R}^d$ is a cone. Cones need not be convex; the union of the coordinate axes in $\mathbb{R}^2$ is a nonconvex cone. Cones need not "go to infinity"; the only counterexample is the cone $\{0\}$.

In our work, we consider only polyhedral cones, which are cones that are unions of polyhedra. Polyhedral cones are closed by definition.

**Definition 4.2** A *convex polyhedral cone* is a cone that can be written as the intersection of a finite number of halfspaces. A *polyhedral cone* is a finite union of convex polyhedral cones.

The injectivity threshold of $x \mapsto \mathrm{ReLU}(Wx)$ is connected with random polyhedral geometry through Theorem 3.6. The seminal paper Amelunxen et al. [2013] focuses on a very similar question:

**Problem 4.3** Let $C$ and $K$ be *convex* polyhedral cones in $\mathbb{R}^d$. Draw a random orthogonal basis $\mathbf{Q} \in \mathbb{R}^{d\times d}$. What is the probability

$$\mathbb{P}\left[C \cap \mathbf{Q}K\right] \tag{4.1}$$

in terms of the cones $C$ and $K$?

Note that if we let $K$ to be a subspace, we recover almost the same problem as in Theorem 3.6, except the cone there is not convex. The motivation of Amelunxen et al. [2013] does not have anything to do with injectivity; they investigate phase transitions of random convex optimization problems.

In this work, we need the notion of the *intrinsic volumes* of a cone. The standard reference for this is Chapter 6 of Schneider and Weil [2008]. We will also make use of the exposition in Amelunxen and Lotz [2017].

**Definition 4.4** For a convex polyhedral cone $C \subset \mathbb{R}^d$, define $\mathrm{span}(C)$ to be the smallest subspace of $\mathbb{R}^d$ containing $C$. The *faces* of $C$ are the intersections of $C$ with supporting hyperplanes, and $C$ itself. For example, a quadrant in $\mathbb{R}^2$ has one face of dimension 0, two faces of dimension 1, and one face of dimension 2.

The *relative interior* of a face $F$ is the (topological) interior of $F$ in $\mathrm{span}(C)$.

The projection of any point $x$ to any convex closed $C \subset \mathbb{R}^d$ is well-defined as the unique point $y \in C$ for which $\|x - y\|_2$ is minimal.

**Definition 4.5** Let $0 \leq k \leq d$. Let $g \sim N(0, I_d)$ be a standard normal vector in $\mathbb{R}^d$. The *intrinsic volume of dimension* $k$ of a *convex* polyhedral cone $C \subset \mathbb{R}^d$, denoted $v_k(C)$, is the probability that the projection of $g$ to $C$ lands in the relative interior of a face of dimension $k$.

Due to the scaling property of cones in Definition 4.1, sampling $g$ from the sphere $\mathbb{S}^{d-1}$ results in the same probability as when $g$ is a Gaussian vector. In general, we can use any rotationally invariant measure on $\mathbb{R}^d$.

**Example 4.6** Let us calculate the intrinsic volumes of a quadrant $Q$ in $\mathbb{R}^d$. See Figure 4.1 for an illustration.

The points with both coordinates positive are projected to themselves, and this is the only way a point lands in a face of dimension 2, thus $v_2(Q) = \frac{1}{4}$.

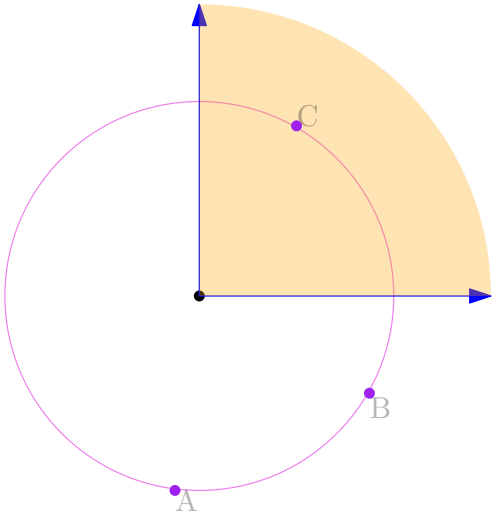



**Figure 4.1:** The point $A$ is projected to the unique face of dimension 0, which is the origin. The point $B$ is projected to the $x$-axis, and the point $C$ is projected to the unique face of dimension 2, the cone itself.

The points with exactly one positive coordinate are projected to the relative interiors of the corresponding axes, which are faces of dimension 1. Hence, $v_1(Q) = \frac{1}{2}$. Finally, the points with negative coordinates are projected to the origin, which makes $v_0(Q) = \frac{1}{4}$.

**Remark 4.7** Note that the probabilities $v_0(C), v_1(C), \ldots, v_d(C)$ form a probability distribution on $\{0, 1, \ldots, d\}$. The *statistical dimension* of Amelunxen et al. [2013] is defined as the mean of this distribution. It has some very useful properties, which are outside the scope of this work. We just note that the statistical dimension appears in answers to questions similar to Problem 4.3, and that it is the correct way to generalize the notion of dimension of a subspace to arbitrary convex cones.

For general polyhedral cones as in Definition 4.2, the intrinsic volumes are not defined via the projection probability as in Definition 4.5. The correct generalization is the following:

**Definition 4.8** For $k \geq 1$, the intrinsic volumes of a polyhedral cone $C = \bigcup_{i \in I} C_I$, where $(C_i)_{i \in I}$ are convex polyhedral cones, are defined by inclusion-exclusion:

$$v_k(C) = \sum_{\varnothing \neq J \subseteq I} (-1)^{|J|+1} \, v_k\left(\bigcap_{i \in J} C_i\right). \tag{4.2}$$

The inclusion-exclusion comes from the valuation properties of $v_k$ and the uniqueness of the extension of any valuation on convex sets to the *convex ring*. Note that this definition does not correspond to the probability of a

projection of a Gaussian vector landing in a face of dimension $k$ of a general nonconvex cone. We discuss this further in Appendix B.2.

The tool used in Amelunxen et al. [2013] and Amelunxen and Lotz [2017] to solve problems such as Problem 4.3 is the following formula:

**Theorem 4.9** [Kinematic Crofton formula] Let $C \subset \mathbb{R}^m$ be a convex cone. Let $V$ be a random $n$-dimensional subspace of $\mathbb{R}^m$. Then

$$\mathbb{P}\left[V \cap C \neq \{0\}\right] = 2 \sum_{i=0}^{\lfloor \frac{n-1}{2} \rfloor} v_{m-n+2i+1}(C). \tag{4.3}$$

This is not yet useful in the context of injectivity, as Theorem 4.9 cannot be applied to the probability in Theorem 3.6 because the cone $C_{m,n}$ is non-convex.

**Proposition 4.10** Let $C \subset \mathbb{R}^m$ be a cone. Let $V$ be a random $n$-dimensional subspace of $\mathbb{R}^m$. The probability $p_{m,n}$ from Equation (3.1) is equal to

$$p_{m,n} = 1 - \mathbb{P}\left[V \cap (C \cap \mathbb{S}^{m-1}) \neq \varnothing\right] = 1 - \mathbb{E}\left[\mathbb{1}^s(V \cap C)\right], \tag{4.4}$$

where $\mathbb{1}^s$ is the spherical indicator function, defined as

$$\mathbb{1}^s(A) = \begin{cases} 1 & \text{if } A \cap \mathbb{S}^{m-1} \neq \varnothing; \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

**Proof** Any intersection of cones is a cone, thus $V \cap C$ is always a cone. It intersects the sphere if and only if it is not equal to the trivial cone $\{0\}$. The probability of an event is equal to the expectation of its indicator function.□

The *spherical cinematic formulas* from Schneider and Weil [2008], allow us to compute expectations of the form $\mathbb{E}\left[F(V \cap C)\right]$, when:

- $C$ is a finite union of convex cones, and
- $F$ is additive, which means it satisfies the functional equation

$$F(A \cap B) + F(A \cup B) = F(A) + F(B) \tag{4.6}$$

for all $A, B \subset \mathbb{R}^m$.

The cone $C_{m,n}$ is a finite union of orthants in $\mathbb{R}^m$, however $\mathbb{1}^s$ does not satisfy the additivity Equation (4.6). The unique additive function which agrees with $\mathbb{1}^s$ on convex cones is defined through the *Euler characteristic* of subsets of the sphere $\mathbb{S}^{m-1}$.

**Definition 4.11** Let $\chi$ be the Euler characteristic function on $\mathbb{R}^m$. Define

$$\chi^s(A) = \chi(A \cap \mathbb{S}^{m-1}) \tag{4.7}$$

to be the *spherical Euler characteristic* of a set $A \in \mathbb{R}^m$.

Equations 6.62 and 6.63 in Schneider and Weil [2008] then give us the following theorem:

**Theorem 4.12** [Spherical Crofton formula] Let $C \subset \mathbb{R}^m$ be a (possibly non-convex) cone. Let $V$ be a random $n$-dimensional subspace of $\mathbb{R}^m$. Then

$$\mathbb{E}\left[\, \chi^s(V \cap C) \,\right] = 2 \sum_{i=0}^{\lfloor \frac{n-1}{2} \rfloor} v_{m-n+2i+1}(C). \tag{4.8}$$

Motivated by Theorem 4.12, let us define

$$q_{m,n} = \mathbb{E}\left[\, \chi^s(V \cap C_{m,n}) \,\right] \overset{?}{\approx} \mathbb{E}\left[\, \mathbb{1}^s(V \cap C_{m,n}) \,\right] = 1 - p_{m,n} \tag{4.9}$$

In the rest of this chapter, we estimate the order of $q_{m,n}$ explicitly. Our goal is to show that the phase transition from $|q_{m,n}| \to \infty$ to $q_{m,n} \to 0$ happens when $m/n = c_{\text{Euler}} \approx 8.34$.

### 4.1.1 Using $\chi$ as a surrogate for $\mathbb{1}$

The phase transition threshold for $q_{m,n}$ can be thought of as the "Euler characteristic threshold", in analogy with the injectivity threshold from Chapter 3. We conjecture that the Euler characteristic threshold might be a good candidate for the injectivity threshold. In particular, we have very good reasons to think $p_{m,n} \to 1$ implies $q_{m,n} \to 0$, and some vague intuition for the reverse implication.

In the theory of Gaussian random fields [Adler and Taylor, 2007], the Euler characteristic is used as an approximation to the indicator function of so-called *excursion sets*. If $f$ is sampled from a Gaussian process, an excursion set $A_u(f)$ is the preimage of an interval of the form $\langle u, +\infty \rangle$.

We paraphrase how Adler et al. [2015] justify the Euler characteristic heuristic:

Suppose that $u$ is large, so that the probability of $A_u(f)$ being nonempty is small. Then the excursion set $A_u(f)$ is most likely to be made up of a few isolated small regions, with neither holes, handles nor hollows. In fact, if $u$ is large, it is likely that there would be at most one component to this set. Then the Euler characteristic, which is equal to 1 on simply connected nonempty sets, approximates the indicator function of $A_u(f)$ well.

Thus, the approximation in Equation (4.9) is likely to be correct when $p_{m,n}$ is large, which would imply that the injectivity threshold is a lower bound to the Euler characteristic threshold.

We could also assume that there is no extreme cancellation in the expectation of the Euler characteristic when $p_{m,n}$ is small. If this assumption could be made rigorous, $p_{m,n} \to 0$ would imply $q_{m,n} \neq 0$, and the thresholds would be equal.

## 4.2 Calculating $q_{m,n}$

For clarity, we give an explicit formula for for $q_{m,n+1}$ instead of $q_{m,n}$. The difference does not matter, since we are interested in the behaviour of $q_{m,n}$ in terms of the ratio $m/n$.

