A CONCENTRATION RESULT FOR MULTILAYER FEEDFORWARD NEURAL NETWORKS

VERA KOPONEN

Abstract. We consider for an arbitrary fixed ρ and for each positive integer n a multilayer feedforward artificial neural network with ρ layers, n neurons in the first layer (the input layer) and only one neuron, the output neuron, in the last layer. Very roughly formulated, the main result is that if the distribution of weights of connections from a layer to the next are, for all large n, approximated well by a fixed continuous (but otherwise arbitrary) curve which does not depend on n, and if the values of the n input neurons are independently and identically distributed with a continuous probability density function, then there is a number ψ such that for all $\varepsilon > 0$ the probability that the value of the output neuron is in $[\psi-\varepsilon, \psi+\varepsilon]$ tends to 1 as n tends to infinity.

arXiv:2608.15335v1 [cs.AI] 15 Aug 2026

1. Introduction

Artificial neural networks (NNs) [24] are widely used in many different contexts and often successfull in predicting outcomes or classifying objects (e.g. images). But they do not produce any explanations, or justifications, of their predictions, or classifications [28]. This raises ethical and legal questions about their use, and it motivates efforts to understand the internal behaviour of NNs.

We will consider the most basic type of NN, the multilayer feedforward NN in which every neuron is assigned a unique nonnegative real number, its value (or signal strength), once all of its input neurons have been assigned values. We also assume that the NN has only one output neuron, and that every neuron, except the input neurons, is connected to all neurons in the previous layer. In this setting three things determine the value of the output neuron. Firstly, the values of the input neurons. Secondly, the weights (which are nonnegative real numbers) of the connections from one layer to the next. Thirdly, the so-called activation function which is a specification of how to assign a value to a (not input) neuron based on the values of neurons in the previous layer and the weights of connections from neurons in the previous layer.

We wish to understand how the value of the output neuron correlates with the properties of the NN such as: the number of layers, the number of neurons in each layer, the distribution of weights of connections from one layer to the next, and the values of the input neurons. For a specific NN the value out the output neuron can be exactly computed given complete information about the values of the input neurons and complete information about the NN itself, including knowing the weight of each individual connection. But here we are interested in finding more general correlations for large NNs. The focus on large NNs is motivated by the fact that such are common and that they are harder to analyse as many weighted connections are involved in producing the output of the NN. That is, we wish to be able to make conclusions about the distribution of the values of the output neuron given only partial information about the NN. In this study such partial information will be the number of layers, some very rough bounds on the number of neurons in each layer, information about the approximate distribution of weights. In addition we will assume that the activation functions of neurons are

Date: 15 August, 2026.

"sufficently nice", where the use of averages and taking maxima or minima falls in the category of "sufficiently nice".

When using an NN we probably do not know the exact values given to the input neurons. But we may have some knowledge, or at least an educated guess, about the probabilistic distribution of values to the input neurons. So given partial information about an NN (ask explained above), a probability distribution on the set of possible combinations of values to the input neurons and an interval I, we can ask what the probability is that the value of the output neuron belongs to I.

In this article we prove a "concentration result" for the output neuron, formally stated as Theorem 3.3 and informally stated below. We assume throughout that for some constant $\mathfrak{m} \geq 1$ the values of neurons and weights of connections are always in the interval $[0, \mathfrak{m}]$.

Main result, in a simplified formulation. *Suppose that for every large enough integer n, $\mathcal{B}_n$ is an NN with $\rho \geq 2$ layers, n neurons in the first layer, the input layer, and with a single output neuron in layer ρ. For $i = 2, \ldots, \rho$, let $W_i : [0, \mathfrak{m}] \to [0, \mathfrak{m}]$ be continuous with $\int_{[0,\mathfrak{m}]} W_i(x)dx = 1$, and suppose that the following hold:*

1. *For every interval $I \subseteq [0, \mathfrak{m}]$ any non input neuron b, the proportion of connections to b with weight in I from neurons in the previous layer is between $\int_I W_i(x)dx - \delta$ and $\int_I W_i(x)dx + \delta$.*
2. *For all $i = 2, \ldots, \rho$, the activation function for neurons in layer i is defined by using a continuous aggregation function F_i, in the sense of Definition 2.2 below, and a continuous function $g_i : [0, \mathfrak{m}]^2 \to [0, \mathfrak{m}]$, as explained in (A8)*

If the values of the neurons of the input layer are independently and identically distributed with a continuous probability density function, then there is $\psi \in [0, \mathfrak{m}]$ such that for all $\varepsilon > 0$, the probability that the value of the output neuron is in $[\psi - \varepsilon, \psi + \varepsilon]$ tends to 1 as $n \to \infty$.

The proof of the main result shows that if $\mathcal{B}_n$ is as assumed above and n is large, then there is an NN $\mathcal{B}'_n$ with only 2 layers and n input neurons such that with high probability the value of the output neuron of $\mathcal{B}'_n$ is in $[\psi - \varepsilon, \psi - \varepsilon]$ for the same ψ as above. This is Corollary 3.4.

It seems to be impossible to get the conclusion of the theorem above without having some knowledge about the weights of the NN, as for example expressed by (1) above, because the values of non input neurons depend on the weights of connections and the values of the input neurons. One can find out the distribution of weights of connections to any neuron from neurons in the previous layer with relatively small computational resources: Suppose that the NN has ρ layers (where ρ is fixed) and at most n neurons in each layer. Given any non input neuron we can sort, by size, the weights of the at most n incomming connections in time $O(n \log n)$. There are at most $(\rho - 1)n$ non input neurons so the time needed for repeating this for all non input neurons is bounded by $O(n^2 \log n)$. After sorting the weights of the incomming connections to a neuron b we can inspect if their distribution follows, up to a small deviation, a sufficiently uncomplicated continuous curve (represented by W_i if b belongs to the i^{th} layer) relative to the number of incomming connections to b. Thus checking if condition (1) holds is computationally feasible.

Remark 4.6 explains how the main theorem can straightforwardly be generalized to a context where each neuron has several *features* where each feature has its own value.

Related work and further research. Convergence properties of graph neural networks (GNNs) have been studied by Adam-Day et. al. in [1, 2] for certain probability distributions on the underlying graph. However the GNNs of [1, 2] do not consider

weights of connections from one layer to the next. Once the random graph is chosen in [1, 2] the values of the node features are updated a certain number of steps by means of aggregating (locally or globally) over the values of node features in the previous step and using a Lipschitz continuous updating function.

From a technical point of view the most closely related work is probably [19]. This study is also inspired by studies on convergence phenomena such as [8, 14, 16, 17, 18, 20, 21, 22, 23, 25, 26, 27] in the field of Statistical Relational Artificial Intelligence [5, 9, 11, 15].

The main result of the present study has rather strong assumptions. For example, it is assumed that if b_1 and b_2 are neurons in the same non input layer then the weights of the incomming connections to b_1 follow (roughly) the same distribution as the weights of the incomming connection sto b_2. I believe that this assumption can be replaced by weaker ones which allow neurons in the same layer to have different distributions of incomming weights. Another direction of generalization is obviously to consider other probability distributions on the values of the input neurons. Under such weaker assumptions, the value of the output neuron need not converge (as $n \to \infty$) almost surely to a single value. Instead, if we are given an interval $I \subseteq [0, \mathfrak{m}]$ we may be able to estimate the probability that the value of the output neuron belongs to I.

Preliminaries. By $\mathbb{N}^+$ we denote the set of positive integers. For a positive integer n, $[n] := \{1, \dots, n\}$. If S is a set then $|S|$ is its cardinality which we also call size if S is finite. Finite sequence may be denoted by $\bar{s}$, for some symbol s, and then $|\bar{s}|$ denotes the length of the sequence.

In order to formalize the notions of a multilayer feedforward neural network (NN) and its states the concept of continuous, or general, structure (see e.g. [4, 6]) will be used. This concept generalizes the concept of "discrete" structure (see e.g. [10, 12, 13]) which includes for example graphs in the usal sense of discrete mathematics. But besides the basic definitions given below no further knowledge of model theory is necessary to understand the results and proofs.

By a *signature* we mean a finite set of so-called *relation symbols*, each with an associated *arity* which is a positive integer. By a *continuous structure*, or just *structure*, say $\mathcal{A}$, we mean a nonempty set A called the *domain* of $\mathcal{A}$ combined with an *interpretation* of every $R \in \sigma$ which, if R has arity k, is a function from A^k to the set of real numbers $\mathbb{R}$; this function, the interpretation of R in $\mathcal{A}$, is denoted by $R^{\mathcal{A}}$. In this study the range of $R^{\mathcal{A}}$ will always be included in $[0, \mathfrak{m}]$ for some fixed, but arbitrary, $\mathfrak{m} \geq 1$.

The following is a direct consequence of [3, Corollary A.1.14] which in turn follows from a bound given by Chernoff [7]:

Lemma 1.1. *Let Z be the sum of n independent 0/1-valued random variables, each one with probability p of having the value 1, where $p > 0$. For every $\varepsilon > 0$ there is $c_\varepsilon > 0$, depending only on ε, such that the probability that $|Z - pn| > \varepsilon pn$ is less than $2e^{-c_\varepsilon pn}$. (If $p = 0$ then the same statement holds if '$2e^{-c_\varepsilon pn}$' is replaced by 'e^{-n}'.)*

2. Aggregation functions

In order to define activation functions for neurons in a way that does not depend on the number of neurons of an NN we will use the concept of aggregation function. We fix some real $\mathfrak{m} \geq 1$.

