Parameter identifiability of a deep feedforward ReLU neural network

Joachim Bona-Pellissier[1], François Bachoc[2], and François Malgouyres[2]

[1]Institut de Mathématiques de Toulouse, UMR5219. Université de Toulouse, CNRS. UT1, F-31042 Toulouse, France

[2]Institut de Mathématiques de Toulouse, UMR5219. Université de Toulouse, CNRS. UPS IMT, F-31062 Toulouse Cedex 9, France

December 17, 2021

Abstract

The possibility for one to recover the parameters –weights and biases– of a neural network thanks to the knowledge of its function on a subset of the input space can be, depending on the situation, a curse or a blessing. On one hand, recovering the parameters allows for better adversarial attacks and could also disclose sensitive information from the dataset used to construct the network. On the other hand, if the parameters of a network can be recovered, it guarantees the user that the features in the latent spaces can be interpreted. It also provides foundations to obtain formal guarantees on the performances of the network.

It is therefore important to characterize the networks whose parameters can be identified and those whose parameters cannot.

In this article, we provide a set of conditions on a deep fully-connected feedforward ReLU neural network under which the parameters of the network are uniquely identified –modulo permutation and positive rescaling– from the function it implements on a subset of the input space.

Keywords: ReLU networks, Equivalent parameters, Symmetries, Parameter recovery, Deep Learning

1 Introduction

The development of Machine Learning and in particular of Deep Learning in the last decade has led to many breakthroughs in fields such as image classification [27], object recognition [46, 47], speech recognition [23, 49, 22], natural language processing [36, 37, 26], anomaly detection [45] or climate sciences [2]. Deep neural networks are now widely used in real-life tasks stemming from those fields and beyond. This development and the diversity of contexts in which neural networks are used require to investigate theoretical properties that permit to guarantee that they can be used safely, are robust to attack, and can be used widely without giving access to sensitive information.

One key problem in these regards is the relation between the parameters and the function implemented by the network. If a parameterization of a network uniquely defines a function, the reverse is not true. Which other parameterizations define the same function, and what do they have in common? Which information on the parameters of a network are we able to infer from the knowledge of its function on a given domain? Addressing these questions is important for different reasons: industrial property, privacy, robustness and efficiency guarantee (see Section 2 for further discussions and references).

In this article, we consider fully-connected feedforward neural networks with K layers, $K \geq 2$, with the ReLU activation function (see Section 3 for details). The weights and bias parameterizing a neural network are gathered in a list $\mathbf{M}$ of matrices and a list $\mathbf{b}$ of vectors. The corresponding function is denoted[1] $f_{\mathbf{M},\mathbf{b}} : \mathbb{R}^{n_K} \longrightarrow \mathbb{R}^{n_0}$. We say that two parameterizations $(\mathbf{M},\mathbf{b})$ and $(\tilde{\mathbf{M}},\tilde{\mathbf{b}})$ are *equivalent* if they can be deduced from each other by the permutation of neurons in each hidden layer and by positive rescaling between the inward and outward weights of every neuron of every hidden layer. These two operations, that are precisely defined in Definition 3, are well-known in the literature [43, 44, 48, 52] and will be referred to as 'permutation and positive rescaling'. As is well known and restated for completeness in Proposition 4, if two parameterizations $(\mathbf{M},\mathbf{b})$ and $(\tilde{\mathbf{M}},\tilde{\mathbf{b}})$ are equivalent, then the corresponding networks implement the same function: for all $x \in \mathbb{R}^{n_K}$, $f_{\mathbf{M},\mathbf{b}}(x) = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(x)$. In other words, parameter equivalence implies *functional equivalence* of the networks.