**Theorem 4.13**

$$q_{m,n+1} = \frac{(-1)^n}{2^{m-n-1}} \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} U(m,n,2i) \tag{4.10}$$

where

$$U(m,n,2i) = \binom{m}{n-2i} \sum_{\ell=0}^{n} \left(-\frac{1}{2}\right)^{\ell} \binom{n-2i}{\ell-2i} \sum_{j=0}^{\ell} \binom{m-n+\ell}{j}. \tag{4.11}$$

In particular, the first term equals

$$U(m,n,0) = \binom{m}{n} \sum_{\ell=0}^{n} \left(-\frac{1}{2}\right)^{\ell} \binom{n}{\ell} \sum_{j=0}^{\ell} \binom{m-n+\ell}{j}. \tag{4.12}$$

Due to Theorem 4.12, the expectation of the Euler characteristic is determined by the intrinsic volumes $v_k(C_{m,n+1})$ for $k \geq m-n$. The main step in the proof is the following formula for $v_k$:

**Proposition 4.14** For $k \geq m-n$,

$$v_k(C_{m,n+1}) = (-1)^m \sum_{a=k}^{m} \left(-\frac{1}{2}\right)^{a} \binom{a}{k} \binom{m}{a} \sum_{j=0}^{n-m+a} \binom{a}{j}. \tag{4.13}$$

**Proof (of Proposition 4.14)** For a subset of coordinates $S \subseteq \{1,2,\ldots,m\}$, let $\mathcal{O}_S^m$ be the orthant with exactly the coordinates in $S$ positive. Let $\mathcal{F}_{m,n}$ be the the family of subsets of $\{1,2,\ldots,m\}$ with at most $n$ elements. The cone $C_{m,n+1}$ is the union of orthants indexed by elements of $\mathcal{F}_{m,n}$:

$$C_{m,n+1} = \bigcup_{\substack{S \subset m \\ |S| \leq n}} \mathcal{O}_S^m = \bigcup_{S \in \mathcal{F}_{m,n}} \mathcal{O}_S^m \tag{4.14}$$

By Definition 4.8, the intrinsic volume of an union of orthants is

$$v_k(C_{m,n+1}) = \sum_{\varnothing \neq \mathcal{J} \subseteq \mathcal{F}_{m,n}} (-1)^{|\mathcal{J}|+1} \, v_k \left( \bigcap_{S \in \mathcal{J}} \mathcal{O}_S^m \right). \tag{4.15}$$

For any subset $\mathcal{J} \subseteq \mathcal{F}_{m,n}$, the intersection $\bigcap_{S \in \mathcal{J}} \mathcal{O}_S^m$ is an inclusion of an orthant in $\mathbb{R}^m$; that is, a possibly lower-dimensional orthant. The dimension of $\bigcap_{S \in \mathcal{J}} \mathcal{O}_S^m$ is equal to the number of coordinates where all the orthants $(C_s)_{S \in \mathcal{J}}$ agree. It is useful to formalize this notion:

**Definition 4.15** Let $\mathcal{J}$ be a subfamily of subsets of $\{1, 2, \ldots, m\}$. The *agreement set* of $\mathcal{J}$ is equal to the following set:

$$A(\mathcal{J}) = \left( \bigcup_{S \in \mathcal{J}} S \right)^c \bigsqcup \left( \bigcap_{S \in \mathcal{J}} S \right) \tag{4.16}$$

The above notation allows us to phrase the dimensions of the intersection cones in the summation in Equation (4.15) succinctly. For any $\mathcal{J} \subseteq \mathcal{F}_{m,n}$,

$$\dim \operatorname{span} \bigcap_{S \in \mathcal{J}} \mathcal{O}_S^m = |A(\mathcal{J})|. \tag{4.17}$$

If $A(\mathcal{J}) = \varnothing$, the intersection of the orthants corresponding to sets in $\mathcal{J}$ is a single point $\{0\}$. Else, the intersection is an orthant of dimension $|A(\mathcal{J})|$, embedded in $\mathbb{R}^m$.

The intrinsic volumes $v_k$ of an embedded orthant in $\mathbb{R}^m$ depend only on its actual dimension. The following two lemmas, proved in Appendix B.1, give us explicit formulas for the intrinsic volumes of embedded orthants:

**Lemma 4.16** For $0 \leq k \leq d$, the intrinsic volume $v_k$ of the nonnegative orthant $\mathcal{O}^d_{\{1,2,\ldots,d\}} = \mathbb{R}^d_{\geq 0} \subseteq \mathbb{R}^d$ is

$$v_k(\mathbb{R}^d_{\geq 0}) = \frac{1}{2^d} \binom{d}{k}. \tag{4.18}$$

**Lemma 4.17** For $0 \leq d \leq m$, let $F : \mathbb{R}^d \to \mathbb{R}^m$ be an isometric linear embedding. Then, for $0 \leq k \leq d$,

$$v_k(F\mathbb{R}^d_{\geq 0}) = \frac{1}{2^d} \binom{d}{k}, \tag{4.19}$$

and $v_k(F\mathbb{R}^d_{\geq 0}) = 0$ for $k > d$.

The above two lemmas immediately imply that for all $\mathcal{J} \subseteq \mathcal{F}_{m,n}$,

$$v_k\left(\bigcap_{S\in\mathcal{J}} \mathcal{O}_S^m\right) = \frac{1}{2^{|A(\mathcal{J})|}}\binom{|A(\mathcal{J})|}{k}. \tag{4.20}$$

For $0 \le a \le m$ and $1 \le p \le \sum_{i=0}^{n}\binom{m}{i}$, define $N_{m,n}(p,a)$ to be the number of subfamilies $\mathcal{J} \subseteq \mathcal{F}_{m,n}$ with $|\mathcal{J}| = p$ and $|A(\mathcal{J})| = a$. Then Equation (4.15) can be written as

$$v_k(C_{m,n+1}) = \sum_{p\ge1}\sum_{a=k}^{m} N_{m,n}(p,a)(-1)^{p+1}\frac{1}{2^a}\binom{a}{k} \tag{4.21a}$$

$$= \sum_{a=k}^{m}\frac{1}{2^a}\binom{a}{k}\sum_{p\ge1}(-1)^{p+1}N_{m,n}(p,a) \tag{4.21b}$$

The plan now is to get a non-closed formula for $N_{m,n}(p,a)$, and then calculate $\sum_{p\ge1}(-1)^{p+1}N_{m,n}(p,a)$ directly. We can describe $N_{m,n}(p,a)$ as follows:

- There are $\binom{m}{a}$ ways to choose the set $A \stackrel{\text{def}}{=} A(\mathcal{J})$.
- If $|A| = a$, the cardinality of $R \stackrel{\text{def}}{=} \bigcap_{S\in\mathcal{J}} S$ can be any $0 \le r \le a$. If we fix $r$, there are $\binom{a}{r}$ ways to choose $R$.
- All $p$ sets in $\mathcal{J}$ are supersets of $R$, and do not contain any of the elements in $A\backslash R$. Let $\mathcal{J}/A \stackrel{\text{def}}{=} \{S\backslash A \,:\, S \in \mathcal{J}\}$ be the family of subsets of $\{1,2,\dots,m\}\backslash A$ defined by removing all elements of $A$ from all sets in $\mathcal{J}$. Given $R$ and $A$, this quotiented family uniquely defines $\mathcal{J}$.
- We have $|\mathcal{J}/A| = p$, the agreement set of $\mathcal{J}/A$ is empty, and every element of $\mathcal{J}/A$ has cardinality at most $n-r$.

To count the ways to satisfy the last point, we introduce a generalization.

**Definition 4.18** For nonnegative integers $d, s, p$, define $F(d,s,p)$ to be the number of families of $p$ distinct subsets $S_1,\dots,S_p \subseteq \{1,2,\dots,d\}$, with each subset having at most $s$ elements, and

$$S_1 \cap \ldots \cap S_p = \varnothing, \tag{4.22}$$

$$S_1 \cup \ldots \cup S_p = \{1,2,\dots,d\}. \tag{4.23}$$

Then, following the argument above, we have an expression for $N_{m,n}(p,a)$:

$$N_{m,n}(p,a) = \binom{m}{a}\sum_{r=0}^{a}\binom{a}{r}F(m-a,n-r,p) \tag{4.24}$$

The exact expression for $F(d,s,p)$ is not important here, because we can simplify Equation (4.21b) using a convenient combinatorial identity.

**Lemma 4.19** With $F(d,s,p)$ defined as in Definition 4.18,

$$\sum_{p\geq 1}(-1)^{p+1}F(d,s,p) = \begin{cases}(-1)^d & \text{if } s \geq d;\\ 0 & \text{otherwise.}\end{cases} \tag{4.25}$$

for all $d, s \geq 1$.

We defer the proof of Lemma 4.19 to Appendix A.3.

We now calculate the alternating sum terms in Equation (4.21b).

$$\begin{aligned}\sum_{p\geq 1}(-1)^{p+1}N_{m,n}(p,a) &= \sum_{p\geq 1}(-1)^{p+1}\binom{m}{a}\sum_{r=0}^{a}\binom{a}{r}F(m-a,n-r,p) && (4.26a)\\ &= \binom{m}{a}\sum_{r=0}^{a}\binom{a}{r}\sum_{p\geq 1}(-1)^{p+1}F(m-a,n-r,p) && (4.26b)\\ &= \binom{m}{a}\sum_{r=0}^{a}\binom{a}{r}(-1)^{m-a}\ \mathbb{1}_{n-r\geq m-a} && (4.26c)\\ &= \binom{m}{a}\sum_{r=0}^{a}\binom{a}{r}(-1)^{m-a}\ \mathbb{1}_{r\leq n-m+a} && (4.26d)\end{aligned}$$

Finally,

$$\begin{aligned}v_k(C_{m,n+1}) &= \sum_{a=k}^{m}\frac{1}{2^a}\binom{a}{k}\sum_{p\geq 1}(-1)^{p+1}N_{m,n}(p,a) && (4.27a)\\ &= \sum_{a=k}^{m}\frac{1}{2^a}\binom{a}{k}\binom{m}{a}\sum_{r=0}^{a}\binom{a}{r}(-1)^{m-a}\ \mathbb{1}_{r\leq n-m+a} && (4.27b)\\ &= (-1)^m\sum_{a=k}^{m}\left(-\frac{1}{2}\right)^a\binom{a}{k}\binom{m}{a}\sum_{r=0}^{a}\binom{a}{r}\ \mathbb{1}_{r\leq n-m+a} && (4.27c)\\ &= (-1)^m\sum_{a=k}^{m}\left(-\frac{1}{2}\right)^a\binom{a}{k}\binom{m}{a}\sum_{r=0}^{n-m+a}\binom{a}{r}. && (4.27d)\end{aligned}$$

□

**Proof (of Theorem 4.13)** Given Proposition 4.14, the proof boils down to algebraic manipulations. Theorem 4.12 gives us

$$\begin{aligned}q_{m,n} &= \mathbb{E}\left[\ \chi^s(V\cap C_{m,n})\ \right] && (4.28)\\ &= 2\sum_{i=0}^{\lfloor\frac{n-1}{2}\rfloor} v_{m-n+2i+1}(C_{m,n}). && (4.29)\end{aligned}$$

Then we have

$$q_{m,n+1} \tag{4.30a}$$

$$= 2\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} v_{m-n+2i}(C_{m,n+1}) \tag{4.30b}$$

$$= 2\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor}(-1)^m \sum_{a=m-n+2i}^{m} \left(-\frac{1}{2}\right)^a \binom{a}{m-n+2i}\binom{m}{a} \sum_{j=0}^{n-m+a}\binom{a}{j} \tag{4.30c}$$

$$= 2(-1)^m \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \sum_{a=m-n+2i}^{m} \left(-\frac{1}{2}\right)^a \binom{a}{m-n+2i}\binom{m}{a} \sum_{j=0}^{n-m+a}\binom{a}{j} \tag{4.30d}$$

$$= 2(-1)^m \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \sum_{a=m-n+2i}^{m} \left(-\frac{1}{2}\right)^a \binom{n-2i}{a-m+n-2i}\binom{m}{n-2i} \sum_{j=0}^{n-m+a}\binom{a}{j} \tag{4.30e}$$

$$= 2(-1)^m \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{m}{n-2i} \sum_{a=m-n+2i}^{m} \left(-\frac{1}{2}\right)^a \binom{n-2i}{a-m+n-2i} \sum_{j=0}^{n-m+a}\binom{a}{j} \tag{4.30f}$$

$$= 2(-1)^m \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{m}{n-2i} \sum_{a=m-n+2i}^{m} \left(-\frac{1}{2}\right)^a \binom{n-2i}{a-m+n-2i} \sum_{j=0}^{n-m+a}\binom{a}{j} \tag{4.30g}$$

We shift the index $a = m-n+2i+t$ to make it cleaner:

$$= 2(-1)^m \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{m}{n-2i} \sum_{t=0}^{n-2i} \left(-\frac{1}{2}\right)^{m-n+2i} \left(-\frac{1}{2}\right)^t \binom{n-2i}{t} \sum_{j=0}^{2i+t}\binom{m-n+2i+t}{j} \tag{4.31a}$$

$$= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{(-1)^{n-2i}\binom{m}{n-2i}}{2^{m-n+2i-1}} \sum_{t=0}^{n-2i} \left(-\frac{1}{2}\right)^t \binom{n-2i}{t} \sum_{j=0}^{2i+t}\binom{m-n+2i+t}{j} \tag{4.31b}$$

And again reindexing, with $\ell = t+2i$, using the convention $\binom{a}{b} = 0$ for $b < 0$:

$$= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{(-1)^{n-2i}\binom{m}{n-2i}}{2^{m-n+2i-1}} \sum_{\ell=0}^{n} \left(-\frac{1}{2}\right)^{\ell-2i} \binom{n-2i}{\ell-2i} \sum_{j=0}^{\ell}\binom{m-n+\ell}{j} \tag{4.32a}$$

$$= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{(-1)^{n}\binom{m}{n-2i}}{2^{m-n-1}} \sum_{\ell=0}^{n} \left(-\frac{1}{2}\right)^{\ell} \binom{n-2i}{\ell-2i} \sum_{j=0}^{\ell}\binom{m-n+\ell}{j} \tag{4.32b}$$

$$= \frac{(-1)^n}{2^{m-n-1}} \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{m}{n-2i} \sum_{\ell=0}^{n} \left(-\frac{1}{2}\right)^{\ell} \binom{n-2i}{\ell-2i} \sum_{j=0}^{\ell}\binom{m-n+\ell}{j}. \tag{4.32c}$$

□

## 4.3 Estimating the Euler characteristic threshold

This section is partly inspired by the MathOverflow answer Alekseyev [2021], and the comments by the MathOverflow user fedja to that question. Note however that the precise statements and the proof details here are original contributions of the author of this work.

We will now nonrigorously estimate the expression in Theorem 4.13 using the saddle point method. In principle, there is no great obstacle to doing it rigorously and proving the exact phase transition of $q_{m,n}$ in terms of $m/n$. However, in this chapter we are already working under the Euler characteristic heuristic. Our main goal is not to analyze $q_{m,n}$, but to demonstrate where the phase transition for $p_{m,n}$ could be if the Euler characteristic heuristic is justified.