Definition 2.1. (a) Let $[0, \mathfrak{m}]^{fin}$ denote the set of all finite nonempty sequences of reals from the interval $[0, \mathfrak{m}]$, or equivalently, $[0, \mathfrak{m}]^{fin} = \bigcup_{k=1}^{\infty} [0, \mathfrak{m}]^k$.
(b) A function $F : [0, \mathfrak{m}]^{fin} \to [0, \mathfrak{m}]$ will be called an *aggregation function* if it is symmetric in the sense that if $(x_1, \dots, x_k) \in [0, \mathfrak{m}]^k$ and $(y_1, \dots, y_k)$ is a permutation (reordering) of $(x_1, \dots, x_k)$ then $F(x_1, \dots, x_k) = F(y_1, \dots, y_k)$.

Examples of aggregation functions include the average of a sequence,

$$\mathrm{av}(x_1, \ldots, x_n) := \frac{x_1 + \ldots + x_n}{n}$$

and the maximum of a sequence,

$$\max(x_1, \ldots, x_n) := \text{ the largest number among } x_1, \ldots, x_n.$$

We will require that aggregation functions used by NNs are continuous in the sense defined below. The intuition behind the definition is that if the entries of two sequences (not necessarily of the same length) have similar distribution (if one forgets about the ordering), then a continuous aggregation function should give roughly the same output for both sequences. The definition below is, in theory at least, slightly stronger than the definition of continuous aggregation function in [19] since conditions (c) and (d) in [19, Definition 2.4] are replaced below by a more general condition, labelled (c).

Definition 2.2. Let $F : [0, \mathfrak{m}]^{fin} \to [0, \mathfrak{m}]$ be an aggregation function. We call F *continuous* if the following two conditions hold:

(1) For every $\varepsilon > 0$ there is $\delta > 0$ such that for all n, if $(q_1, \ldots, q_n)$, $(q'_1, \ldots, q'_n) \in [0, \mathfrak{m}]^n$ and $|q_i - q'_i| \leq \delta$ for all $i = 1, \ldots, n$, then $|F(q_1, \ldots, q_n) - F(q'_1, \ldots, q'_n)| \leq \varepsilon$.

(2) For every $\varepsilon > 0$ there are $\delta > 0$ and $M, N \in \mathbb{N}^+$ such that if $\alpha_0, \ldots, \alpha_{M-1} \in [0, 1]$ and $(q_1, \ldots, q_n), (q'_1, \ldots, q'_m) \in [0, 1]^{<\omega}$ (where we may have $n \neq m$) are such that conditions (a)–(d) below hold, then $|F(q_1, \ldots, q_n) - F(q'_1, \ldots, q'_m)| \leq \varepsilon$:
   (a) $n, m \geq N$,
   (b) for all $i = 0, \ldots, M-1$, if $\alpha_i > 0$ then $\alpha_i > \delta$, and (regardless of whether $\alpha_i > 0$ or not)

$$\frac{\left|\left\{l : q_l \in \left[\frac{i}{M}, \frac{i+1}{M}\right]\right\}\right|}{n} \in (\alpha_i - \delta, \alpha_i + \delta),$$

$$\frac{\left|\left\{l : q'_l \in \left[\frac{i}{M}, \frac{i+1}{M}\right]\right\}\right|}{m} \in (\alpha_i - \delta, \alpha_i + \delta), \quad \text{and}$$

   (c) if $0 \leq i < M-2$ and $\alpha_i = \alpha_{i+1} = \alpha_{i+2} = 0$, then

$$\left\{l : q_l \in \left[\tfrac{i+1}{M}, \tfrac{i+2}{M}\right]\right\} = \left\{l : q'_l \in \left[\tfrac{i+1}{M}, \tfrac{i+2}{M}\right]\right\} = \emptyset.$$

Lemma 2.3. *The aggregation functions maximum, minimum and average are continuous.*

Proof. For the average this is proved in [19, Lemma 2.5] since only conditions (a) and (b) of part (2) are used and conditions (a) and (b) above are indentical with (a) and (b) in [19, Definition 2.4].

For maximum and minimum the proof of the corresponding result in [19, Lemma 2.5] can easily be modified to yield continuity of maximum and minimum with Definition 2.2 of continuity. □

Lemma 2.4. *If $F : [0, \mathfrak{m}]^{fin} \to [0, \mathfrak{m}]$ is a continuous aggregation function and $h : [0, \mathfrak{m}] \to [0, \mathfrak{m}]$ is continuous, then $G : [0, \mathfrak{m}]^{fin} \to [0, \mathfrak{m}]$ defined by $G(q_1 \ldots, q_n) := h(F(q_1, \ldots, q_n))$ is a continuous aggregation function.*

3. Model theoretic representation

We fix some $\rho \geq 2$ and consider a sequence of neural networks (NNs), denoted $\mathcal{B}_n$ ('$\mathcal{B}$' for "base structure") for $n \in \mathbb{N}^+$, such that each $\mathcal{B}_n$ has exactly ρ layers and the first layer of $\mathcal{B}_n$ (the layer of input neurons) has exactly n neurons. For each n and $i = 1, \ldots, \rho$, the i^{th} layer of $\mathcal{B}_n$ is denoted by $L_{n,i}$. We also assume that the last layer, $L_{n,\rho}$, is a singleton. We let $L_{n,\rho} := \{o_n\}$ and call o_n the *output neuron* of $\mathcal{B}_n$.

Fix some real number $\mathfrak{m} \geq 1$, where we think of $\mathfrak{m}$ as the maximal value that neurons or connections betwen them can have. For all n, $i = 1, \ldots, \rho - 1$ and all choices of $a \in L_{n,i}$ and $b \in L_{n,i+1}$ there is a *connection* from a to b which carries a *weight* which is a real number in $[0, \mathfrak{m}]$. Once values are assigned to all input neurons of $\mathcal{B}_n$ (the neurons in $L_{n,1}$) all other neurons should automatically be assigned values that depend only on the values of the input neurons, the weight of the connections, and some activation function. The collection of values of all neurons obtained in this way will be thought of as a *state* of $\mathcal{B}_n$, and we are interested in the set of all states that $\mathcal{B}_n$ can be in. To make these ideas mathematically precise we formulate them in a model theoretic language. Since values of neurons and weights of connections are real numbers we will consider *continuous structures* in which relation symbols are interpreted as functions taking values in $[0, \mathfrak{m}]$. So if σ is a signature (i.e. set of relation symbols), $\mathcal{A}$ is a continuous σ-structure, and $R \in \sigma$ has arity k, then the *interpretation of R in $\mathcal{A}$*, denoted $R^{\mathcal{A}}$, is a function from A^k (where A is the domain of $\mathcal{A}$) to $\mathbb{R}$. Since we have, in addition, stipulated that all values must be in $[0, \mathfrak{m}]$ we also require $R^{\mathcal{A}}$ to be a mapping into $[0, \mathfrak{m}]$.

Let $\tau = \{L_1, \ldots, L_\rho, E\}$ where $L_1, \ldots, L_\rho$ have arity ρ and E has arity 2. For each n, $\mathcal{B}_n$ is a τ-structure with domain B_n such that

(A1) For all $b \in B_n$, there is exactly one $i \in [\rho]$ such that $L_i^{\mathcal{B}_n}(b) = 1$, and if $j \neq i$ then $L_j^{\mathcal{B}_n}(b) = 0$. This expresses that each neuron belongs to exactly one layer, and we define
$$L_{n,i} := \{b \in B_n : L_i^{\mathcal{B}_n}(b) = 1\}.$$

(A2) If $i, j \in [\rho]$, $j \leq i$ or $j - i > 1$, then for all $a \in L_{n,i}$ and $b \in L_{n,j}$ we have $E^{\mathcal{B}_n}(a, b) = 0$. This expresses that a synapse with positive weight must go from some layer $L_{n,i}$ to the next layer $L_{n,i+1}$.

We will assume that all $\mathcal{B}_n$, $n \in \mathbb{N}^+$, have "*similar design*" which besides having the same number of layers will mean that they use the same activation functions and that the weights of connections between two fixed consecutive layers have similar distributions for all sufficiently large n. First we fix some functions that will specify a "similarity class" of sequences of NNs.

(A3) Let $\zeta(x)$ be a polynomial such that $\zeta(x) > 0$ for $x > 0$ and $\lim_{x\to\infty} \zeta(x) = \infty$.

(A4) Let $W_2, \ldots, W_\rho : [0, \mathfrak{m}] \to [0, \infty)$ and $g_2, \ldots, g_\rho : [0, \mathfrak{m}]^2 \to [0, \mathfrak{m}]$ be continuous functions such that $\int_{[0,\mathfrak{m}]} W_i(x)dx = 1$ for all $i = 2, \ldots, \rho$.

(A5) Let $F_2, \ldots, F_\rho : [0, \mathfrak{m}]^{fin} \to [0, \mathfrak{m}]$ be continuous aggregation functions in the sense of Defintion 2.2.

Next, we assume that the following conditions hold:

(A6) For all $i = 1, \ldots, \rho$, $|L_{n,i}| \leq \zeta(n)$, so layer i of $\mathcal{B}_n$ has at most $\zeta(n)$ neurons.

(A7) For all $i = 2, \ldots, \rho$, for every interval $I \subseteq [0, \mathfrak{m}]$ with different endpoints, and for every $\varepsilon > 0$, if n is sufficiently large and $b \in L_{n,i}$, then
$$\int_I W_i(x)dx - \varepsilon \leq \frac{\left|\left\{a \in L_{n,i-1} : E^{\mathcal{B}_n}(a, b) \in I\right\}\right|}{\left|L_{n,i}\right|} \leq \int_I W_i(x)dx + \varepsilon,$$
and if $\int_I W_i(x)dx = 0$, then the proportion above is 0. Informally, this means that the distribution of weights of connections from $L_{n,i-1}$ to $b \in L_{n,i}$ is approximated well by W_i for large n.

(A8) For $i = 2, \ldots, \rho$ and any $b \in L_{n,i}$, if $t(n, i-1)$ denotes the number of neurons in layer $i-1$, v_j $(j = 1, \ldots, t(n, i))$ is the value of the j^{th} neuron in $L_{n,i-1}$ (in some arbitrary order) and w_j is the weight of the connection from the j^{th} neuron to b, then the value of b is assigned to be
$$F_i\big(g_i(v_1, w_1), \ldots, g_i(v_{t(n,i-1)}, w_{t(n,i-1)})\big).$$

Hence the above expression defines the activation function for neurons in layer i of $\mathcal{B}_n$ for all n.

Remark 3.1. Fix any $i \in \{2, \dots, \rho\}$. Since $\int_{[0,\mathfrak{m}]} W_i(x)dx = 1$ and W_i is continuous on $[0, \mathfrak{m}]$, hence uniformly continuous on the same interval, there are $\alpha, \beta \in\subseteq [0, \mathfrak{m}]$ such that $\beta - \alpha > 0$ and $W_i(x) > 0$ for all $x \in [\alpha, \beta]$. Let $M \in \mathbb{N}^+$. It follows that if, for all $i = 0, \dots, M-1$, we let $I_i := \big[\alpha + \frac{i(\beta-\alpha)}{M},\ \alpha + \frac{(i+1)(\beta-\alpha)}{M}\big]$, then $\int_{I_i} W_i(x)dx > 0$. It now follows from Assumption (A7) that if n is sufficiently large, then $|L_{n,i-1}| \geq M$. Thus, as $n \to \infty$, then number of neurons in layer $i-1$ tends to infinity. So the number of neurons in all layers except for the output layer tend to infinity as $n \to \infty$.

We assumed that the input layer $L_{n,1}$ of $\mathcal{B}_n$ has exactly n neurons. Without loss of generality we may (to simplify notation) assume that $L_{n,1} = [n] := \{1, \dots, n\}$.

We now formally define the notion of state of $\mathcal{B}_n$, the state space of $\mathcal{B}_n$, and the probability measure on the state space. Let $\sigma := \tau \cup \{P\}$ where P has arity 1. The value of a neuron a of $\mathcal{B}_n$ in a particular state will be represented by the value of $P(a)$ in an expansion to σ of $\mathcal{B}_n$. More precisely, let $\mathbf{S}_n$ consist of all σ-structures $\mathcal{A}$ such that $\mathcal{A}$ is an expansion of $\mathcal{B}_n$ and

(A9) For all $i = 2, \dots, \rho$ and all $b \in L_{n,i}$, if $a_1, \dots, a_{t(n,i-1)}$ is a list without repetitions of all neurons in $L_{n,i-1}$, in any order, then

$$P^{\mathcal{A}}(b) = F_i\big(g_i\big(P^{\mathcal{A}}(a_1), w_1\big), \dots, g_i\big(P^{\mathcal{A}}(a_{t(n,i-1)}), w_{t(n,i-1)}\big)\big).$$

For each $(x_1, \dots, x_n) \in [0, \mathfrak{m}]^n$ let $\mathfrak{S}_n(x_1, \dots, x_n)$ be the unique $\mathcal{A} \in \mathbf{S}_n$ such that $x_k = P^{\mathcal{A}}(k)$ for all $k \in L_{n,1} := [n]$. By (A9), $\mathfrak{S}_n$ is a bijection.

Let $\mu : [0, \mathfrak{m}] \to [0, \infty)$ be a probability density function. Define $\mu_n : [0, \mathfrak{m}]^n \to [0, \infty)$ by $\mu_n(x_1, \dots, x_n) = \mu(x_1) \cdot \ldots \cdot \mu(x_n)$, so μ_n is a probability density function on $[0, \mathfrak{m}]^n$. Let $\mathbb{P}_n$ be the probability measure on $[0, \mathfrak{m}]^n$ which is induced by μ_n in the sense that if $X \subseteq [0, \mathfrak{m}]^n$ is (Lebesque) measurable then

$$\mathbb{P}_n(X) = \int_X \mu_n(x_1, \dots, x_n)dx_1 \dots dx_n.$$

Call $\mathbf{X} \subseteq \mathbf{S}_n$ *measurable* if the inverse image $\mathfrak{S}^{-1}(\mathbf{X})$ is a measurable subset of $[0, \mathfrak{m}]^n$. By mild abuse of notation we also view $\mathbb{P}_n$ as a probability measure on $\mathbf{S}_n$ via the bijection $\mathfrak{S}_n$ as follows: for measurable $\mathbf{X} \subseteq \mathbf{S}_n$,

$$\mathbb{P}_n(\mathbf{X}) := \mathbb{P}_n(\mathfrak{S}_n^{-1}(\mathbf{X})) := \int_{\mathfrak{S}_n^{-1}(\mathbf{X})} \mu_n(x_1, \dots, x_n)dx_1 \dots dx_n.$$

For all $(x_1, \dots, x_n) \in [0, \mathfrak{m}]^n$ let $Out_n(x_1, \dots, x_n)$ be the value of the output neuron of $\mathcal{B}_n$ in the state $\mathfrak{S}_n(x_1, \dots, x_n)$, or formally,

$$Out_n(x_1, \dots, x_n) := P^{\mathfrak{S}_n(x_1,\dots,x_n)}(o_n).$$

Lemma 3.2. *For all n the function $Out_n : [0, \mathfrak{m}]^n \to [0, \mathfrak{m}]$ is continuous. It follows that, for every interval $I \subseteq [0, \mathfrak{m}]$, $Out_n^{-1}(I)$ is Lebesque measurable.*

Proof. Fix any $n \in \mathbb{N}^+$. For any fixed $w \in [0, \mathfrak{m}]$ and $i = 2, \dots, \rho$, let $g_{i,w}(x) = g_i(x, w)$ so $g_{i,w} : [0, \mathfrak{m}] \to [0, \mathfrak{m}]$ is continuous. We prove the claim by induction on ρ. The base case is when $\rho = 2$. If $w_1, \dots, w_n$ enumerate the weights of connections from $L_{n,1} = [n]$ to the unique neuron o_n in $L_{2,2}$ then, by (A9),

$$P^{\mathfrak{S}_n(x_1,\dots,x_n)}(o_n) = F_2(g_{2,w_1}(x_1), \dots, g_{2,w_n}(x_n))$$

where the right hand side is a composition of continuous functions in the usual sense of mathematical analysis, since the aggregation function F_2 restricted to $[0, \mathfrak{m}]^n$ is continuous in this usual sense because of part (1) of Definition 2.2. Hence Out_n is continuous if

$\rho = 2$. By the same argument it follows that if $\rho > 2$, $b \in L_{n,2}$, and $w_1, \ldots, w_n$ enumerate the weights of connections from $L_{n,1} = [n]$ to b then

$$P^{\mathfrak{S}_n(x_1,\ldots,x_n)}(b) = F_2(g_{2,w_1}(x_1), \ldots, g_{2,w_n}(x_n))$$

Then $H_2^b(x_1, \ldots, x_n) := F_2(g_{2,w_1}(x_1), \ldots, g_{2,w_n}(x_n))$ is continuous. where the right hand side is a continuous function from Now suppose that $2 \leq i < \rho$ and that, for all $b \in L_{n,i}$, there is continuous $H_i^b : [0, \mathfrak{m}]^n \to [0, \mathfrak{m}]$ such that

$$P^{\mathfrak{S}_n(x_1,\ldots,x_n)}(b) = H_i^b(x_1, \ldots, x_n).$$

Note that we showed that this holds if $i = 2$. Let $b \in L_{n,i+1}$, let $a_1, \ldots, a_{t(n,i)}$ be an enumeration of $L_{n,i}$, and let w_j be the weight of the connection from a_j to b. By (A9) and the assumption we have

$$P^{\mathfrak{S}_n(x_1,\ldots,x_n)}(b) =$$
$$F_{i+1}\big(g_{i+1,w_1}(H_i^{a_1}(x_1, \ldots, x_n)), \ldots, g_{i+1,w_{t(n,i)}}(H_i^{a_{t(n,i)}}(x_1, \ldots, x_n))\big)$$

where the right hand side is (using part (1) of Definition 2.2) a composition of continuous functions, so the function $H_{i+1}^b(x_1, \ldots, x_n) := P^{\mathfrak{S}_n(x_1,\ldots,x_n)}(b)$ is continuous. This completes the inductive step. □

Theorem 3.3. *Suppose that $\mathcal{B}_n$, $n \in \mathbb{N}^+$, is a sequence of NNs subject to the assumptions stated above ((A1) – (A8)) and that the set of states $\mathbf{S}_n$ and the probability distribution $\mathbb{P}_n$ are as defined above. Then there is $\psi \in [0, \mathfrak{m}]$ such that for all $\varepsilon > 0$ there is $c > 0$ such that for all sufficiently large n,*

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(o_n) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-cn}.$$

Briefly formulated, the next theorem states that under the above stated assumptions *only two layers suffice* to get, up to an arbitrarily small error, the same distribution of the output neuron for large n.