In this section, it will be convenient to denote $m - n = Kn$ for $K > 0$.
Let us estimate the order of the terms $U(m, n, 2i)$ from Equation (4.11):

$$\log_2 |U(m,n,2i)| = \log_2 \binom{m}{n-2i} + \log_2 \left| \sum_{\ell=0}^{n} \left(-\frac{1}{2}\right)^{\ell} \binom{n-2i}{\ell-2i} \sum_{j=0}^{\ell} \binom{m-n+\ell}{j} \right|. \tag{4.33}$$

Consider the second term in the above expression. It is equal to

$$\log_2 \left| \sum_{\ell=0}^{n} \left(-\frac{1}{2}\right)^{\ell} \binom{n-2i}{\ell-2i} \sum_{j=0}^{\ell} \binom{m-n+\ell}{j} \right| \tag{4.34a}$$

$$= -n + \log_2 \left| \sum_{\ell=0}^{n} (-2)^{n-\ell} \binom{n-2i}{n-\ell} \sum_{j=0}^{\ell} \binom{m-n+\ell}{j} \right| \tag{4.34b}$$

$$= -n + \log_2 \left| \sum_{\ell=0}^{n} (-2)^{n-\ell} \binom{n-2i}{n-\ell} \sum_{j=0}^{\ell} \binom{m-n+\ell}{j} \right| \tag{4.34c}$$

We use generating functions to get a better expression for the term inside the logarithm. By the binomial theorem,

$$\sum_{\ell=0}^{n} (-2)^{n-\ell} \binom{n-2i}{n-\ell} = \left[x^{n-\ell}\right] (1-2x)^{n-2i}. \tag{4.35}$$

For the prefix sum of binomial coefficients term, we use a more advanced technique to get it as a coefficient sequence of the right generating function.

**Lemma 4.20**

$$\sum_{j=0}^{\ell} \binom{m-n+\ell}{j} = \left[x^{\ell}\right] (1-x)^{-m+n} (1-2x)^{-1}. \tag{4.36}$$

We defer the proof of Lemma 4.20 to Appendix A.4.

Now we can write

$$\left|\sum_{\ell=0}^{n}(-2)^{n-\ell}\binom{n-2i}{n-\ell}\sum_{j=0}^{\ell}\binom{m-n+\ell}{j}\right| \tag{4.37a}$$

$$=\left|\left(\sum_{\ell=0}^{n-2i}\left[x^{n-\ell}\right](1-2x)^{n-2i}\right)\left(\sum_{\ell=-\infty}^{\infty}\left[x^{\ell}\right](1-x)^{-m+n}(1-2x)^{-1}\right)\right| \tag{4.37b}$$

$$=\left|\left[x^{n}\right](1-2x)^{n-2i-1}(1-x)^{-m+n}\right|, \tag{4.37c}$$

and thus

$$\log_2\left|\sum_{\ell=0}^{n}\left(-\frac{1}{2}\right)^{\ell}\binom{n-2i}{\ell-2i}\sum_{j=0}^{\ell}\binom{m-n+\ell}{j}\right| \tag{4.38a}$$

$$=-n+\log_2\left|\left[x^{n}\right](1-2x)^{n-2i-1}(1-x)^{-m+n}\right| \tag{4.38b}$$

$$=-(2i+1)+\log_2\left|\left[x^{n}\right]\left(x-\frac{1}{2}\right)^{n-2i-1}(1-x)^{-m+n}\right| \tag{4.38c}$$

$$=-(2i+1)+\log_2\left|\frac{1}{2\pi}\int_{\gamma}\frac{\left(z-\frac{1}{2}\right)^{n-2i-1}(1-z)^{-m+n}}{z^{n+1}}\,\mathrm{d}z\right|, \tag{4.38d}$$

where we used Cauchy's integral formula and Taylor's theorem. Here $\gamma$ is a small enough loop around the origin in $\mathbb{C}$.

From now on, we proceed to give a nonrigorous estimate on the integral above using the saddle point method. The saddle point method allows us to estimate integrals of the form

$$I(n)=\int_{\gamma}g(z)e^{nf(z)}\,\mathrm{d}z \tag{4.39}$$

for $\gamma$ a tight loop around the origin. Let $z_0$ be a dominant saddle point of $f$. Then the approximated integral is

$$|I(n)|\approx\left|g(z_0)e^{nf(z_0)}\sqrt{\frac{2\pi}{n|f''(z_0)|}}\right|. \tag{4.40}$$

Recall we denote $m-n=Kn$ for $K>0$. We will apply the saddle point method to the integral

$$\int_{\gamma}\frac{\left(z-\frac{1}{2}\right)^{n-2i-1}(1-z)^{-Kn}}{z^{n+1}}\,\mathrm{d}z\,. \tag{4.41}$$

To bring it into the format of Equation (4.39), we pick

$$g(z) = \left(z - \frac{1}{2}\right)^{-2i-1} z^{-1} \tag{4.42a}$$

$$f(z) = \log\left(\frac{1}{2} - z\right) - \log(z) - K\log(1-z) \tag{4.42b}$$

For $K > 3 + 2\sqrt{2} \approx 5.8$, the relevant saddle point of $f$ is

$$z_0 = \frac{-\sqrt{K^2 - 6K + 1} + K + 1}{4K}, \tag{4.43}$$

see Appendix A.5.

Note that $z_0$ is a positive real number smaller than $\frac{1}{2}$: The saddle point method now gives

$$\log_2 \left| \int_\gamma \frac{\left(z - \frac{1}{2}\right)^{n-2i-1} (1-z)^{-Kn}}{z^{n+1}} \, \mathrm{d}z \right| \tag{4.44}$$

$$\approx \log_2 g(z_0) + n \log_2(e) f(z_0) + O(\log n). \tag{4.45}$$

$$\approx -(2i+1)\log_2 \left|\frac{1}{2} - z_0\right| + n\log_2(e) f(z_0) + O(\log n). \tag{4.46}$$

Apart from the convergence issues when applying Taylor's theorem, our approximation here is not rigorous for two more reasons:

- The function $g(z)$ depends on $i$, which may depend on $n$;
- The function $f(z)$ has multiple saddle points depending on $K$, and the analysis of which one is dominant depends on $K$ and $i$.

However, we believe that these two issues are not crucial, because:

- Experimentally, for moderately large $n$, the terms $U(m, n, 2i)$ with small $i$ are dominant; see Figure 4.2. The terms for larger $i$ are completely negligible due to the $\binom{m}{n-2i}$ factor.
- For small $i$ and moderately large $n$, in the regime of $K$ we are interested in, the experiments agree with the approximation in the formula.

Plugging in the saddle point estimate into the equation for the integral yields

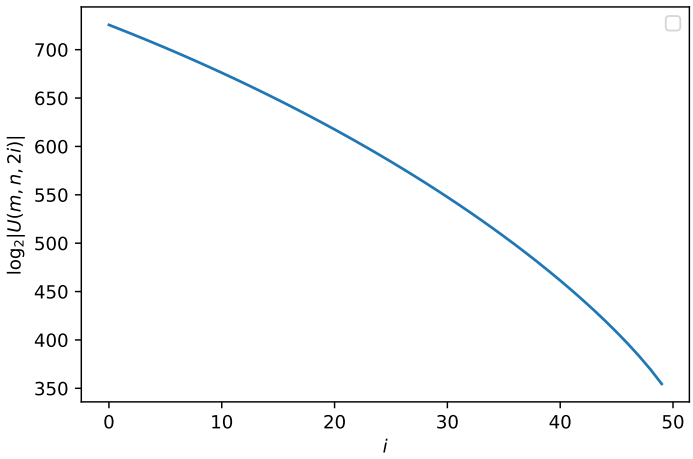



**Figure 4.2:** For $n = 100$ and $m = 834$, the expression for $q_{m,n+1}$ in Theorem 4.13 is dominated by the first term, which corresponds to $i = 0$. Note the logarithmic scale in the figure. The ratio $m/n$ here is close to the predicted phase transition.

$$\log_2 \left| \sum_{\ell=0}^{n} \left(-\frac{1}{2}\right)^{\ell} \binom{n-2i}{\ell-2i} \sum_{j=0}^{\ell} \binom{m-n+\ell}{j} \right| \tag{4.47a}$$

$$= -(2i+1) + \log_2 \left| \frac{1}{2\pi} \int_{\gamma} \frac{\left(z-\frac{1}{2}\right)^{n-2i-1} (1-z)^{-m+n}}{z^{n+1}} \, \mathrm{d}z \right| \tag{4.47b}$$

$$\approx -(2i+1) + \left( -(2i+1) \log_2 \left| \frac{1}{2} - z_0 \right| + n \log_2(e) f(z_0) + O(\log n) \right) \tag{4.47c}$$

$$= -2i - 2i \log_2 \left| \frac{1}{2} - z_0 \right| + n \log_2(e) f(z_0) + O(\log n) \tag{4.47d}$$

and thus finally

$$\log_2 |U(m,n,2i)| \tag{4.48a}$$

$$\approx \log_2 \binom{m}{n-2i} - 2i - 2i \log_2 \left| \frac{1}{2} - z_0 \right| + n \log_2(e) f(z_0) + O(\log n) \tag{4.48b}$$

$$\approx m \, H((n-2i)/m) - 2i - 2i \log_2 \left| \frac{1}{2} - z_0 \right| + n \log_2(e) f(z_0) + o(n) \tag{4.48c}$$

$$= (K+1)n \, H\left( \frac{1}{K+1} \frac{n-2i}{n} \right) - 2i - 2i \log_2 \left| \frac{1}{2} - z_0 \right| + n \log_2(e) f(z_0) + o(n) \tag{4.48d}$$

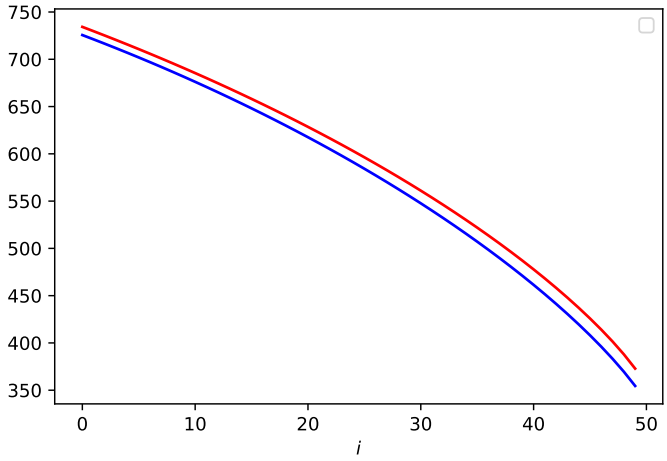



**Figure 4.3:** Our approximation in Equation (4.48d) for $\log_2 |U(m,n,2i)|$ in blue, versus the exactly computed $\log_2 |U(m,n,2i)|$ in red. Here $n = 100$ and $m = 834$, for $0 \leq i \leq n/2$ on the $x$-axis.

The approximation in Equation (4.48) is experimentally sound, as we can see from Figure 4.4. We now simplify the calculation with some technical statements. The following lemma is proved in Appendix A.6.

**Lemma 4.21** The expression

$$(K+1)n\ H\left(\frac{1}{K+1}\frac{n-2i}{n}\right) - 2i - 2i\log_2\left|\frac{1}{2} - z_0\right| + n\log_2(e)f(z_0) \quad (4.49)$$

is stricly decreasing in $i$ on the set $\{0,1,\ldots,\lfloor n/2 \rfloor\}$. In particular, the approximated $\log_2 |U(m,n,2i)|$ is maximal in

$$\log_2 |U(m,n,0)| \approx n\left((K+1)\ H\left(\frac{1}{K+1}\right) + \log_2(e)f(z_0)\right) \quad (4.50)$$

The above approximation is experimentally sound; the error in Figure 4.4 is very small compared to $n$. Lemma 4.21 motivates the following approxima-

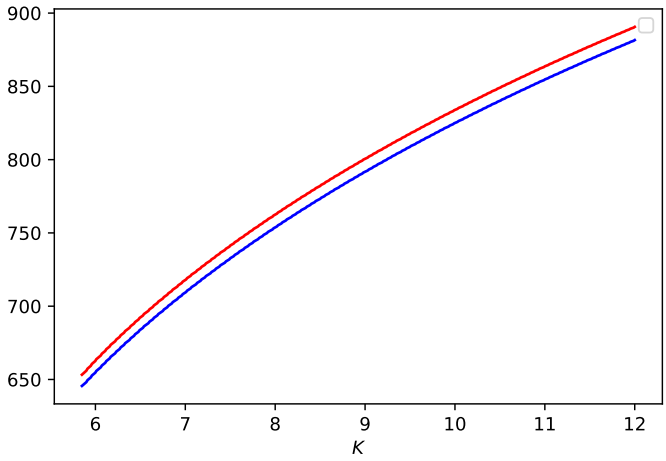



**Figure 4.4:** Our approximation in Equation (4.50) for $\log_2 |U(m,n,0)|$ in blue, versus the exactly computed $\log_2 |U(m,n,0)|$ in red. Here $n = 100$ and $m = (K+1)n$ for $K$ on the $x$-axis.

tion of the expression in Theorem 4.13:

$$\log_2 |q_{m,n+1}| \tag{4.51a}$$

$$= -m + n + 1 + \log_2 \left| \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} U(m,n,2i) \right| \tag{4.51b}$$

$$\approx -m + n + 1 + \log_2 |U(m,n,0)| \tag{4.51c}$$

$$\approx -Kn + n \left( (K+1)\, H\left(\frac{1}{K+1}\right) + log_2(e) f(z_0) \right) + o(n) \tag{4.51d}$$

$$= n \left( -K + (K+1)\, H\left(\frac{1}{K+1}\right) + \log_2(e) f(z_0) \right) + o(n). \tag{4.51e}$$

When $n \to \infty$, the final expression above goes to $-\infty$ or $+\infty$, depending on the sign of the coefficient of $n$. We calculate the point where the coefficient change sign in the following proposition, which we prove in Appendix A.7.