Corollary 3.4. *Suppose $\mathcal{B}_n$, $n \in \mathbb{N}^+$, is a sequence of NNs subject to the assumptions stated above and that $\mathbb{P}_n$ is as stated above. Let ψ be as in Theorem 3.3. Then there is a sequence $\mathcal{B}'_n$, $n \in \mathbb{N}^+$, of NNs such that the following hold:*

- *Each $\mathcal{B}'_n$ has exactly 2 layers and n neurons in the first layer.*
- *All assumptions above hold for $\mathcal{B}'_n$ in place of $\mathcal{B}_n$ with the same W_2 and g_2 but with a modified continuous aggregation function F'_2 in place of F_2 which $\mathcal{B}_n$ uses.*
- *If $\mathbb{P}'_n$ is the probability distribution induced by μ_n on the state space $\mathbf{S}'_n$ of $\mathcal{B}'_n$ and o'_n denotes the output neuron of $\mathcal{B}'_n$, then for all $\varepsilon > 0$ there is $c > 0$ such that for all sufficiently large n,*

$$\mathbb{P}'_n\big(\{\mathcal{A} \in \mathbf{S}'_n : P^{\mathcal{A}}(o'_n) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-cn}.$$

4. Proofs

We prove Theorem 3.3 by induction on the number ρ of layers, where $\rho \geq 2$ (by assumption).

4.1. The base case: 2 layers.

In this section we prove the following lemma which gives Theorem 3.3 in the case when $\rho = 2$:

Lemma 4.1. *Suppose that $\rho = 2$, so each $\mathcal{B}_n$ has two layers, $L_{n,1}$ and $L_{n,2} = \{o_n\}$. There is $\psi \in [0, \mathfrak{m}]$ such that for all $\varepsilon > 0$ there is $c > 0$ such that for all sufficiently large n,*

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(o_n) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-cn}.$$

Moreover, ψ depends only on W_2, F_2, and g_2.

$$I_{\Delta,i} := \Big[\frac{i\mathfrak{m}}{\Delta},\ \frac{(i+1)\mathfrak{m}}{\Delta}\Big].$$

Lemma 4.2. *Let $n, \Delta \in \mathbb{N}^+$, $i \in \{0, \dots, \Delta - 1\}$, and $k \in [n]$. Then*

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(k) \in I_{\Delta,i}\}\big) =$$
$$\mathbb{P}_n\big(\{(x_1, \dots, x_n) \in [0,\mathfrak{m}]^n : x_k \in I_{\Delta,i}\}\big) = \int_{I_{\Delta,i}} \mu(x)dx.$$

Hence the probability does not depend on n or on k, but only on Δ and i (and μ which we have fixed).

Proof. From the definitions it directly follows that $x_k \in I_{\Delta,i}$ if and only if $P^{\mathfrak{S}_n(x_1,\dots,x_n)}(k) \in I_{\Delta,i}$. Since for all $X \subseteq [0,\mathfrak{m}]^n$, $\mathbb{P}_n(X) = \mathbb{P}_n(\mathfrak{S}_n(X))$ it suffices to prove the second equality of the lemma.

By the definition of $\mathbb{P}_n$ and since $\mu_n(x_1, \dots, x_n) = \mu(x_1) \cdot \dots \cdot \mu(x_n)$ where $\mu : [0,\mathfrak{m}] \to [0,\infty)$ is a probability density function, we get, by using Fubini's theorem,

$$\mathbb{P}_n\big(\{(x_1, \dots, x_n) \in [0,\mathfrak{m}]^n : x_k \in I_{\Delta,i}\}\big) =$$
$$\int_{[0,\mathfrak{m}]^{k-1} \times I_{\Delta,i} \times [0,\mathfrak{m}]^{n-k}} \mu_n(x_1, \dots, x_n)dx_1 \dots dx_n =$$
$$\int_{[0,\mathfrak{m}]^{k-1} \times I_{\Delta,i} \times [0,\mathfrak{m}]^{n-k}} \mu(x_1) \cdot \dots \cdot \mu(x_n)dx_1 \dots dx_n$$
$$\Big(\prod_{i=1}^{k-1} \int_{[0,\mathfrak{m}]} \mu(x_i)dx_i\Big) \cdot \int_{I_{\Delta,i}} \mu(x_k)dx_k \cdot \Big(\prod_{i=k+1}^{n} \int_{[0,\mathfrak{m}]} \mu(x_i)dx_i\Big) =$$
$$\int_{I_{\Delta,i}} \mu(x_k)dx_k.$$

□

Lemma 4.3. *Let $n \in \mathbb{N}^+$, $k \in [n]$, and let $I_k \subseteq [0,\mathfrak{m}]$ be an interval. The event $\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(k) \in I\}$ is independent from from all events $\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(l) \in I_i\}$ where $i \neq k$ and $I_i \subseteq [0,\mathfrak{m}]$ is an interval.*

Proof. By the definitions,

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(i) \in I_i \text{ for all } i \in [n]\}\big) =$$
$$\mathbb{P}_n\big(\{(x_1, \dots, x_n) \in [0,\mathfrak{m}]^n : x_i \in I_i \text{ for all } i \in [n]\}\big).$$

Therefore it suffices to prove that the event $\{(x_1, \dots, x_n) \in [0,\mathfrak{m}]^n : x_k \in I_k\}$ is independent from from all events $\{(x_1, \dots, x_n) \in [0,\mathfrak{m}]^n : x_i \in I_i\}$ where $i \neq k$ and $I_i \subseteq [0,\mathfrak{m}]$ is an interval. But this follows from by using Fubini's therem since $\mu_n(x_1, \dots, x_n) = \mu(x_1) \cdot \dots \cdot \mu(x_k)$. □

Let

$$\beta_{\Delta,i} := \int_{I_{\Delta,i}} \mu(x)dx \qquad \text{and} \qquad \gamma_{\Delta,i} := \int_{I_{\Delta,i}} W_2(x)dx.$$

Define

$$B_n(\Delta, i) := \big\{k \in [n] : E^{\mathcal{B}_n}(k, o_n) \in I_{\Delta,i}\big\}. \tag{4.1}$$

It follows from (A7) that for all $\varepsilon > 0$, if n is large enough then

$$\gamma_{\Delta,i} - \varepsilon \leq \frac{|B_n(\Delta, i)|}{n} \leq \gamma_{\Delta,i} + \varepsilon. \tag{4.2}$$

So if $\gamma_{\Delta,i} > 0$, then $|B_n(\Delta, i)| \to \infty$ as $n \to \infty$ and the rate och growth is linear in n. Also, by (A7), if $\gamma_{\Delta,i} = 0$, then $B_n(\Delta, i) = \emptyset$.

For all $\mathcal{A} \in \mathbf{S}_n$, $\Delta \in \mathbb{N}^+$ and $i, j \in \{0, \dots, \Delta - 1\}$, let

$$C_n(\mathcal{A}, \Delta, i, j) := \big\{k \in B_n(\Delta, i) : P^{\mathcal{A}}(k) \in I_{\Delta,j}\big\}. \tag{4.3}$$

For $\Delta \in \mathbb{N}^+$ and $\varepsilon > 0$ define

$$\begin{aligned} &\mathbf{X}_n(\Delta, \varepsilon) := \\ &\big\{\mathcal{A} \in \mathbf{S}_n : \text{ for all } i, j \in \{0, \dots, \Delta - 1\} \text{ such that } \gamma_{\Delta,i} > 0, \\ &(\beta_{\Delta,j} - \varepsilon)|B_n(\Delta, i)| \le \big|C_n(\mathcal{A}, \Delta, i, j)\big| \le (\beta_{\Delta,j} + \varepsilon)|B_n(\Delta, i)|\big\}. \end{aligned} \tag{4.4}$$

Lemma 4.4. *For all $n, \Delta \in \mathbb{N}^+$ and $\varepsilon > 0$, $\mathbf{X}_n(\Delta, \varepsilon)$ is measurable. If n is sufficiently large, then $\mathbb{P}_n\big(\mathbf{X}_n(\Delta, \varepsilon)\big) \ge 1 - e^{-cn}$ for some constant $c > 0$ that depends only on ε.*

Proof. Fix any $i, j \in \Delta$. Define

$$\begin{aligned} &\mathbf{X}_n(\Delta, \varepsilon, i, j) := \\ &\big\{\mathcal{A} \in \mathbf{S}_n : (\beta_{\Delta,j} - \varepsilon)|B_n(\Delta, i)| \le \big|C_n(\mathcal{A}, \Delta, i, j)\big| \le (\beta_{\Delta,j} + \varepsilon)|B_n(\Delta, i)|\big\}. \end{aligned}$$

From the definition of measurable subset of $\mathbf{S}_n$ it follows that every subset of $\mathbf{S}_n$ of the form

$$\mathbf{X}_n(k) := \big\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(k) \in I_{\Delta,i}\big\} \text{ for some } k \in [n]$$

is measurable. Since $\mathbf{X}_n(\Delta, \varepsilon, i, j)$ can be formed from sets of the form $\mathbf{X}_n(k)$ by taking complements, intersections, and unions, a finite number of times, it follows that $\mathbf{X}_n(\Delta, \varepsilon, i, j)$ is measurable. As $\mathbf{X}_n(\Delta, \varepsilon)$ is a finite union of sets of the form $\mathbf{X}_n(\Delta, \varepsilon, i, j)$ it follows that also $\mathbf{X}_n(\Delta, \varepsilon)$ is measurable.