**Proposition 4.22** The expression

$$-K + (K+1)H\left(\frac{1}{K+1}\right) + \log_2(e) f(z_0) \tag{4.52}$$

is decreasing on $K > 3 + 2\sqrt{2} \approx 5.8$, and has an unique root

$$K_0 \approx 7.34463. \tag{4.53}$$

Recall that we defined $m = n + Kn$. In view of Proposition 4.22 and the computations in Equation (4.51), we have the following estimate on the phase transition.

**Proposition 4.23** The phase transition for $q_{m,n}$ as defined in Equation (4.9) and calculated in Theorem 4.13 happens at $m = c_{\text{Euler}}n$, for

$$c_{\text{Euler}} = K_0 + 1 \approx 8.34. \tag{4.54}$$

The phase transition is in the sense of: for all $\varepsilon > 0$, when $n \to \infty$,

$$m < (c_{\text{Euler}} - \varepsilon)n \implies |q_{m,n}| \to \infty \tag{4.55a}$$
$$m > (c_{\text{Euler}} + \varepsilon)n \implies q_{m,n} \to 0 \tag{4.55b}$$

Note that we do not claim this as a theorem due to several nonrigorous steps in this section. We did the computations because of Equation (4.9), so that now we can pose the following conjecture for the injectivity threshold.

**Conjecture 4.24** The injectivity phase transition coincides with the expected Euler characteristic phase transition: for all $\varepsilon > 0$, when $n \to \infty$,

$$m < (c_{\text{Euler}} - \varepsilon)n \implies p_{m,n} \to 0 \tag{4.56a}$$
$$m > (c_{\text{Euler}} + \varepsilon)n \implies p_{m,n} \to 1, \tag{4.56b}$$

for $c_{\text{Euler}} \approx 8.34$.

Conjecture 4.24 is consistent with our findings in Chapter 3, where we gave a proof that the injectivity threshold is at most 9.09, and experimental evidence that it is larger than 5.

Chapter 5

# Deep injective networks

The natural extension of Chapter 3 is to try to get injectivity guarantees for deep random networks. We can apply Theorem 3.12 on each layer to prove that the network with $(d_0, d_1, \dots, d_L) = (n, Cn, C^2n, \dots, C^{L-1}n)$ is injective with probability $1 - o(n)$, for large enough $C$. This is not satisfying, because the layer widths increase exponentially.

In this chapter, we prove that the layer widths may stay constant after the first layer, with moderate restrictions on the depth:

**Theorem 5.1** Let $C \geq 2L \log L + \Theta(L)$. For $(d_0, \dots, d_L) = (n, Cn, \dots, Cn)$, the network with random Gaussian weights and zero biases is injective with probability $1 - o(n)$.

We first develop a generalization of Theorem 3.4 to deep networks in Section 5.1. We use this tool to prove Theorem 5.1 in Section 5.3, using technical bounds from Section 5.2 and Chapter 6. We conclude with the connection of injectivity of deep networks with well-known contractive properties of neural networks in Section 5.4.

## 5.1 Characterizing injectivity

The goal of this section is to prove the key fact Lemma 5.8: a criterion for injectivity in deep random ReLU networks. In this section, the only assumption on the biases is that they are independent of the weight matrices.

For simplicity of exposition, we assume $d_\ell \geq 3d_0$ for all $1 \leq \ell \leq L$. This is not restrictive, as Theorem 3.11 shows that $f$ is not injective for $d_1 < 3.4d_0$ if $d_0$ is large enough.

For single-layer networks, Theorem 3.4 is a practical criterion of injectivity when the weights are independent Gaussians. One may ask whether there is a similar criterion for deep networks.

**Example 5.2** Let $W_1 \in \mathbb{R}^{d_1 \times d_2}$ and $W_2 \in \mathbb{R}^{d_2 \times d_3}$ be matrices with independent $N(0,1)$ entries. If we apply Theorem 3.4 on both layers, we get that the neural network $x \mapsto \mathrm{ReLU}(W_2 \,\mathrm{ReLU}(W_1 x))$ is injective whenever:

- $W_1 \mathbb{R}^{d_1}$ doesn't intersect any orthant $O_S^{d_2}$ with $|S| < d_1$; and
- $W_2 \mathbb{R}^{d_2}$ doesn't intersect any orthant $O_S^{d_3}$ with $|S| < d_2$.

However, if $d_2 > d_1$, the above criterion loses a lot of power. The second layer doesn't actually take $\mathbb{R}^{d_2}$ as the input, but only a much smaller subset of $\mathbb{R}^{d_2}$: the image of $x \mapsto \mathrm{ReLU}(W_1 x)$, which should look like a piecewise linear $d_1$-dimensional "submanifold".

It turns out we can do much better than in Example 5.2. Let us first introduce notation for prefixes of a neural network.

**Definition 5.3** Let $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$ be a neural network defined as in Equation (2.5):

$$f = \mathrm{ReLU} \circ B^{(L)} \circ \ldots \circ \mathrm{ReLU} \circ B^{(1)}. \tag{5.1}$$

For $1 \le \ell \le L$, denote by $f_\ell$ the *prefixes* of the neural network $f$:

$$f_\ell = \mathrm{ReLU} \circ B^{(\ell)} \circ \ldots \circ \mathrm{ReLU} \circ B^{(1)}. \tag{5.2}$$

By convention, define also $f_0 = \mathrm{id}$. Note that $f_L = f$.

The following lemma formalizes the key fact in this section:

**Proposition 5.4 (Informal)** Any layer can only create new noninjectivities on half-open orthants with few pluses. That is, if $f_\ell(x) = f_\ell(x') \in O_S^{d_\ell}$ with $|S|$ small, then $f_{\ell-1}(x) = f_{\ell-1}(x')$.

**Lemma 5.5** Let $x \in \mathbb{R}^{d_0}$ be an input vector, and $2 \le \ell \le L$. If $f_\ell(x)$ has at least $2d_0 + 1$ positive coordinates, then almost surely there doesn't exist an $x' \in \mathbb{R}^{d_0}$ for which $f_\ell(x) = f_\ell(x')$, but $f_{\ell-1}(x) \neq f_{\ell-1}(x')$.

**Proof** Every layer $\mathrm{ReLU} \circ B^{(i)}$ is a piecewise affine map, which is affine on the "wedges" mapping to any half-open orthant. More precisely, given an orthant $O_S^{d_i}$ such that the preimage $(B^{(i)})^{-1} O_S^{d_i}$ is nonempty, define the affine map $B_S^{(i)}$ such that $B_S^{(i)} x = \mathrm{ReLU}(B^{(i)} x)$ on this preimage.

The condition $f_\ell(x) = f_\ell(x')$ can be written as

$$B_{S_\ell}^{(\ell)} B_{S_{\ell-1}}^{(\ell-1)} \ldots B_{S_1}^{(1)}(x) = B_{S'_\ell}^{(\ell)} B_{S'_{\ell-1}}^{(\ell-1)} \ldots B_{S'_1}^{(1)}(x') \tag{5.3}$$

for some sets $S_i, S'_i \subseteq \{1, 2, \ldots, d_i\}$ for $1 \le i \le \ell$.
Note that $S_\ell = S'_\ell$, because $f_\ell(x) = f_\ell(x')$.

Write the affine map $B^{(\ell)}_{S_\ell}$ as the sum of a linear map and a bias term:

$$B^{(\ell)}_{S_\ell}(t) = W^{(\ell)}_{S_\ell} t + b^{(\ell)} \tag{5.4}$$

The matrix $W^{(\ell)}_{S_\ell}$ is just $W^{(\ell)}$ with some rows zeroed out. The rows replaced by zeros correspond to the entries of $f_\ell(x)$ which are zero, which are indexed by the complement of $S_\ell$. Moreover, the nonzero rows are of full rank almost surely, since the entries are independent Gaussians. As $f(x)$ has at least $2d_0 + 1$ positive coordinates, $|S_\ell| \geq 2d_0 + 1$. This means the nullity of $W^{(\ell)}_{S_\ell}$ is at most $d_{\ell-1} - 2d_0 - 1$.

We can rewrite Equation (5.3) as

$$B^{(\ell-1)}_{S_{\ell-1}} \dots B^{(1)}_{S_1}(x) - B^{(\ell-1)}_{S'_{\ell-1}} \dots B^{(1)}_{S'_1}(x') \in \operatorname{Ker} W^{(\ell)}_{S_\ell}. \tag{5.5}$$

Consider the left-hand side of Equation (5.5) as a function of $x$ and $x'$. Its image is contained in a linear subspace $V \subseteq \mathbb{R}^{d_{\ell-1}}$ with $\dim V \leq 2d_0 + 1$, because it is an affine transform of the vector $[x, x']^T \in \mathbb{R}^{2d_0}$.

If the matrix $W^{(\ell)}_{S_\ell}$ was independent of the affine subspace in question, we would be done, because $\dim V + \dim \operatorname{Ker} W^{(\ell)}_{S_\ell} \leq d_{\ell-1}$. The issue is that the sets corresponding to the orthants in different layers are not independent.

However, we can condition on the sets $(S_i, S'_i)_{1 \leq i \leq \ell}$. The matrix $W^{(\ell)}_{S_\ell}$ and the affine maps before are conditionally independent, given the sets corresponding to the orthants. Thus we have

$$\mathbb{P}\left[V \cap \operatorname{Ker} W^{(\ell)}_{S_\ell} \neq \{0\} \mid (S_i, S'_i)_{1 \leq i \leq \ell}\right] = 0, \tag{5.6}$$

and the probability is still zero after union bounding over the finite number of possibilities for the sets $S_i$ and $S'_i$. □

**Remark 5.6** If the biases are zero, the bound $2d_0 + 1$ can be replaced by $2d_0$. In that case, the affine transform that yields $V$ is linear, thus the right-hand side of Equation (5.5) is contained in a subspace $V$ with $\dim V \leq 2d_0$.

If we apply Lemma 5.5 to all layers, we get an actionable criterion for proving injectivity of $f$.

**Corollary 5.7** If $f_\ell(x)$ has at least $2d_0 + 1$ positive coordinates for all $x \in \mathbb{R}^{d_0}$ and $1 \leq \ell \leq L$, then $f$ is injective.

As mentioned in Remark 5.6, we have shown a slightly stronger statement when the biases are zero. The following statement will be our main tool in the following sections.

**Lemma 5.8** If $f$ has zero biases and $f_\ell(x)$ has at least $2d_0$ positive coordinates for all $x \in \mathbb{R}^{d_0} \backslash \{0\}$ and $1 \leq \ell \leq L$, then $f$ is injective.

We can leave out the origin from the input space $\mathbb{R}^{d_0}$, since if $f_\ell(x) = f_\ell(0)$ for $x \neq 0$, we can apply Lemma 5.5 on $x$ instead of 0.

Lemma 5.8 gets an important obstacle to injectivity out of the way, as a priori different inputs could take completely different paths through the network and map to the same output in the last layer. Thus, the problem is now similar to the one-layer case, where orthants with less than $d_0$ pluses are the only obstacle to injectivity.

For $L = 1$, Theorem 3.4 shows that the bound $2d_0$ in Lemma 5.8 can be replaced with $d_0$. It is open whether the bound can be improved for $L \geq 2$.

## 5.2 Activation regions of deep networks

From now on, assume the biases are all zero, so the affine transforms $B^{(\ell)}$ in Equation (2.5) are random matrices $W^{(\ell)}$ with independent Gaussian entries.

Let $W \in \mathbb{R}^{m \times n}$ be any matrix. Then the ReLU network layer $x \mapsto \mathrm{ReLU}(Wx)$ splits the input space into several *activation regions*, defined by the signs of the preactivation neurons $w_i^T x$.

**Lemma 5.9** For $m \geq n$, let $W \in \mathbb{R}^{m \times n}$ be a matrix with independent $N(0,1)$ entries. The number of activation regions of the network $x \mapsto \mathrm{ReLU}(Wx)$ is exactly

$$2 \sum_{i=0}^{n-1} \binom{m-1}{i} = 2T(m-1, n-1) \leq \left(\frac{em}{n}\right)^n \tag{5.7}$$

**Proof** This is equivalent to Lemma 6.4. □

In deeper networks, there are subtle differences between various intuitive definitions of activation regions. We formally define them as follows:

**Definition 5.10 (Activation region)** An activation region $A$ of $f$ is a maximal subset of $\mathbb{R}^{d_0}$ on which all the prefixes $f_1, f_2, \ldots, f_L$ are affine maps.

This definition is equivalent to how Hanin and Rolnick [2019] define activation regions. Another concept used in the literature are *linear regions*, which are the regions on which $f$ itself is affine. In a ReLU network with zero biases, all activation regions are cones.

The natural question to ask is how many activation regions are there in a random network with given layer widths. There is a simple bound which is exponential in the number of layers $L$.

**Lemma 5.11** The number of activation regions of $f_\ell$ is at most $\left(e^\ell \prod_{i=1}^{\ell} \frac{d_i}{d_0}\right)^{d_0}$.

We defer the proof of Lemma 5.11 to Appendix A.2.

## 5.3 Injectivity of random deep networks

Consider the action of $W^{(\ell)} f_{\ell-1}$ on the input space $\mathbb{R}^n$. Each activation region $\mathcal{A}$ of $f_{\ell-1}$ is mapped to a cone contained in a $d_0$-dimensional subspace. Due to rotational invariance of $W^{(\ell)}$, as seen in Proposition 3.3, the image of $\mathcal{A}$ is contained in a random $d_0$-dimensional subspace.

Inspired by this, we introduce one more technical ingredient from Chapter 6.

**Lemma 6.6** Given a random subspace $V \subseteq \mathbb{R}^m$, with $\dim V = n \leq m$, the probability that $V$ intersects a fixed orthant nontrivially is

$$\frac{1}{2^{m-1}} \sum_{i=0}^{n-1} \binom{m-1}{i} \leq 2^{-m} \left(\frac{em}{n}\right)^n. \tag{6.9}$$

We can now prove Theorem 5.1.

**Theorem 5.1** Let $C \geq 2L \log L + \Theta(L)$. For $(d_0, \ldots, d_L) = (n, Cn, \ldots, Cn)$, the network with random Gaussian weights and zero biases is injective with probability $1 - o(n)$.

**Proof** We prove that, with large probability, $f(x)$ has more than $2n$ positive coordinates for all $x \in \mathbb{R}^n \backslash \{0\}$. This will set us up to use Lemma 5.8 to prove injectivity.

Let $\mathcal{R}_L$ be the "preactivation image" of $f$, that is,

$$\mathcal{R}_L \stackrel{\text{def}}{=} W^{(L)} f_{L-1}(\mathbb{R}^n \backslash \{0\}). \tag{5.8}$$

For any activation region $\mathcal{A} \subset \mathbb{R}^n$ of $f$ that is not equal to the zero cone $\{0\}$, consider its image

$$\mathcal{R}_L(\mathcal{A}) \stackrel{\text{def}}{=} W^{(L)} f_{L-1}(\mathcal{A}) \subseteq \mathbb{R}^{Cn}. \tag{5.9}$$

The preactivation image $\mathcal{R}_L$ is the union of the images $\mathcal{R}_L(\mathcal{A})$, as $\mathcal{R}_L$ contains the image of the activation region containing the origin $0 \in \mathbb{R}^n$ if and only if that region is not just the zero cone $\{0\} \subset \mathbb{R}^n$.

We calculate the probability over the randomness of $f$:

$$\mathbb{P}\left[\mathcal{R}_L \text{ intersects a half-open orthant with at most } 2n \text{ pluses}\right] \tag{5.10a}$$
$$\leq \sum_{\mathcal{A}} \mathbb{P}\left[\mathcal{R}_L(\mathcal{A}) \text{ intersects a half-open orthant with at most } 2n \text{ pluses}\right] \tag{5.10b}$$
$$\leq \left(e^L C^L\right)^n \mathbb{P}\left[\mathcal{R}_L(\mathcal{A}) \text{ intersects a half-open orthant with at most } 2n \text{ pluses}\right] \tag{5.10c}$$
$$\leq \left(e^L C^L\right)^n T(Cn, 2n)\, \mathbb{P}\left[\mathcal{R}_L(\mathcal{A}) \text{ intersects a fixed half-open orthant}\right] \tag{5.10d}$$
$$\leq \left(e^L C^L\right)^n T(Cn, 2n)\, \mathbb{P}\left[\text{ span}\, \mathcal{R}_L(\mathcal{A}) \text{ intersects a fixed half-open orthant nontrivially }\right] \tag{5.10e}$$
$$< \left(e^L C^L\right)^n \left(\frac{eCn}{2n}\right)^{2n} 2^{-Cn} \left(\frac{eCn}{n}\right)^n \tag{5.10f}$$
$$= \left(2^{-C-2} e^{L+3} C^{L+3}\right)^n. \tag{5.10g}$$

We used the union bound, then Lemma 5.11, then the union bound again, and finally Lemma 6.6.

If we put $C \geq 2L \log L + 30L$, the term inside the parentheses becomes smaller than $\frac{1}{2}$, thus

$$\mathbb{P}\left[\mathcal{R}_L \text{ intersects a half-open orthant with at most } 2n \text{ pluses}\right] \leq 2^{-n}. \tag{5.11}$$

This implies that with probability $1 - 2^{-n}$, for all $x \in \mathbb{R}^n \backslash \{0\}$, the output $f(x)$ has more than $2n$ positive coordinates. Moreover, the same is analogously true for each prefix $f_1, \ldots, f_{L-1}$ of $f$, so Lemma 5.8 finishes the proof. □

If we can improve Lemma 5.11, we can get a better depenendency in Theorem 5.1, possibly even injectivity for $C \gtrsim \log L$. This is discussed in Appendix A.2. However, improving Lemma 5.11 will not give us injectivity for a constant $C$ independent of $L$, since the number of activation regions grows with $L$. In the next section, we show an alternative line of attack.

## 5.4 Angle convergence and injectivity

As discussed in Labatie [2019], Daneshmand et al. [2020], Hanin and Rolnick [2018], Schoenholz et al. [2016], and possibly elsewhere, the image of a deep random ReLU neural network resembles a half-line through the origin. Variants of this phenomenon are called "pathology of one-dimensional signal", "mean shift" and "rank collapse" in the literature. The rank collapse paper Daneshmand et al. [2020] gives evidence that a fixed set of input points,

passed through a deep random ReLU network, almost surely converges to a rank-one matrix. In Schoenholz et al. [2016], it is shown that large-width random layers increase the correlation of two fixed inputs with high probability, which means that the angle of two inputs decreases as they are passed through the network. This phenomenon, which we call *angle convergence*, turns out be be very useful for injectivity of neural networks.

Given the rich literature for the angle convergence of a fixed set of inputs, it is plausible that an uniform version of angle convergence could be true:

**Conjecture 5.12 (Informal)** The image of a deep random ReLU network is contained in a cone of small angle.

In this section, we formalize this conjecture and prove that angle convergence implies injectivity of deep networks of width independent of depth.

**Definition 5.13** Let $\varepsilon > 0$. For any nonzero vector $z \in \mathbb{R}^m$, define the set

$$\mathcal{B}_{\text{angle}}(z, \varepsilon) = \left\{ y \in \mathbb{R}^m \backslash \{0\} : z^T y \geq (1 - \varepsilon^2) \|z\| \|y\| \right\} \tag{5.12}$$

of vectors with $\varepsilon$-small angle with $z$. Note that $\mathcal{B}_{\text{angle}}(z, \varepsilon)$ is a cone with the origin removed, and that it doesn't depend on the norm of $z$.

As in Section 5.3, consider the network with $(d_0, d_1, \ldots, d_L) = (n, Cn, \ldots, Cn)$, zero biases and random Gaussian layers. We can prove noninjectivities cannot happen too "locally" in the angle sense:

**Proposition 5.14** Let $x \in \mathbb{R}^n$ be an input vector, and let $\ell \geq 2$. With probability $1 - o(n)$, there doesn't exist an $x' \in \mathbb{R}^n$ for which $f_\ell(x) = f_\ell(x')$, but $f_{\ell-1}(x) \neq f_{\ell-1}(x')$ and $W^{(\ell)} f_{\ell-1}(x') \in \mathcal{B}_{\text{angle}}(W^{(\ell)} f_{\ell-1}(x), \frac{1}{2})$.

**Remark 5.15** The statement of Proposition 5.14 looks as if it is not symmetric in $x$ and $x'$, but it is actually symmetric. The condition

$$W^{(\ell)} f_{\ell-1}(x') \in \mathcal{B}_{\text{angle}} \left( W^{(\ell)} f_{\ell-1}(x), \frac{1}{2} \right) \tag{5.13}$$

is equivalent to (using Equation (5.12)):

$$f_{\ell-1}(x')^T \left( W^{(\ell)} \right)^T W^{(\ell)} f_{\ell-1}(x) \geq \frac{3}{4} \left\| W^{(\ell)} f_{\ell-1}(x') \right\| \left\| W^{(\ell)} f_{\ell-1}(x) \right\|. \tag{5.14}$$

**Proof** The plan is to show that $f_\ell(x)$ has at least $2n$ positive coordinates, and then use use Lemma 5.5 and Remark 5.6. For a vector $x \in \mathbb{R}^m$, denote $x_{\text{pos}} = \text{ReLU}(x)$ and $x_{\text{neg}} = x - x_{\text{pos}}$.

The key technical part is the following statement, proved in Appendix A.8:

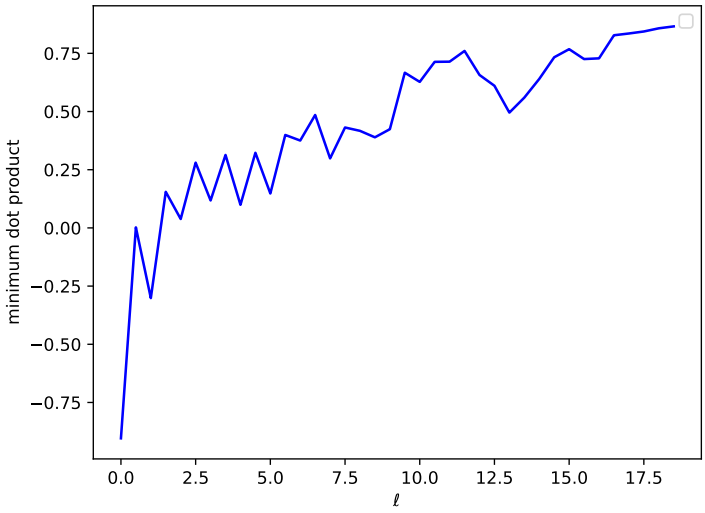



**Figure 5.1:** The minimum pairwise dot product over $k$ random points in $\mathbb{R}^n$, passed through the network with layer dimensions $(n, Cn, \ldots, Cn)$, and normalized to the unit sphere after each layer. Here $n = 10$, $C = 8$, $L = 20$, $k = 200$. We plot the minimum dot products after both activation and preactivation layers, as can be seen from the "jumps" in the plot. The results for preactivations and activations are indexed by $\{0.5, 1.5, \ldots, 19.5\}$ and $\{0, 1, \ldots, 20\}$, respectively.

**Lemma 5.16** Let $m = Cn$, for $C$ a large enough constant. Consider a random vector $z \in \mathbb{R}^m$ drawn from any rotationally invariant distribution. With probability $1 - o(n)$, all $y \in \mathcal{B}_{\text{angle}}\left(z, \frac{1}{2}\right)$ have at least $2n$ positive coordinates.

The matrix $W^{(\ell)}$ has a rotationally invariant distribution, so we can apply Lemma 5.16 to $z = W^{(\ell)} f_{\ell-1}(x)$. Then the conditions of Lemma 5.5 (for zero biases) are satisfied, hence there aren't any new noninjectivities in layer $\ell$.□

We propose the following concrete line of attack on the injectivity question using Proposition 5.14.

**Conjecture 5.17** There is a constant $C$ such that for every large enough $L$: for any $x, x' \in \mathbb{R}^n$, the network $f$ with $(d_0, d_1, \ldots, d_L) = (n, Cn, \ldots, Cn)$, we have $W^{(L)} f_{L-1}(x) \in \mathcal{B}_{\text{angle}}\left(W^{(L)} f_{L-1}, \frac{1}{2}\right)$.

The conjecture would imply injectivity (with large probability in $n$) for $C$ independent of $L$, since we could apply Theorem 5.1 for the first several layers, and Proposition 5.14 for the rest of the network.

Conjecture 5.17 is stronger than the comparable statements in the literature, since it concerns convergence in angle uniformly over the whole input space. In Figure 5.1, we give some evidence in support of the conjecture by covering the input space using an epsilon-net.

Chapter 6

# Random vectors in a halfspace and related results

One of the ingredients of the proofs in Chapters 3 and 5 is the formula for the probability that a given number of independent standard Gaussian vectors lie in a single halfspace. Here we discuss this formula and related results, inspired mostly by Bürgisser and Cucker [2013].

## 6.1 Number of regions in a hyperplane arrangement

**Definition 6.1** A *halfspace* $H$ in $\mathbb{R}^d$ is a subset of $\mathbb{R}^d$, defined by a *normal vector* $w \in \mathbb{R}^d \setminus \{0\}$:

$$H = \{x \in \mathbb{R}^d \ : \ w^T x \geq 0\}. \tag{6.1}$$

The boundary of a halfspace $H$ with the normal vector $w$ is a *hyperplane* $h$, defined as

$$h = \{x \in \mathbb{R}^d \ : \ w^T x = 0\}. \tag{6.2}$$

Each hyperplane corresponds to two halfspaces, depending on the direction of the normal vector.

A set of hyperplanes $h_1, \ldots, h_k$ divide $\mathbb{R}^d$ into connected regions. Formally, given a set of hyperplanes $h_1, \ldots, h_k$, a *region* is defined as an intersection $\bigcap_{i=1}^k H_i$, for a suitable selection of halfspaces $H_i$ corresponding to the hyperplanes $h_i$.

**Definition 6.2** A set of hyperplanes $h_1, \ldots, h_k$ in $\mathbb{R}^d$ is in *general position* if

$$\dim \left( \bigcap_{i \in I} h_i \right) = d - |I| \tag{6.3}$$

for all $I \subseteq \{1, 2, \ldots, k\}$ with $|I| \leq d$.