From Lemma 4.2 it follows that, for all $k \in [n]$,

$$\mathbb{P}_n\big(\big\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(k) \in I_{\Delta,i}\big\}\big) = \beta_{\Delta,j} := \int_{I_{\Delta,i}} \mu(x)dx. \tag{4.5}$$

First suppose that such that $\gamma_{\Delta,i} > 0$. Then, as noted above, $\lim_{n\to\infty} |B_n(\Delta, i)| = \infty$ so it follows from (4.5), Lemma 4.3, and Lemma 1.1 that there is $d > 0$ such that if n is large enough, then

$$\mathbb{P}_n\big(\mathbf{X}_n(\Delta, \varepsilon, i, j)\big) \ge 1 - e^{-dn}. \tag{4.6}$$

Now suppose that $\gamma_{\Delta,i} = 0$. Then, as noted above, $B_n(\Delta, i) = \emptyset$ and as, for all $\mathcal{A} \in \mathbf{S}_n$, $C_n(\mathcal{A}, \Delta, i, j) \subseteq B_n(\Delta, i)$ we get $C_n(\mathcal{A}, \Delta, i, j) = \emptyset$. Hence $\mathbf{X}_n(\Delta, \varepsilon, i, j) = \mathbf{S}_n$ so (4.6) holds for any choice of $d > 0$. Since

$$\mathbf{X}_n(\Delta, \varepsilon) = \bigcup_{0 \le i,j \le \Delta - 1} \mathbf{X}_n(\Delta, \varepsilon, i, j)$$

and there are only Δ^2 choices of i and j it follows that there is $c > 0$ such that if n is sufficiently large, then

$$\mathbb{P}_n\big(\mathbf{X}_n(\Delta, \varepsilon)\big) \ge 1 - e^{-cn}.$$

□

Note that

$$L_{n,1} = \bigcup_{i=0}^{\Delta-1} B_n(\Delta, i)$$

and observe that for every $\mathcal{A} \in \mathbf{S}_n$ we have

$$B_n(\Delta, i) = \bigcup_{j=0}^{\Delta-1} C_n(\mathcal{A}, \Delta, i, j) \quad \text{and} \quad L_{n,1} = \bigcup_{i=0}^{\Delta-1} \bigcup_{j=0}^{\Delta-1} C_n(\mathcal{A}, \Delta, i, j).$$

For $\mathcal{A} \in \mathbf{X}_n(\Delta, \delta)$ we have

$$\beta_{\Delta,j} - \delta \leq \frac{|C_n(\mathcal{A}, \Delta, i, j)|}{|B_n(\Delta, i)|} \leq \beta_{\Delta,j} + \delta. \tag{4.7}$$

Note that if $\beta_{\Delta,j}, \gamma_{\Delta,i} > 0$ then

$$\frac{|C_n(\mathcal{A}, \Delta, i, j)|}{n} = \frac{|B_n(\Delta, i)|}{n} \cdot \frac{|C_n(\mathcal{A}, \Delta, i, j)|}{|B_n(\Delta, i)|} \leq \gamma_{\Delta,i}\beta_{\Delta,j} + 3\delta.$$

In a similar way we get $\frac{|C_n(\mathcal{A},\Delta,i,j)|}{n} \geq \gamma_{\Delta,i}\beta_{\Delta,j} - 3\delta$. Thus

$$\beta_{\Delta,j}, \gamma_{\Delta,i} > 0 \implies \gamma_{\Delta,i}\beta_{\Delta,j} - 3\delta \leq \frac{|C_n(\mathcal{A}, \Delta, i, j)|}{n} \leq \gamma_{\Delta,i}\beta_{\Delta,j} + 3\delta.$$

On the other hand,

$$\text{if } \beta_{\Delta,j} = 0 \text{ or } \gamma_{\Delta,i} = 0 \quad \text{then} \quad \frac{|C_n(\mathcal{A}, \Delta, i, j)|}{n} = 0.$$

Fix M and $l \in \{0, \ldots, M-1\}$. For $\mathcal{A} \in \mathbf{X}_n(\Delta, \delta)$ and $k \in [n]$ let

$$q_k(\mathcal{A}) := g_2(P^{\mathcal{A}}(k), E^{\mathcal{B}_n}(k, o_n)) \tag{4.8}$$

and

$$\omega(\mathcal{A}, M, l) := \frac{\big|\big\{k \in [n] : q_k(\mathcal{A}) \in I_{M,l}\big\}\big|}{n}. \tag{4.9}$$

Define

$$\Gamma_1 := \big\{(i,j) \in \{0, \ldots, \Delta-1\} : \beta_{\Delta,j}, \gamma_{\Delta,i} > 0 \text{ and } g_2(I_{\Delta,j} \times I_{\Delta,i}) \subseteq I_{M,l}\big\} \quad \text{and}$$
$$\Gamma_2 := \big\{(i,j) \in \{0, \ldots, \Delta-1\} : \beta_{\Delta,j} = 0 \text{ or } \gamma_{\Delta,i} = 0, \text{ and}$$
$$g_2(I_{\Delta,j} \times I_{\Delta,i}) \cap I_{M,l} \neq \emptyset \text{ and } g_2(I_{\Delta,j} \times I_{\Delta,i}) \not\subseteq I_{M,l}\big\}.$$

Suppose that $\mathcal{A} \in \mathbf{X}_n(\Delta, \delta)$. Then we have

$$\begin{aligned} \omega(\mathcal{A}, M, l) &\geq \sum_{(i,j)\in\Gamma_1} \frac{|C_n(\mathcal{A}, \Delta, i, j)|}{n} \geq \sum_{(i,j)\in\Gamma_1} (\gamma_{\Delta,i}\beta_{\Delta,j} - 3\delta) \\ &\geq \sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j}. \end{aligned} \tag{4.10}$$

Note that it follows from the definitions that if $(i,j) \in \{0, \ldots, \Delta-1\}^2$ and $(i,j) \notin \Gamma_1 \cup \Gamma_2$, then $\frac{|C_n(\mathcal{A},\Delta,i,j)|}{n} = 0$. So we also have

$$\begin{aligned} \omega(\mathcal{A}, M, l) &\leq \sum_{(i,j)\in\Gamma_1} \frac{|C_n(\mathcal{A}, \Delta, i, j)|}{n} + \sum_{(i,j)\in\Gamma_2} \frac{|C_n(\mathcal{A}, \Delta, i, j)|}{n} \\ &\quad + \sum_{(i,j)\in\{0,\ldots,\Delta-1\}^2\setminus(\Gamma_1\cup\Gamma_2)} \frac{|C_n(\mathcal{A}, \Delta, i, j)|}{n} \\ &\leq \sum_{(i,j)\in\Gamma_1} (\gamma_{\Delta,i}\beta_{\Delta,j} + 3\delta) + \sum_{(i,j)\in\Gamma_2} (\gamma_{\Delta,i}\beta_{\Delta,j} + 3\delta) \\ &= \sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j} + \sum_{(i,j)\in\Gamma_2} \gamma_{\Delta,i}\beta_{\Delta,j} + 3\delta|\Gamma_1| + 3\delta|\Gamma_2| \\ &\leq \sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j} + \sum_{(i,j)\in\Gamma_2} \gamma_{\Delta,i}\beta_{\Delta,j} + 6\delta\Delta^2. \end{aligned} \tag{4.11}$$

Let

$$\Omega_0 := \big\{(x,y) \in [0,\mathfrak{m}]^2 : \mu(x) > 0 \text{ and } W(x) > 0\big\}, \quad \text{and}$$
$$\text{let } \Omega \text{ be the closure of } \Omega_0.$$

Then Ω is a finite union of closed rectangles. Now let

$$X_l := g_2^{-1}(I_{M,l}) \cap \Omega.$$

By λ^2 we denote the Lebesgue measure on $\mathbb{R}^2$ restricted to $[0,\mathfrak{m}]^2$.

Case 1: $\lambda^2(X_l) > 0$.
As g_2 is continuous on $[0,\mathfrak{m}]^2$ it follows that X_l is a compact subset of $[0,\mathfrak{m}]^2$. Let Y_l be the boundary of X_l relative to $[0,\mathfrak{m}]^2$. That is, Y_l contains all $(x,y) \in [0,\mathfrak{m}]^2$ such that for every $\delta' > 0$, the ball with radius δ' centered at (x,y) contains a point in X_l and a point not in X_l. Then Y_l is a finite union of curves, so $\lambda^2(Y_l) = 0$. Note that $(i,j) \in \Gamma_2$ if and only if $I_{\Delta,j} \times I_{\Delta,i} \cap Y_l \neq \emptyset$. Let

$$Z := \bigcup_{(i,j)\in\Gamma_2} I_{\Delta,j} \times I_{\Delta,i}.$$

Since each $I_{\Delta,j} \times I_{\Delta,i}$ is a square with side $\mathfrak{m}/\Delta$ it follows that we can make $\lambda^2(Z)$ as small as we like by choosing Δ sufficiently large. Let

$$V := \bigcup_{(i,j)\in\Gamma_1} I_{\Delta,j} \times I_{\Delta,i}$$

Then, for any $\delta' > 0$, if Δ is sufficiently large then

$$\lambda^2(V) \geq \lambda^2(X_l) - \delta'.$$

As we assume that $\lambda^2(X_l) > 0$ and since $\lambda^2(X_l)$ does not depend on Δ or n it follows that if Δ is sufficiently large then $\lambda^2(V)$ is larger than some positive contant that does not depend on Δ or n. But as noted above, we can make $\lambda^2(Z)$ as small as we like if Δ is sufficiently large, so the proportion $\lambda^2(Z)/\lambda^2(V_l)$ can be made as small as we like if Δ is large enough. Since both Z and V are unions of squares with side length $\mathfrak{m}/\Delta$ it follows that the proportion $|\Gamma_2|/|\Gamma_1|$ can be made as large as we like if Δ is large enough.