We cite the following standard fact:

**Theorem 6.3** Let $m \geq n$. Let $h_1, \ldots, h_m$ be hyperplanes in general position in $\mathbb{R}^n$ The number of regions regions is exactly

$$2\sum_{i=0}^{n-1}\binom{m-1}{i} = 2T(m-1, n-1) \leq \left(\frac{em}{n}\right)^n \tag{6.4}$$

A proof for the first equality can be found in Bürgisser and Cucker [2013], Lemma 13.7. The bound on $T(m-1, n-1)$ follows from Appendix A.1. In fact, a stronger claim holds: the general position is the worst case, in the sense that no arrangement of hyperplanes has more regions than the quantity in Equation (6.4).

In Chapters 3 and 5, we discussed regions of sets of halfspaces defined by independent standard normal vectors.

**Lemma 6.4** For $m \geq n$, let $W \in \mathbb{R}^{m\times n}$ be a matrix with independent $N(0,1)$ entries. Define the hyperplanes $h_1, \ldots, h_m$ using the rows $w_1, \ldots, w_m$ of the matrix $W$ as normal vectors. Then the number of regions $h_1, \ldots, h_m$ divide $\mathbb{R}^n$ into is equal to

$$2\sum_{i=0}^{n-1}\binom{m-1}{i} = 2T(m-1, n-1) \leq \left(\frac{em}{n}\right)^n \tag{6.5}$$

**Proof** We only need to prove the halfspaces are in general position. Any intersection of a subset of $h_i$ is the kernel of the corresponding submatrix of $W$: for any $I \subseteq \{1, 2, \ldots, m\}$,

$$\bigcap_{i\in I} h_i = \left\{x \in \mathbb{R}^n \ : \ w_i^T x = 0 \text{ for all } i \in I\right\}. \tag{6.6}$$

As the rows of $W$ are independent normal vectors, any subset of $|I| \leq n$ rows is linearly independent with probability one, hence the dimension of its kernel is equal to $n - |I|$, as required by Definition 6.2. □

Bürgisser and Cucker [2013] show that Lemma 6.4 has a straightforward corollary defining the probability of existence of any particular region.

**Proposition 6.5** For $m \geq n$, let $W \in \mathbb{R}^{m\times n}$ be a matrix with independent $N(0,1)$ entries. For a given *sign pattern* $\sigma \in \{-1, +1\}^m$, we have

$$\mathbb{P}\left[\text{there exists } x \in \mathbb{R}^n \ : \ \sigma_i w_i^T x > 0 \text{ for all } 1 \leq i \leq m\right] = \frac{1}{2^{m-1}}\sum_{i=0}^{n-1}\binom{m-1}{i}. \tag{6.7}$$

**Proof** As before, define the hyperplanes $h_1, \ldots, h_m$ using the rows $w_1, \ldots, w_m$ of the matrix $W$ as normal vectors. Any fixed region defines a sign pattern for the rows of the matrix $W$, depending in which halfspace associated with $h_i$ the region belongs, for each $1 \leq i \leq m$. No two regions define the same sign pattern.

Any sign pattern is equally likely, because the distribution of the entries of the matrix $W$ is invariant to flips $w_i \mapsto -w_i$. There are $2^m$ possible sign patterns and $2T(m-1, n-1)$ regions, hence the probability in question is

$$\frac{1}{2^m} 2T(m-1, n-1) = \frac{1}{2^{m-1}} \sum_{i=0}^{n-1} \binom{m-1}{i}. \tag{6.8}$$

□

## 6.2 The probability of a random subspace intersecting a fixed orthant

The following lemma can be found in [Morrison, 2010], although we believe it has appeared before. We give our own proof, reducing it to Proposition 6.5.

**Lemma 6.6** Given a random subspace $V \subseteq \mathbb{R}^m$, with $\dim V = n \leq m$, the probability that $V$ intersects a fixed orthant nontrivially is

$$\frac{1}{2^{m-1}} \sum_{i=0}^{n-1} \binom{m-1}{i} \leq 2^{-m} \left(\frac{em}{n}\right)^n. \tag{6.9}$$

**Proof** As in Proposition 3.3, we can identify a random $n$-dimensional subspace of $\mathbb{R}^m$ with the column space of $W \in \mathbb{R}^{m \times n}$ with independent $N(0,1)$ entries.

Any orthant $\mathcal{O}$ in $\mathbb{R}^m$ corresponds to a sign pattern $\sigma \in \{-1, +1\}^m$. Any element of the column span is of the form $Wx$ for $x \in \mathbb{R}^n$. The condition $Wx \in \mathcal{O}$ is equivalent to the condition

$$\sigma_i w_i^T x \geq 0 \text{ for all } 1 \leq i \leq m. \tag{6.10}$$

As $m \geq n$ and the distribution of $W$ (or equivalently, over subspaces) is atomless, we have

$$\mathbb{P}\left[\text{there exists } x \in \mathbb{R}^n \ : \ \sigma_i w_i^T x \geq 0 \text{ for all } 1 \leq i \leq m\right] \tag{6.11}$$

$$= \mathbb{P}\left[\text{there exists } x \in \mathbb{R}^n \ : \ \sigma_i w_i^T x > 0 \text{ for all } 1 \leq i \leq m\right] \tag{6.12}$$

$$= \frac{1}{2^{m-1}} \sum_{i=0}^{n-1} \binom{m-1}{i} \tag{6.13}$$

by Proposition 6.5. □

Note that the distribution of $W$ is rotationally invariant, so without loss of generality we may assume the orthant is the nonnegative orthant $\mathbb{R}_{\geq 0}$. However, we do not use this in our proof.

## 6.3 Almost evenly distributed spherical random vectors

In the MathOverflow post [Baghal, 2021], the following question was raised:

**Question 6.7** Consider $n$ independent random vectors $z_1, \cdots, z_n$ drawn from the uniform distribution on the $d$-dimensional sphere $\mathbb{S}^{d-1}$. What is the best lower bound on $n$ for which with high probability there exists a constant $c > 0$ such that

$$cn \leq |\{1 \leq i \leq n : \langle z_i, v\rangle > 0\}| \tag{6.14}$$

for all $v \in \mathbb{R}^d \backslash \{0\}$?

We answered the question in [Paleka, 2021] using the formula in Lemma 6.4. Our result shows that the minimal $n$ is at most linear in $d$.

**Proposition 6.8** If $n = 160d$, with probability $1 - o(d)$ we have

$$|\{1 \leq i \leq n : \langle z_i, v\rangle > 0\}| \geq \frac{n}{4} \tag{6.15}$$

**Proof** We use that the number of "distinct" vectors $v \in \mathbb{S}^{d-1}$ with respect to the classifiers $\operatorname{sgn}\langle \cdot, z_i\rangle$ is

$$2T(n-1, d-1) \leq \left(\frac{ne}{d}\right)^d, \tag{6.16}$$

because with probability 1 the kernels of the classifiers $\langle \cdot, z_i\rangle$ define a generic hyperplane arrangement. Thus, instead of considering all vectors on the sphere $\mathbb{S}^{d-1}$, we can union bound over at most $\left(\frac{ne}{d}\right)^d$ representative vectors.

Let $X = |\{1 \leq i \leq n : \langle z_i, e_1\rangle > 0\}|$ for a basis vector $e_1$, and note that $X$ is a sum of $n$ independent Bernoulli variables. Moreover, it has the same distribution if we replace $e_1$ by any other vector, because the distribution of the $z_i$ is rotationally invariant.

Then

$$\mathbb{P}\left[\text{there exists } v \in \mathbb{S}^{d-1} \text{ such that } |\{1 \leq i \leq n : \langle z_i, v\rangle > 0\}| < \frac{n}{4}\right] \tag{6.17a}$$

$$\leq \left(\frac{ne}{d}\right)^d \mathbb{P}\left[X < \frac{n}{4}\right] \tag{6.17b}$$

$$\leq \exp\left(d\log(n) - \frac{n}{16} + d - d\log d\right) \tag{6.17c}$$

$$= \exp(\log(160)d - 9d) \xrightarrow{d\to\infty} 0, \tag{6.17d}$$

where we used the union bound in the first inequality, and the Chernoff bound on $\mathbb{P}[X < n/4]$ in the second inequality, in the form

$$\mathbb{P}\left[X \leq (1-\delta)\mathbb{E}[X]\right] \leq \exp\left(-\frac{1}{2}\delta^2\mathbb{E}[X]\right). \tag{6.18}$$

□

As pointed out in the comments to Baghal [2021] by the MathOverflow user Sinan Saghal, for $n < d$ the bound cannot hold uniformly on $\mathbb{R}^d \setminus \{0\}$. This is because for any $n < d$ vectors in $\mathbb{R}^d$, there exists a vector orthogonal to all of them. Hence, the optimal $n$ in Question 6.7 is linear in $d$, with the constant yet to be determined.

Appendix A

# Deferred proofs

## A.1 Bounds on $T(a, b)$

Let $a, b$ be positive integers such that $a \geq 2b$. We defined the prefix-sum of binomial coefficients

$$T(a, b) = \sum_{i=0}^{b} \binom{a}{i}. \tag{A.1}$$

By definition, $T(a, b) \geq \binom{a}{b}$.

For $0 \leq x \leq 1$, let $H(x) = -x \log_2(x) - (1 - x) \log_2(x)$. We have the upper bound

$$T(a, b) = \sum_{i=0}^{b} \binom{a}{b} \leq 2^{aH(b/a)}. \tag{A.2}$$

For a proof, see e.g. Theorem 3.1 in Galvin [2014]. The bound is approximately sharp for $a/b$ constant.

We mostly use a slightly weaker bound, because is it nicer to plug in:

$$T(a - 1, b - 1) \leq \frac{1}{2} \left( \frac{ea}{b} \right)^b. \tag{A.3}$$

## A.2 Number of activation regions

There are several papers dealing with the maximum number of. Most prominently, Hanin and Rolnick [2019] deal with random networks where the weights and biases are independent with a continuous probability distribution. Given some smoothness assumptions, they show that the number of activation regions should scale as

$$\frac{(O(d_0 + d_1 + \ldots + d_L)))^{d_0}}{d_0!}, \tag{A.4}$$

which is exponential only in the input dimension, and polynomial in the depth. Unfortunately, their proof does not generalize to the zero bias case, which makes it incompatible with the general theme of this work.

Therefore, we opt to use the simple bound, which is exponential in depth. Since we didn't find a short proof of this statement in the literature, we present a simple proof by induction.

**Lemma 5.11** The number of activation regions of $f_\ell$ is at most $\left(e^\ell \prod_{i=1}^{\ell} \frac{d_i}{d_0}\right)^{d_0}$.

**Proof** We use induction. The case $\ell = 1$ follows from Lemma 5.9. The prefix network $f_{\ell-1}$ maps each activation region $\mathcal{A}$ in $\mathbb{R}_0^d$ to a cone in $\mathbb{R}^{d_{\ell-1}}$, contained in a subspace $V$ with $\dim V = d_0$.

Due to rotational invariance, the weight matrix $W^{(\ell)}$ acts on $V$ like a $\mathbb{R}^{d_0 \times d_\ell}$ i.i.d Gaussian matrix. This means that the activation region $\mathcal{A}$ splits up into at most $\left(\frac{d_\ell}{d_0}\right)^{d_0}$ regions, hence the bound in question. □

If the results in Hanin and Rolnick [2019] would generalize to the zero bias case, we could plausibly get a better dependency of the expansivity ratio in Theorem 5.1, for example $C \geq \log L$.

## A.3 Proof of Lemma 4.19

Recall the definition of $F(d, s, p)$ and the statement we want to prove.

**Definition 4.18** For nonnegative integers $d, s, p$, define $F(d, s, p)$ to be the number of families of $p$ distinct subsets $S_1, \ldots, S_p \subseteq \{1, 2, \ldots, d\}$, with each subset having at most $s$ elements, and

$$S_1 \cap \ldots \cap S_p = \varnothing, \tag{4.22}$$

$$S_1 \cup \ldots \cup S_p = \{1, 2, \ldots, d\}. \tag{4.23}$$

Note that $F$ is well-defined for $d = 0$: for all $s \geq 0$ and $p \geq 2$, we have $F(0, s, 1) = 1$ and $F(0, s, p) = 0$.

**Lemma 4.19** With $F(d, s, p)$ defined as in Definition 4.18,

$$\sum_{p \geq 1} (-1)^{p+1} F(d, s, p) = \begin{cases} (-1)^d & \text{if } s \geq d; \\ 0 & \text{otherwise.} \end{cases} \tag{4.25}$$

for all $d, s \geq 1$.

To prove Lemma 4.19, we introduce two "relaxations" of $F$, which will be used to express $F$ using inclusion-exclusion. The definition of $G$ drops the empty intersection condition, and the definition of $H$ drops the set cover condition too.

**Definition A.1** For nonnegative integers $d, s, p$, define $G(d, s, p)$ to be the number of families of $p$ distinct subsets $S_1, \ldots, S_p \subseteq \{1, 2, \ldots, d\}$, with each subset having at most $s$ elements, and

$$S_1 \cup \ldots \cup S_p = \{1, 2, \ldots, d\}. \tag{A.5}$$

**Definition A.2** For nonnegative integers $d, s, p$, define $H(d, s, p)$ to be the number of families of $p$ distinct subsets $S_1, \ldots, S_p \subseteq \{1, 2, \ldots, d\}$, with each subset having at most $s$ elements.

It's easy to see that $H$ has a simple closed form:

$$H(d, s, p) = \binom{T(d, s)}{p}. \tag{A.6}$$

We will proceed by proving Lemma 4.19 directly, without calculating $F$ as an intermediate step. Note that in principle it is possible to calculate $F$ explicitly by inclusion-exclusion, but the resulting expressions are tedious to deal with.

**Proof (of Lemma 4.19)** We can use inclusion-exclusion to decompose $F$ over the empty intersection condition. The number of set families counted by $G$ and having $k$ fixed elements in their intersection is exactly $G(d - k, s - k, p)$ if $k \leq \min(d, s)$, and zero otherwise. Thus

$$F(d, s, p) = \sum_{k=0}^{\min(d,s)} (-1)^k \binom{d}{k} G(d - k, s - k, p). \tag{A.7}$$

Analogously, we can decompose $H$ over the set cover condition. The number of set families counted by $H$ and missing $k$ fixed elements in their union is exactly $H(d - k, s, p)$, hence

$$G(d, s, p) = \sum_{k=0}^{d} (-1)^k \binom{d}{k} H(d - k, s, p). \tag{A.8}$$

The binomial theorem applied on $(1 - 1)^{T(d,s)} = 0$ yields

$$\sum_{p \geq 1} (-1)^{p+1} H(d, s, p) = 1, \tag{A.9}$$

since Equation (A.6) gives $H(d, s, p) = \binom{T(d,s)}{p}$, and $T(d, s) \geq 1$ for $d, s \geq 0$.

Hence

$$\sum_{p\geq 1}(-1)^{p+1}G(d,s,p) \tag{A.10}$$

$$=\sum_{k=0}^{d}(-1)^k\binom{d}{k}\sum_{p\geq 1}(-1)^{p+1}H(d-k,s,p) \tag{A.11}$$

$$=\sum_{k=0}^{d}(-1)^k\binom{d}{k} \tag{A.12}$$

$$=\begin{cases}1 & \text{if } d=0;\\ 0 & \text{otherwise.}\end{cases} \tag{A.13}$$

Finally, we can calculate the alternating sum in Equation (4.25):

$$\sum_{p\geq 1}(-1)^{p+1}F(d,s,p) \tag{A.14}$$

$$=\sum_{k=0}^{\min(d,s)}(-1)^k\binom{d}{k}\sum_{p\geq 1}(-1)^{p+1}G(d-k,s-k,p) \tag{A.15}$$

$$=\sum_{k=0}^{\min(d,s)}(-1)^k\binom{d}{k}\,\mathbb{1}_{d=k} \tag{A.16}$$

$$=\begin{cases}(-1)^d & \text{if } s\geq d;\\ 0 & \text{otherwise.}\end{cases} \tag{A.17}$$

□

## A.4 Proof of Lemma 4.20

**Lemma 4.20**

$$\sum_{j=0}^{\ell}\binom{m-n+\ell}{j}=\left[x^\ell\right](1-x)^{-m+n}(1-2x)^{-1}. \tag{4.36}$$

**Proof** Note that multiplying a generating function by $(1-x)^{-1}$ is equivalent to a "partial sum" of the coefficients. Using the binomial theorem,

$$\sum_{j=0}^{\ell}\binom{m-n+\ell}{j}=\left[z^\ell\right](1+z)^{m-n+\ell}(1-z)^{-1} \tag{A.18}$$

Recall the Lagrange–Bürmann formula [Wikipedia, 2020] in the form:

**Theorem A.3 (Lagrange–Bürmann formula)** Let $\phi$ be a formal power series with $\phi(0)\neq 0$. Let $f$ and $g$ be formal power series such that $f(w)\phi(w)=w$ and $f(g(z))=z$. Then, for any formal power series $H$ and for any integer $k$,

$$\left[w^k\right]H(w)\phi(w)^{k-1}(\phi(w)-w\phi'(w))=\left[z^k\right]H(g(z)). \tag{A.19}$$

We apply the formula with $k = \ell$, $H(x) = (1-w)^{-m+n-\ell+1}(1-2w)^{-1}$, $\phi(w) = 1-w$, $f(w) = \frac{w}{1-w}$ and $g(z) = \frac{z}{1+z}$ to get

$$\left[z^\ell\right](1+z)^{m-n+\ell}(1-z)^{-1} \tag{A.20}$$

$$= \left[w^\ell\right](1-w)^{-m+n-\ell+1}(1-2w)^{-1}(1-w)^{\ell-1}(1-w+w) \tag{A.21}$$

$$= \left[w^\ell\right](1-w)^{-m+n}(1-2w)^{-1}, \tag{A.22}$$

which is exactly what we wanted to prove. □

## A.5 Saddle points

Recall Equation (4.42b):

$$f(z) = \log\left(\frac{1}{2} - z\right) - \log(z) - K\log(1-z). \tag{A.23}$$

All saddle points of $f$ are solutions to the equation

$$f'(z) = 0 \;\leftrightarrow\; \frac{1}{z-\frac{1}{2}} - \frac{1}{z} + \frac{K}{1-z} = 0 \tag{A.24}$$

$$\leftrightarrow\; 2Kz^2 - (K+1)z + 1 = 0, \tag{A.25}$$

with the solution closest to the origin being

$$z_0 = \frac{-\sqrt{K^2-6K+1}+K+1}{4K}. \tag{A.26}$$

## A.6 Proof of Lemma 4.21

**Lemma 4.21** The expression

$$(K+1)n\; H\left(\frac{1}{K+1}\frac{n-2i}{n}\right) - 2i - 2i\log_2\left|\frac{1}{2} - z_0\right| + n\log_2(e)f(z_0) \tag{4.49}$$

is stricly decreasing in $i$ on the set $\{0, 1, \ldots, \lfloor n/2 \rfloor\}$. In particular, the approximated $\log_2|U(m,n,2i)|$ is maximal in

$$\log_2|U(m,n,0)| \approx n\left((K+1)\; H\left(\frac{1}{K+1}\right) + \log_2(e)f(z_0)\right) \tag{4.50}$$

**Proof** The derivative of the binary entropy $H$ with respect to its argument is

$$\frac{\mathrm{d}}{\mathrm{d}p}H(p) = -\log_2\frac{p}{1-p}. \tag{A.27}$$

We differentiate the expression in Equation (4.49) with respect to $i$:

$$\frac{\mathrm{d}}{\mathrm{d}i}\left((K+1)n\ H\left(\frac{1}{K+1}\frac{n-2i}{n}\right)-2i-2i\log_2\left|\frac{1}{2}-z_0\right|+n\log_2(e)f(z_0)\right) \tag{A.28a}$$

$$=-(K+1)n\log_2\left(\frac{\frac{1}{K+1}\frac{n-2i}{n}}{1-\frac{1}{K+1}\frac{n-2i}{n}}\right)\left(-\frac{2}{(K+1)n}\right)-2-2\log_2\left|\frac{1}{2}-z_0\right| \tag{A.28b}$$

$$=2\log_2\left(\frac{n-2i}{Kn+2i}\right)-2-2\log_2\left|\frac{1}{2}-z_0\right| \tag{A.28c}$$

$$=2\log_2\left(\frac{n-2i}{Kn+2i}\frac{1}{|1-2z_0|}\right)<0, \tag{A.28d}$$

where the last inequality is due to the elementary computation

$$K|1-2z_0|=K\left(1-2\frac{-\sqrt{K^2-6K+1}+K+1}{4K}\right) \tag{A.29}$$

$$=\frac{1}{2}\left(K-1+\sqrt{K^2-6K+1}\right)>1. \tag{A.30}$$

□

## A.7 Proof of Proposition 4.22

**Proposition 4.22** The expression

$$-K+(K+1)H\left(\frac{1}{K+1}\right)+\log_2(e)f(z_0) \tag{4.52}$$

is decreasing on $K>3+2\sqrt{2}\approx 5.8$, and has an unique root

$$K_0\approx 7.34463. \tag{4.53}$$

We do not prove this completely rigorously due to messy calculations; as the expression in question is a "reasonable" function of $K$, we think graphical evidence in Figure A.1 is almost sufficient.

**Proof (Nonrigorous)** Recall the definitions of the function $f$ and the saddle point $z_0$:

$$f(z)=\log\left(\frac{1}{2}-z\right)-\log(z)-K\log(1-z) \tag{A.31a}$$

$$z_0=\frac{-\sqrt{K^2-6K+1}+K+1}{4K}. \tag{A.31b}$$

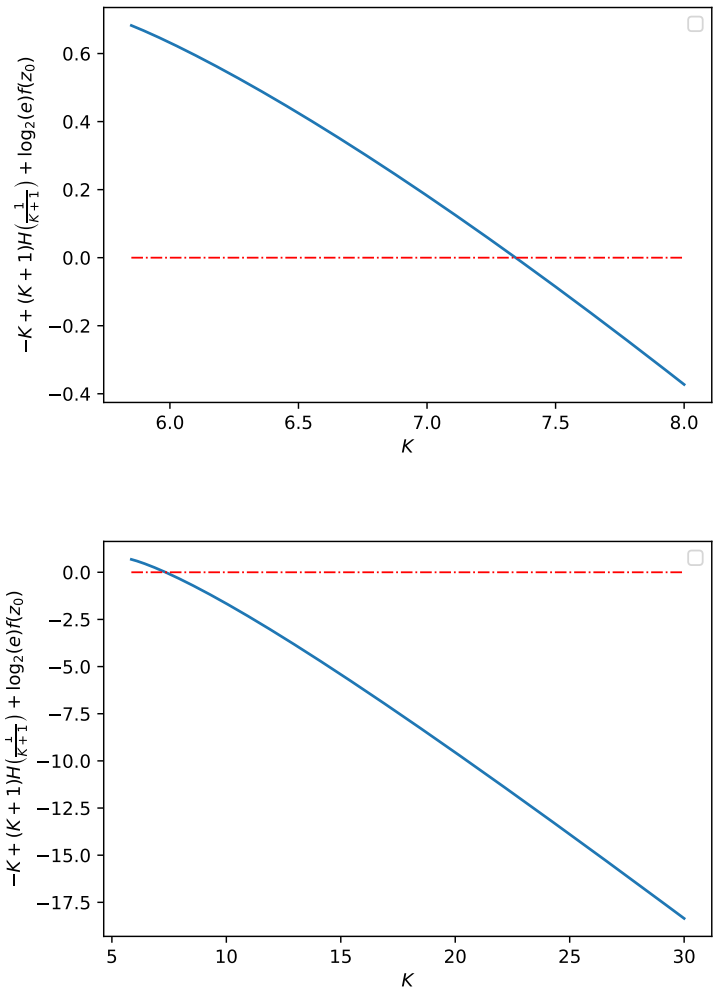



**Figure A.1:** The behaviour of the expression in Equation (4.52), on the critical interval and globally.

The derivative of $z_0$ with respect to $K$ is:

$$\frac{\mathrm{d}z_0}{\mathrm{d}K} = \frac{\mathrm{d}}{\mathrm{d}K}\frac{-\sqrt{K^2-6K+1}+K+1}{4K} = \frac{-\sqrt{K^2-6K+1}-3K+1}{4K^2\sqrt{K^2-6K+1}}. \tag{A.32}$$

The derivative of the expression in question is then

$$\frac{\mathrm{d}}{\mathrm{d}K}\left(-K+(K+1)H\left(\frac{1}{K+1}\right)+\log_2(e)\left(\log\left(\frac{1}{2}-z_0\right)-\log(z_0)-K\log(1-z_0)\right)\right) \tag{A.33}$$

$$= -1-\frac{\log_2 K}{K+1}+\log_2(e)\left(2\log(2)-\log\left(\frac{\sqrt{K^2-6K+1}+3K-1}{K}\right)\right), \tag{A.34}$$

which can be seen to be less than zero for $K > 3+2\sqrt{2} \approx 5.8$. The graph crosses 0 in a unique point $K_0$, which can be numerically estimated to be close to 7.34463. □

## A.8 Proof of Lemma 5.16

**Lemma 5.16** Let $m = Cn$, for $C$ a large enough constant. Consider a random vector $z \in \mathbb{R}^m$ drawn from any rotationally invariant distribution. With probability $1-o(n)$, all $y \in \mathcal{B}_{\text{angle}}\left(z, \frac{1}{2}\right)$ have at least $2n$ positive coordinates.

**Proof** We want to show $\mathcal{B}_{\text{angle}}(z, \frac{1}{2})$ intersects a small number of half-open orthants. This will imply that it has low probability of intersecting a half-open orthant with few pluses.

Consider a fixed orthant $O_S^m \subseteq \mathbb{R}^m$.

$$\mathbb{P}_{z\sim\text{Unif}(\mathbb{S}^{m-1})}\left[\mathcal{B}_{\text{angle}}\left(z,\frac{1}{2}\right)\bigcap O_S^m \neq \{0\}\right] \tag{A.35a}$$

$$= \mathbb{P}_{z\sim\text{Unif}(\mathbb{S}^{m-1})}\left[\mathcal{B}_{\text{angle}}\left(z,\frac{1}{2}\right)\text{ intersects the nonnegative orthant}\right] \tag{A.35b}$$

$$= \mathbb{P}_{z\sim\text{Unif}(\mathbb{S}^{m-1})}\left[\text{ there exists } y\in\mathbb{S}^{m-1},\ y\geq 0 \text{ for which } z^T y \geq \frac{3}{4}\right] \tag{A.35c}$$

$$= \mathbb{P}_{z\sim\text{Unif}(\mathbb{S}^{m-1})}\left[\left\|z_{\text{pos}}\right\|^2 \geq \frac{3}{4}\right] \tag{A.35d}$$

$$= \mathbb{P}_{z\sim\text{Unif}(\mathbb{S}^{m-1})}\left[\left\|z_{\text{neg}}\right\| \leq \frac{1}{2}\right] \tag{A.35e}$$

Now, $\left\|z_{\text{neg}}\right\|$ intuitively concentrates very well around $\frac{\sqrt{2}}{2}$, so we expect exponential decay in Equation (A.35e).