Let

$$\xi_\Delta := \min\big\{\gamma_{\Delta,i}\beta_{\Delta,j} : (i,j) \in \Gamma_1\big\}.$$

By the definition of Γ_1 it follows that $\gamma_{\Delta,i}\beta_{\Delta,j} > 0$ for all $(i,j) \in \Gamma_1$. Hence $\xi_\Delta > 0$. As $|\Gamma_2|/|\Gamma_1|$ can be made as large as we like if Δ is large enough it follows that for every $\delta' > 0$, if Δ is large enough then

$$\sum_{(i,j)\in\Gamma_2} \gamma_{\Delta,i}\beta_{\Delta,j} \;\leq\; \delta' \sum_{(i,j)\in\Gamma_1} \xi_\Delta \;\leq\; \delta' \sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j}.$$

From the definition of $\beta_{\Delta,i}$ and $\gamma_{\Delta,i}$ it follows that $\sum_{i=0}^{\Delta-1} \beta_{\Delta,i} = 1$ and $\sum_{j=0}^{\Delta-1} \gamma_{\Delta,j} = 1$. Hence $\sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j} \leq 1$ and we get

$$\sum_{(i,j)\in\Gamma_2} \gamma_{\Delta,i}\beta_{\Delta,j} \leq \delta'.$$

Substituting this in (4.11) gives

$$\omega(\mathcal{A}, M, l) \leq \sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j} + \delta' + 6\delta\Delta^2.$$

Combining with (4.10) gives

$$\sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j} \leq \omega(\mathcal{A}, M, l) \leq \sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j} + \delta' + 6\delta\Delta^2$$

where $\delta'' := \delta' + 6\delta\Delta^2$ can be made as small as we like by taking $\delta' > 0$ small enough, then Δ large enough, and finally $\delta > 0$ small enough to make $6\delta\Delta^2$ as small as we like. Observe that we have also proved that $\alpha_l := \sum_{(i,j)\in\Gamma_1} \gamma_{\Delta,i}\beta_{\Delta,j}$ is positive if $\lambda^2(X_l) > 0$.

Case 2: $\lambda^2(X_l) = 0$**.** We can argue similarly as in Case 1 with X_l in the role of Y_l and conclude that $\omega(\mathcal{A}, M, l)$ can be made as small as we like if δ' (as in Case 1) is chosen small enough and Δ large enough. Therefore we define $\alpha_l := 0$ and it follows that $\alpha_l - \delta'' \leq \omega(\mathcal{A}, M, l) \leq \alpha_l + \delta''$ with δ'' as above.

We have proved that, for all $l = 0, \ldots, M-1$, and regardless of whether $\lambda^2(X_l)$ is positive or zero, we have

$$\alpha_l - \delta'' \leq \omega(\mathcal{A}, M, l) \leq \alpha_l + \delta''.$$

We assume that the aggregation function F_2 is continuous. Let $\varepsilon > 0$. Then there are $\delta^* > 0$, $M, N \in \mathbb{N}^+$ such that if $\alpha_0, \ldots, \alpha_{M-1} \in [0,1]$ and $(q_1, \ldots, q_n), (q'_1, \ldots, q'_m) \in [0, \mathfrak{m}]^{fin}$ are such that conditions (a)–(c) of Definition 2.2 are satisfied with δ^* in place of δ, then $|F_2(q_1, \ldots, q_n) - F_2(q'_1, \ldots, q'_m)| \leq \varepsilon$. Let $\delta^* > 0$, M and N be such that the conclusion of the previous sentence holds. Then we choose $\delta' > 0$ small enough, Δ large enough and then $\delta > 0$ small enough so that $\delta'' := \delta' + 6\delta\Delta^2$ is smaller than δ^*. For all sufficiently large n, $\mathbf{X}_n(\Delta, \delta)$ is nonempty. Let $\alpha_1, \ldots, \alpha_{M-1}$ be chosen as above. If $\mathcal{A} \in \mathbf{X}_n(\Delta, \delta)$, and $\mathcal{A}' \in \mathbf{X}_m(\Delta, \delta)$ then, with

$$\omega(\mathcal{A}', M, l) := \frac{|k \in [m] : q_k(\mathcal{A}') \in I_{M,l}|}{m},$$

we get, for all sufficiently large n, m,

$$\alpha_l - \delta^* \leq \omega(\mathcal{A}, M, l), \omega(\mathcal{A}', M, l) \leq \alpha_l + \delta^* \quad \text{ for all } l = 0, \ldots, M-1.$$

Without loss of generality we can assume that $n, m \geq N$, so conditions (a) and (b) of Definition 2.2 are satisfied if $q_k := q_k(\mathcal{A})$ and $q'_k := q_k(\mathcal{A}')$. Suppose for a moment that also condition (c) of Definition 2.2 holds for sequences constructed like these. Then $|F_2(q_1, \ldots, q_n) - F_2(q'_1, \ldots, q'_m)| \leq \varepsilon$. Since $P^{\mathcal{A}}(o_n) = F_2(q_1, \ldots, q_n)$ and $P^{\mathcal{A}'}(o_m) = F_2(q'_1, \ldots, q'_m)$ we get $|P^{\mathcal{A}}(o_n) - P^{\mathcal{A}'}(o_m)| \leq \varepsilon$. As $\mathbb{P}_n\big(\mathbf{X}_n(\Delta, \delta)\big) \geq 1 - e^{-cn}$ (where $c > 0$) for all choices of Δ and δ and all sufficiently large n, and since ε can be taken as small as we like, it follows that there is $\psi \in [0, \mathfrak{m}]$ such that, for every open interval I containing ψ,

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(o_n) \in I\}\big) \geq 1 - e^{-cn}.$$

It remains to show that if $\bar{q} := (q_1, \ldots, q_n)$ and $\bar{q}' := (q'_1, \ldots, q'_m)$ where q_k and q'_k are defined as above, then condition (c) of Definition 2.2 holds. Since the verification is the same for $\bar{q}$ and $\bar{q}'$ we only do it for $\bar{q}$.

Verification of condition (c) in Definition 2.2. Suppose that $\alpha_l = \alpha_{l+1} = \alpha_{l+2} = 0$. Suppose for a contradiction that $g_2(\Omega) \cap I_{M,l+1} \neq \emptyset$. Then there is $z \in \Omega$ such that $g_2(z) \in I_{M,l+1}$. Recall that Ω is a finite union of closed rectangles. Let Ω_z denote the connected component to which z belongs. If there would be $z' \in \Omega_z$ such that $g_2(z') \in I_{M,l+1}$ and $g_2(z') \neq g_2(z)$, then, by the continuity of g_2, $g_2^{-1}(I_{M,l+1}) \cap \Omega_z$ would have positive measure, hence also X_l would have have positive measure which contradicts that $\alpha_{l+1} = 0$. Hence $g_2(\Omega_z) \cap I_{M,l+1}$ is a singleton. There is some k such that $g_2^{-1}(I_{M,k}) \cap \Omega_z$ has positive measure, and hence $\alpha_k > 0$ and $g_2(\Omega_z) \cap I_{M,k} \neq \emptyset$. Suppose that $k < l$. (The case $l + 2 < k$ is treated analogously.) Since we assume that $\alpha_l = 0$ we can argue as we did for α_{l+1} and conclude that $g_2(\Omega_z) \cap I_{M,l}$ is either empty or a singleton. Since $g_2(\Omega_z) \cap I_{M,k} \neq \emptyset$, $g_2(\Omega_z) \cap I_{M,l+1} \neq \emptyset$, and g_2 is continuous, it follows that $g_2 \restriction \Omega_z$ must assume all values in $I_{M,l}$. This contradicts the previous conclusion that $g_2(\Omega_z) \cap I_{M,l}$ is either empty or a singleton.

Hence we conclude that $g_2(\Omega) \cap I_{M,l+1} = \emptyset$. It follows that $q_k(\mathcal{A}) \notin I_{M,l+1}$ for all $k \in [n]$, so condition (c) of Definition 2.2 holds.

4.2. **The induction step.** Suppose that $\rho \geq 3$.

Induction hypothesis. *If $\mathcal{B}_n$, $n \in \mathbb{N}^+$, is a sequence of NNs with $\rho - 1$ layers, then there is $\psi \in [0, \mathfrak{m}]$ such that for every $\varepsilon > 0$ there is $c > 0$ such that for all sufficiently large n,*

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(o_n) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-cn}.$$

Moreover, ψ depends only on W_i, F_i and g_i for $i = 2, \ldots, \rho - 1$.

Suppose that $\mathcal{B}_n$, $n \in \mathbb{N}^+$, s a sequence of NNs with ρ layers.

Lemma 4.5. *There is $\psi \in [0, \mathfrak{m}]$ such that for every $\varepsilon > 0$ there is $c > 0$ such that for all sufficently large n,*

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : \text{ for all } a \in L_{n,\rho-1}, P^{\mathcal{A}}(a) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-cn}.$$

Moreover, ψ depends only on W_i, F_i and g_i for $i = 2, \ldots, \rho - 1$.

Proof. For every n and an arbitrary choice of $a_n \in L_{n,\rho-1}$ we construct an NN $\mathcal{B}'_n$ with only $\rho - 1$ layers as follows. Remove layer ρ (i.e. $L_{n,\rho}$) from $\mathcal{B}_n$ and remove all neurons in $L_{n,\rho-1}$ except for a_n. So the layer $\rho - 1$ of $\mathcal{B}'_n$ is $L'_{n,\rho-1} := \{a_n\}$. All other layers are left unchanged and $\mathcal{B}'_n$ will keep W_i, F_i and g_i for $i = 2, \ldots, \rho - 1$. So each $\mathcal{B}'_n$ is an NN with $\rho - 1$ layers and a unique output neuron. Let $\mathbf{S}'_n$ be the set of states of $\mathcal{B}'_n$ and $\mathbb{P}'_n$ the probability measure on $\mathbf{S}'_n$ defined just as $\mathbb{P}_n$ was defined on $\mathbf{S}_n$. By the induction hypothesis there is $\psi \in [0, \mathfrak{m}]$ such that for every $\varepsilon > 0$ there is $c > 0$ such that for all sufficiently large n,

$$\mathbb{P}'_n\big(\{\mathcal{A} \in \mathbf{S}'_n : P^{\mathcal{A}}(a_n) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-cn}$$

and ψ depends only on W_i, F_i and g_i for $i = 2, \ldots, \rho - 1$. Because of the definitions of $\mathcal{B}'_n$, $\mathbf{S}'_n$ and $\mathbb{P}'_n$, we have

$$\mathbb{P}'_n\big(\{\mathcal{A} \in \mathbf{S}'_n : P^{\mathcal{A}}(a_n) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) =$$
$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(a_n) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big),$$

so

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(a_n) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-cn}$$

Since a_n can be any neuron in $L_{n,\rho-1}$ and $|L_{n,\rho-1}| \leq \zeta(n)$ it follows that

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : \exists a \in L_{n,\rho-1}, P^{\mathcal{A}}(a) \notin (\psi - \varepsilon, \psi + \varepsilon)\}\big) \ \leq \ \zeta(n) \cdot e^{-cn}.$$

Since ζ is a polynomial there is $c' > 0$ such that for all sufficiently large n, $\zeta(n) \cdot e^{-cn} \leq e^{-c'n}$, and therefore

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : \forall a \in L_{n,\rho-1}, P^{\mathcal{A}}(a) \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-c'n}.$$

This completes the proof of the lemma. □

We will prove that there is $\varphi \in [0, 1]$ such that for every $\varepsilon > 0$ there is $c > 0$ such that for all sufficiently large n,

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P^{\mathcal{A}}(o_n) \in (\varphi - \varepsilon, \varphi + \varepsilon)\}\big) \geq 1 - e^{-cn}.$$

Moreover, φ will only depend on W_i, F_i and g_i for $i = 2, \ldots, \rho$. Let

$$L_{n,\rho-1} = \{a_1, \ldots, a_{t(n)}\}, \text{ so } |L_{n,\rho-1}| = t(n).$$

Let us use the following abbreviations:

$$W := W_\rho, \ F := F_\rho, \text{ and } g := g_2.$$

For any n, $\mathcal{A} \in \mathbf{S}_n$ and $k \in [t(n)]$, let

$$q_k(\mathcal{A}) := g(P^{\mathcal{A}}(a_k), E^{\mathcal{B}_n}(a_k, o_n))$$

and, for $M \in \mathbb{N}^+$ and $l \in \{0, \ldots, M-1\}$,

$$\omega(\mathcal{A}, M, l) := \frac{\big|\{k \in [t(n)] : q_k(\mathcal{A}) \in I_{M,l}\}\big|}{t(n)}.$$

Let ψ be as in Lemma 4.5, so ψ depends only on W_i, F_i and g_i for $i = 2, \ldots, \rho-1$. For $n \in \mathbb{N}^+$ and $\delta > 0$ define

$$\mathbf{X}_n(\delta) := \big\{\mathcal{A} \in \mathbf{S}_n : \text{ for all } a \in L_{n,\rho-1},\ P^{\mathcal{A}}(a) \in (\psi - \delta, \psi + \delta)\big\}.$$

By Lemma 4.5, there is $c > 0$ such that

$$\mathbb{P}_n\big(\mathbf{X}_n(\delta)\big) \geq 1 - e^{-cn} \quad \text{ for all sufficiently large } n.$$

Recall that F is a continuous aggregation function and, for $\mathcal{A} \in \mathbf{S}_n$,

$$P^{\mathcal{A}}(o_n) \ = \ F\big(q_1(\mathcal{A}), \ldots, q_{t(n)}(\mathcal{A})\big).$$

Therefore it suffices to prove that for every $M \in \mathbb{N}^+$ there are $\alpha_0, \ldots, \alpha_{M-1} \in [0,1]$ such that condition (c) of Definition 2.2 holds and, for every $\delta > 0$ such that $\delta < \alpha_l$ if $\alpha_l > 0$, and if $\delta' > 0$ is sufficiently small, n is sufficiently large, and $\mathcal{A} \in \mathbf{X}_n(\delta')$, then for all $l = 0, \ldots, M-1$

$$\alpha_l - \delta \leq \omega(\mathcal{A}, M, l) \leq \alpha_l + \delta. \tag{4.12}$$

Since $P^{\mathcal{A}}(a) \in (\psi - \delta', \psi + \delta')$ if $\mathcal{A} \in \mathbf{X}_n(\delta')$ and $a \in L_{n,\rho-1}$, where $\delta' > 0$ can be chosen as small as we like, we essentially get a one dimensional problem, in contrast to the corresponding two dimensional problem in the base case. Let

$$\Omega_0 := \big\{x \in [0, \mathfrak{m}] : W(x) > 0\big\} \ \text{ and } \ \text{let } \Omega \text{ be the closure of } \Omega_0.$$

Note that Ω is a finite union of closed intervals. Define $g_\psi : [0, \mathfrak{m}] \to [\mathfrak{m}]$ by $g_\psi(x) := g(\psi, x)$. For $l = 0, \ldots, M-1$, let

$$X_l := g_\psi^{-1}(I_{M,l}) \cap \Omega \quad \text{ and } \quad \alpha_l := \int_{X_l} W(x)dx.$$

Note that X_l is a finite union of closed intervals. By Assumption (A7), the proportion of $a_k \in L_{n,\rho}$ such that $E^{\mathcal{B}_n}(a_k, o_n) \in X_l$ is in $(\alpha_l - \delta', \alpha_l + \delta')$ if n is large enough. Since g is continuous we can choose $\delta' > 0$ small enough so that if n is large enough and $\mathcal{A} \in \mathbf{X}_n(\delta')$, then (4.12) holds for all $l = 0, \ldots, M-1$. Hence conditions (a) and (b) of Definition 2.2 are satisfied by $\bar{q} := (q_1(\mathcal{A}), \ldots, q_{t(n)})$ if n is sufficiently large.

It remains to check that condition (c) of Definition 2.2 holds. By a similar argument as in the base case we get that if $\alpha_l = \alpha_{l+1} = \alpha_{l+2}$, then $g_\psi(X_{l+1}) \cap I_{M,l+1} = \emptyset$, so $g(\psi, X_{l+1}) \cap I_{M,l+1} = \emptyset$. As $g(\psi, X_{l+1})$ and $I_{M,l+1}$ are closed the distance between them is positive. By taking $\delta' > 0$ sufficiently small we may assume that if $x \in [\psi - \delta', \psi + \delta']$ then $g(x, X_{l+1}) \cap I_{M,l+1} = \emptyset$. It follows that $q_k(\mathcal{A}) \notin I_{M,l+1}$ for all $\mathcal{A} \in \mathbf{X}_n(\delta')$ and $k \in [t(n)]$. Thus condition (c) of Definition 2.2 holds for $\bar{q}$.

4.3. **Proof of Corollary 3.4.** Suppose $\mathcal{B}_n$, $n \in \mathbb{N}^+$, is a sequence of NNs subject to the assumptions stated in Section 3 and that $\mathbb{P}_n$ is as stated in the same section. Let ψ be as in Theorem 3.3.

The proof of Theorem 3.3 (see Lemma 4.5) shows that there is $\varphi \in [0, \mathfrak{m}]$ such that for all $\delta > 0$ there is $c > 0$ such that

$$\mathbb{P}_n\big(\big\{\mathcal{A} \in \mathbf{S}_n : \text{ for all } a \in L_{n,2}, P^{\mathcal{A}}(a) \in (\varphi - \delta, \varphi + \delta)\big\}\big) \geq 1 - e^{-cn}. \tag{4.13}$$

Let $h : [0, \mathfrak{m}] \to [0, \mathfrak{m}]$ be any continuous function such that $h(\varphi) = \psi$. Let $F_2' : [0, \mathfrak{m}]^{fin} \to [0, \mathfrak{m}]$ be the aggregation function defined by $F_2'(x_1, \ldots, x_n) = h(F_2(x_1, \ldots, x_n))$

for all n and $(x_1, \ldots, x_n) \in [0, \mathfrak{m}]^n$. By Lemma 2.4, F_2' is a continuous aggregation function.

Define a sequence $\mathcal{B}_n'$, $n \in \mathbb{N}^+$, of NNs with only two layers as follows:

(1) Layer 1 of $\mathcal{B}_n'$ is identical with layer 1 of $\mathcal{B}_n$ and layer 2 of $\mathcal{B}_n'$ is $L_{n,2}' := \{o_n'\}$ where o_n' is an aribtrary choice of neuron in layer 2 of $\mathcal{B}_n$.
(2) The weights of connections from $L_{n,1}$ to o_n' in $\mathcal{B}_n'$ are the same as the weights of the same connections in $\mathcal{B}_n$.
(3) The function $g_2 : [0, \mathfrak{m}]^2 \to [0, \mathfrak{m}]$ of $\mathcal{B}_n$ is also used by $\mathcal{B}_n'$, but instead of F_2, $\mathcal{B}_n'$ uses the continuous aggregation function F_2', so to determine the value of o_n', $\mathcal{B}_n'$ uses the activation function

$$F_2'(g_2(v_1, w_1), \ldots, g_2(v_n, w_n))$$

where v_i is the value of the i^{th} neuron in layer 1 and w_i is the weight of the connection from the i^{th} neuron in layer 1 to o_n'

We let the state space $\mathbf{S}_n'$ of $\mathcal{B}_n'$ be defined as in Section 3 (but with $\mathbf{S}'$ and $\mathcal{B}_n'$ in place of $\mathbf{S}_n$ and $\mathcal{B}_n$) and we let $\mathbb{P}_n'$ be the probability distribution induced by μ_n on $\mathbf{S}_n'$ in the same way as in that section.