We finish with a standard computation. Write the uniform distribution on the sphere as a normalized standard Gaussian, so $z \propto (g_1, g_2, \ldots, g_m)$ and

$$\left\|z_{\text{neg}}\right\|^2 = \frac{\sum_{i \in S} g_i^2}{\sum_{i=1}^m g_i^2}. \tag{A.36}$$

Let $S \subseteq \{1, 2, \ldots, m\}$ be the set of nonzero coordinates of $z_{\text{neg}}$.
Then $|S| \sim B\left(m, \frac{1}{2}\right)$, so using Chernoff gives

$$\mathbb{P}\left[|S| > 0.505m\right] \leq \exp\left(-\frac{m}{6} \cdot (0.005)^2\right). \tag{A.37}$$

Conditioning on the complement of the above, we assume $|S| \leq 0.505m$. Using the standard chi-square bounds (see Lemma 1 from Laurent and Massart [2000]):

$$\mathbb{P}\left[\sum_{i \in S} g_i^2 \leq |S| - 2\sqrt{|S|x}\right] \leq \exp\left(-x\right) \tag{A.38}$$

$$\mathbb{P}\left[\sum_{i=1}^m g_i^2 \geq m + 2\sqrt{mx} + 2x\right] \leq \exp\left(-x\right) \tag{A.39}$$

Pick $x = \frac{m}{40000}$, and condition on the complements again. After straightforward calculations, we get

$$\mathbb{P}\left[\left\|z_{\text{neg}}\right\|^2 \leq \frac{1}{4}\right] \leq c_1 \exp(-c_0 m) = c_1 \exp(-c_0 C)^n \tag{A.40}$$

with $c_1 = 3$ and $c_0 = 10^{-6}$.

Let $\Omega_{m,n}$ be the family of sets in $\{1, 2, \ldots, m\}$ with $|S| \leq 2n$. The sets in $\Omega_{m,n}$ correspond to all half-open orthants with at most $2n$ pluses in $\mathbb{R}^m$.

Due to the bounds in Appendix A.1, we have $|\Omega_{m,n}| = T(m, 2n) \leq (eC/2)^{2n}$. We union bound over all orthants corresponding to the sets in $\Omega_{m,n}$:

$$\mathbb{P}_{z \sim \text{Unif}(\mathbb{S}^{m-1})}\left[\exists y \in \mathcal{B}_{\text{angle}}\left(z, \frac{1}{2}\right) \text{ with } \leq 2n \text{ positive coordinates}\right] \tag{A.41}$$

$$\leq \sum_{S \in \Omega} \mathbb{P}_{z \sim \text{Unif}(\mathbb{S}^{m-1})}\left[\mathcal{B}_{\text{angle}}\left(z, \frac{1}{2}\right) \bigcap O_S^m \neq \{0\}\right] \tag{A.42}$$

$$\leq (eC/2)^{2n} c_1 \exp(-Cc_0)^n \tag{A.43}$$

$$\leq c_1 (e^2 C^2 \exp(-Cc_0))^n, \tag{A.44}$$

and the expression under the exponent can be arbitrarily small if we take $C$ a large enough constant. □

Note that we can show a stronger statement: if we replace $\frac{1}{2}$ in Lemma 5.16 with any $\varepsilon < \frac{\sqrt{2}}{2}$, the probability in Lemma 5.16 still decays exponentially.

Appendix B

# More on intrinsic volumes

## B.1 Intrinsic volumes of orthants

**Lemma 4.16** For $0 \leq k \leq d$, the intrinsic volume $v_k$ of the nonnegative orthant $\mathcal{O}^d_{\{1,2,\dots,d\}} = \mathbb{R}^d_{\geq 0} \subseteq \mathbb{R}^d$ is

$$v_k(\mathbb{R}^d_{\geq 0}) = \frac{1}{2^d}\binom{d}{k}. \tag{4.18}$$

**Proof** Sample a point $g \sim N(0, I_d)$. The projection of $g$ to $\mathbb{R}^d_{\geq 0}$ is equal to $g_{\text{pos}} = \text{ReLU}(g)$. The relative dimension of the face of $\mathbb{R}^d_{\geq 0}$ that $g_{\text{pos}}$ belongs to is equal to the number of positive coordinates of $g$. As the coordinates of a standard normal vector are independent Bernoulli variables with parameter $\frac{1}{2}$, the relative dimension has the distribution of a $B(d, \frac{1}{2})$ binomial variable.

The intrinsic volume $v_k$ is just the probability that the relative dimension is $k$, which is equal to $\frac{1}{2^d}\binom{d}{k}$ for a $B(d, \frac{1}{2})$ variable. □

**Lemma 4.17** For $0 \leq d \leq m$, let $F : \mathbb{R}^d \to \mathbb{R}^m$ be an isometric linear embedding. Then, for $0 \leq k \leq d$,

$$v_k(F\mathbb{R}^d_{\geq 0}) = \frac{1}{2^d}\binom{d}{k}, \tag{4.19}$$

and $v_k(F\mathbb{R}^d_{\geq 0}) = 0$ for $k > d$.

**Proof** Any isometric linear embedding $F : \mathbb{R}^d \to \mathbb{R}^m$ can be decomposed as $F = QI_{d\times m}$, where $Q \in \mathbb{R}^{m\times m}$ is orthogonal and $I_{d\times m} \in \mathbb{R}^{d\times m}$ is the identity embedding from $\mathbb{R}^d$ to the first $d$ coordinates of $\mathbb{R}^m$.

If $g \sim N(0, I_m)$, $Q^T g$ has the same distribution as $g$, because $Q$ is orthogonal and the distribution is rotationally invariant. Therefore we may assume $Q = I$ and $F = I_{d\times m}$ without loss of generality.

Sample a point $g \sim N(0, I_m)$. The projection of $g$ to $I_{m\times d}\mathbb{R}^d_{\geq 0}$ is equal to

$$(\mathrm{ReLU}(g_1), \ldots, \mathrm{ReLU}(g_d), 0, 0, \ldots, 0) \tag{B.1}$$

The first $d$ coordinates of a standard normal vector in $\mathbb{R}^m$ are distributed as $N(0, I_d)$. Hence, relative dimension of the face of $I_{m\times d}\mathbb{R}^d_{\geq 0}$ the projection gets mapped to is again distributed as $B(d, \frac{1}{2})$, and we can finish as in the proof of Lemma 4.16. □

## B.2 Defining intrinsic volumes for nonconvex cones

This section is a short summary of Section 6.5 in Schneider and Weil [2008].

We call a subset of $\mathbb{S}^{m-1}$ *spherically convex* if it's an intersection of a convex cone with $\mathbb{S}^{m-1}$. Let $\mathcal{K}^s$ be the set of all spherically convex sets. The *spherical convex ring* $\mathcal{R}^s$, as defined in Schneider and Weil [2008], is the family of all finite unions of spherically convex subsets of $\mathbb{S}^{m-1}$. As the intersection of two convex sets is convex, we can say that $\mathcal{K}^s$ generates $\mathcal{R}^s$ by finite unions and intersections.

We say that a real-valued function is *additive* on some domain if it satisfies Equation (4.6). For example, a map $F : \mathcal{K}^s \to \mathbb{R}$ is additive if

$$F(A \cap B) + F(A \cup B) = F(A) + F(B) \tag{B.2}$$

for all $A, B \in \mathcal{K}^s$ such that $A \cup B \in \mathcal{K}^s$.

We say that a map $F : \mathcal{K}^s \to \mathbb{R}$ is a *valuation* whenever it is additive and $F(\varnothing) = 0$. The following classical theorem allows us to extend valuations from the generating family $\mathcal{K}^s$ to the full space $\mathcal{R}^s$:

**Theorem B.1 (Groemer's extension theorem)** Every continuous valuation $F : \mathcal{K}^s \to \mathbb{R}$ has an additive extension to $\mathcal{R}^s$.

For a proof, see Theorem 14.4.2 in Schneider and Weil [2008]. They prove a generalization for valuations with values in any topological vector space.

Using the additive functional equation, we can easily get the expression for the extended valuation. For any $S \in \mathcal{R}^s$, we can represent $S = \cup_{i=1}^n K_i$ for some $K_1, \ldots, K_n \in \mathcal{K}^s$.

$$F(S) = \sum_{\varnothing \neq J \subseteq I} (-1)^{|J|+1} F\left(\bigcap_{i\in J} K_i\right). \tag{B.3}$$

We can see that $F(S)$ is well-defined using the additivity of $F$ on $\mathcal{K}^s$. For any two representations of $S$ as a finite union of spherically convex sets, we can expand them to the "lowest common denominator" representation, and then

apply additivity of $F$ to prove that the right-hand sides of Equation (B.3) are the same.

The intrinsic volumes $v_k$, as defined in Definition 4.5, are not defined on spherically convex sets, but convex cones. However, we can uniquely identify each spherically convex set with a nonempty convex cone. Thus by abuse of notation we can define

$$v_k(S) \stackrel{\text{def}}{=} v_k(C) \tag{B.4}$$

for $S \in \mathcal{K}^s$ and $C$ the unique convex cone such that $C \cap \mathbb{S}^{d-1} = S$.

We can easily check that $v_k$ for $k \geq 1$ are additive on $\mathcal{K}^s$; for a reference, see Theorem 6.5.2 in Schneider and Weil [2008]. The exceptional $v_0$ as defined in Definition 4.5 is not additive, but this is not relevant for our work, as we do not use $v_0$ in Chapter 4.

The defining Equation (4.6) for $v_k(C)$ in Definition 4.8 corresponds to applying Equation (B.3) to the valuation $v_k$.

# Bibliography


Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer New York, 2007. doi: 10.1007/978-0-387-48116-6. URL http://dx.doi.org/10.1007/978-0-387-48116-6.

Robert J. Adler, Jonathan E. Taylor, and Keith J. Worsley. *Applications of Random Fields and Geometry*. 2015. doi: 10.1007/978-0-387-48116-6. URL https://web.stanford.edu/class/stats317/hrf.pdf.

Max Alekseyev. Asymptotics of an alternating sum involving the prefix sum of binomial coefficients. MathOverflow, 2021. URL https://mathoverflow.net/q/402256.

Dennis Amelunxen and Martin Lotz. Intrinsic volumes of polyhedral cones: A combinatorial perspective, 2017.

Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data, 2013.

Sina Baghal. Almost evenly distributed spherical random vectors. MathOverflow, 2021. URL https://mathoverflow.net/q/397443.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models, 2017.

Joan Bruna, Arthur Szlam, and Yann LeCun. Signal recovery from pooling representations, 2013.

Peter Bürgisser and Felipe Cucker. *Condition: The geometry of numerical algorithms*, volume 349. Springer Science & Business Media, 2013.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming, 2020.

Charles Clum. Private communication, 2021a.

Charles Clum. Doctoral thesis. To be published, 2021b.

Hadi Daneshmand, Jonas Kohler, Francis Bach, Thomas Hofmann, and Aurelien Lucchi. Batch normalization provably avoids rank collapse for randomly initialised deep networks, 2020.

Claudio Gallicchio and Simone Scardapane. Deep randomized neural networks. *Studies in Computational Intelligence*, page 43–68, 2020. ISSN 1860-9503. doi: 10.1007/978-3-030-43883-8_3. URL http://dx.doi.org/10.1007/978-3-030-43883-8_3.

David Galvin. Three tutorial lectures on entropy and counting, 2014.

LLC Gurobi Optimization. Gurobi optimizer reference manual, 2021. URL http://www.gurobi.com.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks, 2018.

Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture, 2018.

Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

Reinhard Heckel and Mahdi Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2018.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, Nov 2021. ISSN 1939-3539. doi: 10.1109/tpami.2020.2992934. URL http://dx.doi.org/10.1109/TPAMI.2020.2992934.

Antoine Labatie. Characterizing well-behaved vs. pathological deep neural networks, 2019.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. ISSN 00905364. URL http://www.jstor.org/stable/2674095.

Qi Lei, Ajil Jalal, Inderjit S. Dhillon, and Alexandros G. Dimakis. Inverting deep generative models, one layer at a time, 2019.

Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need, 2020.

Kent E. Morrison. The probability that a subspace contains a positive vector, 2010.

Daniel Paleka. Almost evenly distributed spherical random vectors. MathOverflow, 2021. URL https://mathoverflow.net/q/397464.

Michael Puthawala, Konik Kothari, Matti Lassas, Ivan Dokmanić, and Maarten de Hoop. Globally injective ReLU networks, 2020.

Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network?, 2020.

R. Schneider and W. Weil. *Stochastic and Integral Geometry*. Probability and its applications. Springer, Springer series in statistics, 2008.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation, 2016.

Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. A correspondence between random neural networks and statistical field theory, 2017.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128(7):1867–1888, Mar 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01303-4. URL http://dx.doi.org/10.1007/s11263-020-01303-4.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Wikipedia. Lagrange–Bürmann formula- – Wikipedia, the free encyclopedia, 2020. URL https://en.wikipedia.org/w/index.php?title=Lagrange_inversion_theorem&oldid=1041106502#Lagrange%E2%80%93B%C3%BCrmann_formula. [Online; accessed 21-November-2021].

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks, 2020.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, 2014.