Let $\varepsilon > 0$. As h is continuous and maps φ to ψ, it follows from (4.13) and the construction of $\mathcal{B}_n'$ that if $\delta > 0$ is chosen sufficiently small in (4.13), then

$$\mathbb{P}_n'\big(\{\mathcal{A} \in \mathbf{S}_n' : P^{\mathcal{A}}(o_n') \in (\psi - \varepsilon, \psi + \varepsilon)\}\big) \geq 1 - e^{-cn}.$$

Remark 4.6. (Generalization to the context of several features with their own values) Theorem 3.3 and Corollary 3.4 can without any new ideas, but at the expense of messier book keeping and notation, be generalized to the following context. Let every neuron have κ "features" where each feature has a value in $[0, \mathfrak{m}]$. The features can be represented by relation symbols $P_1, \ldots, P_\kappa$. For each layer $i = 2, \ldots, \rho$ and index $j = 1, \ldots, \kappa$ of a feature, there are continuous aggregation functions $F_{i,j} : [0, \mathfrak{m}]^{fin} \to [0, \mathfrak{m}]$ and continuous $g_{i,j} : [0, \mathfrak{m}]^{\kappa+1} \to [0, \mathfrak{m}]$ such that the value of the feature j of a neuron b of $\mathcal{B}_n$ in layer i is equal to

$$F_{i,j}\big(g_{i,j}(v_{1,1}, \ldots, v_{\kappa,1}, w_1), \ldots, g_{i,j}(v_{1,t(n,i-1)}, \ldots, v_{\kappa,t(n,i-1)}, w_{t(n,i-1)})\big)$$

if $v_{j,1}, \ldots, v_{j,t(n,i-1)}$ $(j = 1, \ldots, \kappa)$ enumerates the values of the feature j of neurons in layer $i-1$ and $w_1, \ldots, w_{t(n,i-1)}$ enumerates the weights of connections from the corresponding neurons in layer $i-1$ to b. In this context we assume that for all $j = 1, \ldots, \kappa$, there is a continuous probability density function $\mu_j : [0, \mathfrak{m}] \to [0, \infty)$ and we replace $\mu_n : [0, \mathfrak{m}]^n \to [0, \infty)$ defined as in Section 3 by the probability density function $\mu_n : [0, \mathfrak{m}]^{\kappa n} \to [0, \infty)$ defined by

$$\mu_n(x_1, \ldots, x_{\kappa n} = \mu_1(x_1) \cdot \ldots \cdot \mu_1(x_n) \cdot \ldots \cdot \mu_\kappa(x_{\kappa(n-1)+1}) \cdot \ldots \cdot \mu_\kappa(x_{\kappa n}).$$

The version of Theorem 3.3 in this context states that for every feature P_j, $j = 1, \ldots, \kappa$, there is $\psi_j \in [0, \mathfrak{m}]$ such that for all $\varepsilon > 0$ there is $c > 0$ such that for all sufficiently large n,

$$\mathbb{P}_n\big(\{\mathcal{A} \in \mathbf{S}_n : P_j^{\mathcal{A}}(o_n) \in (\psi_j - \varepsilon, \psi_j + \varepsilon)\}\big) \geq 1 - e^{-cn}.$$

References

[1] Sam Adam-Day, Theodor-Mihai Iliant, İsmail İlkan Ceylan, Zero-One Laws of Graph Neural Networks, in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023, 70733–70756, https://doi.org/10.52202/075280-3099.

[2] Sam Adam-Day, Michael Benedikt, İsmail İlkan Ceylan, Ben Finkelshtein, Almost Surely Asymptotically Constant Graph Neural Networks, in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024, 124843–124886, https://doi.org/10.52202/079017-3965.

[3] Noga Alon and Joel H. Spencer, *The Probabilistic Metod, Second Edition*, John Wiley & Sons (2000).

[4] I. Ben Yaacov, A. Berenstein, C. W. Henson and A. Usvyatsov, Model theory for metric structures, in Z. Chatzidakis, D. Machpherson, A. Pillay and A. Wilkie (editors), *Model theory with applications to algebra and analysis, London Mathematical Society Lecture Notes Series*, Vol. 350 (2008) 315–427.
[5] Guy Van den Broeck, Kristian Kersting, Sriraam Natarajan, David Poole (editors), *An Introduction to Lifted Probabilistic Inference*, MIT Press (2021), https://doi.org/10.7551/mitpress/10548.001.0001.
[6] C. C. Chang and H. Jerome Keisler, *Continuous Model Theory*, Princeton University Press (1966).
[7] H. Chernoff, A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Annals of Mathematical Statistics*, Vol. 23 (1952) 493–509.
[8] Fabio G. Cozman, Denis D. Mauá, The finite model theory of Bayesian network specifications: Descriptive complexity and zero/one laws, *International Journal of Approximate Reasoning*, vol. 110 (2019), 107–126, https://doi.org/10.1016/j.ijar.2019.04.003.
[9] Luc De Raedt, Kristian Kersting, Sriraam Natarajan, David Poole, *Statistical Relational Artificial Intelligence: Logic, Probability, and Computation*, Morgan & Claypool Publishers (2016), https://doi.org/10.2200/S00692ED1V01Y201601AIM032.
[10] Heinz-Dieter Ebbinghaus, Jörg Flum, *Finite Model Theory*, second revised and enlarged edition, Springer, 1999, https://doi.org/10.1007/3-540-28788-4.
[11] Lise Getoor, Ben Taskar (editors), *Introduction to Statistical Relational Learning*, MIT Press (2007), https://doi.org/10.7551/mitpress/7432.001.0001.
[12] Grädel, Kolaitis, Libkin, Marx, Spencer, Vardi, Venema, Weinstein, *Finite Model Theory and Its Applications*, Springer (2007).
[13] Wilfrid Hodges, *Model Theory*, Cambridge University Press (1993).
[14] Manfred Jaeger, Convergence results for relational Bayesian networks, in *Proceedings of the Thirteenth Annual IEEE Symposium on Logic in Computer Science (LICS 1998)*, IEEE Computer Society (1998), 44–55, https://doi.org/10.1109/LICS.1998.705642.
[15] Angelika Kimmig, Lilyana Mihalkova, Lise Getoor, Lifted graphical models: A survey, *Machine Learning*, vol. 99 (2015), 1–45, https://doi.org/10.1007/s10994-014-5443-2.
[16] Vera Koponen, Conditional probability logic, lifted Bayesian networks, and almost sure quantifier elimination, *Theoretical Computer Science*, vol. 848 (2020), 1–27, https://doi.org/10.1016/j.tcs.2020.08.006.
[17] Vera Koponen, Random expansions of finite structures with bounded degree, *Annals of Pure and Applied Logic*, vol. 177, no. 3 (2026), 103665, https://doi.org/10.1016/j.apal.2025.103665.
[18] Koponen, V., Karlsson, E., Convergence Laws for Expansions of Linear Preorders. In: Kozen, D., de Queiroz, R. (eds) *Logic, Language, Information, and Computation. WoLLIC 2025*, Lecture Notes in Computer Science, vol 15942. Springer (2026), https://doi.org/10.1007/978-3-031-99536-1_7
[19] Vera Koponen, *Information and Computation*, Vol. 310 (2026) 105441, https://doi.org/10.1016/j.ic.2026.105441.
[20] Vera Koponen, Domain size asymptotics for Markov logic networks, arXiv:2509.04192v2 [cs.AI], 2026, https://doi.org/10.48550/arXiv.2509.04192.
[21] Vera Koponen, Yasmin Tousinejad, Random expansions of trees with bounded height, *Theoretical Computer Science*, vol. 1040 (2025), 115201, https://doi.org/10.1016/j.tcs.2025.115201.
[22] Vera Koponen, Felix Weitkämper, Asymptotic elimination of partially continuous aggregation functions in directed graphical models, *Information and Computation*, vol. 293 (2023), 105061, https://doi.org/10.1016/j.ic.2023.105061.
[23] Vera Koponen, Felix Weitkämper, On the relative asymptotic expressivity of inference frameworks, *Logical Methods in Computer Science*, vol. 20, no. 4 (2024), 13:1–13:52, https://doi.org/10.46298/lmcs-20(4:13)2024.
[24] Bernhard Mehlig, *Machine Learning with Neural Networks: An Introduction for Scientists and Engineers*, Cambridge University Press (2021).
[25] Yasmin Tousinejad, Vera Koponen, Random coloured digraphs defined by a Markov logic network, https://arxiv.org/abs/2606.23715.
[26] Felix Q. Weitkämper, An asymptotic analysis of probabilistic logic programming, with implications for expressing projective families of distributions, *Theory and Practice of Logic Programming*, vol. 21, no. 6 (2021), 802–817, https://doi.org/10.1017/S1471068421000314.
[27] Felix Weitkämper, Probabilities of the Third Type: Statistical Relational Learning and Reasoning with Relative Frequencies, *Journal of Artificial Intelligence Research*, vol. 80 (2024), 1407–1436, https://doi.org/10.1613/JAIR.1.15679.
[28] Yu Zhang, Peter Tino, Ales Leonardis, Ke Tang, A survey on neural network interpretability, *IEEE Transactions on Emerging Topics in Computational Intelligence*, Vol 5 (2021) 726–742.

Vera Koponen, Department of Mathematics, Uppsala University, Sweden.
Email address: vera.koponen@math.uu.se