The main contribution of this article is an *identifiablity statement* (see Theorem 6) which establishes a 'weak' converse of this statement. We consider a set $\Omega \subset \mathbb{R}^{n_K}$ and two parameterizations $(\mathbf{M},\mathbf{b})$ and $(\tilde{\mathbf{M}},\tilde{\mathbf{b}})$ sharing the same architecture (number of layers and of neurons per layer). We establish a *sufficient condition* $\mathbf{P}$ such that, if for all $x \in \Omega$, $f_{\mathbf{M},\mathbf{b}}(x) = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(x)$ and the condition $\mathbf{P}$ is met, then the two parameterizations $(\mathbf{M},\mathbf{b})$ and $(\tilde{\mathbf{M}},\tilde{\mathbf{b}})$ are equivalent. The motivation for the introduction of the set Ω is that, in practice, we may only test the values of $f_{\mathbf{M},\mathbf{b}}$ and $f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}$ on a subset of $\mathbb{R}^{n_K}$. Typically, Ω is a subset of the support of the input distribution law. Such a setting also allows to show that two networks which coincide on a given domain actually coincide on the whole input space $\mathbb{R}^{n_K}$. Indeed, if the functions implemented by the networks coincide on Ω and if the sufficient condition $\mathbf{P}$ is satisfied, then the parameters are equivalent and thus by Proposition 4 the functions also coincide on the rest of the input space $\mathbb{R}^{n_K}$. This can be useful to bound the generalization error.

We also reformulate this identifiability statement (see Corollary 7) in a way that illustrates its interest with regard to risk minimization. The corollary considers a random variable X generating the input and an output of the form $Y = f_{\mathbf{M},\mathbf{b}}(X)$, for some parameters $(\mathbf{M},\mathbf{b})$. It states that, when the condition $\mathbf{P}$ is met, any estimated neural network $(\tilde{\mathbf{M}},\tilde{\mathbf{b}})$ for which the population risk equals 0 belongs to the equivalence class of $(\mathbf{M},\mathbf{b})$. In words, the only way to have a perfect prediction is to perfectly recover $(\mathbf{M},\mathbf{b})$, up to permutation and positive rescaling.

We describe the related works in Section 2. In addition to the works providing identifiability, stability or stable recovery statements, we give a few pointers on privacy, robustness and guarantees of efficiency that motivate our study from an applied perspective. We define in Section 3 the considered neural networks and provide the (known) properties that are useful in our context. The sufficient condition $\mathbf{P}$ and the main theorems are in Section 4. The sketch of the proofs is in Section 5 and the details are in the Appendix.

2 Related work

2.1 Identifiability, stability and stable recovery

2.1.1 Identifiability

Identifiability of the parameters of neural networks has been the topic of a fair amount of work. For smooth activation functions, some results were already established in the 1990s. For shallow networks, results exist for activation functions amongst which tanh [55, 3], the logistic sigmoid [29], or the Gaussian and rational functions [25]. For deep networks, [16] shows that with tanh as activation function, with only a few generic conditions on the parameters, two networks that implement the same function have the same architecture and the same parameters up to some permutations and sign-flip operations.

In the case of ReLU networks, we have seen that two operations are well known to preserve the function implemented by the network: permutation and positive rescaling. These operations define equivalence classes on the set of parameters, and we can at best identify the parameters of a network up to these equivalences. It is shown in [44] that these operations are the only generic operations of this kind for ReLU networks with nonincreasing number of neurons per layer. Indeed, they show that for any fully-connected ReLU network architecture with nonincreasing number of neurons per layer, for any nonempty open set Ω, there exists a parameterization $(\mathbf{M},\mathbf{b})$ such that for any other parameterization $(\tilde{\mathbf{M}},\tilde{\mathbf{b}})$ satisfying some generic assumption, if $f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}$ coincides with $f_{\mathbf{M},\mathbf{b}}$ on Ω, then

[1] For clarity of the proofs, we index the layers from K (input) to 0 (output). The input layer is not counted hence the 'K layers'.

$(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ in the equivalence class of $(\mathbf{M}, \mathbf{b})$.

In this work, in order to establish identifiability, we take advantage of the piecewise linear geometry of the functions implemented by ReLU networks to identify the parameters. Indeed, it is well known that the function defined by a deep ReLU network is continuous piecewise-linear, i.e. we can partition the input space into polyhedral regions, sometimes called ‘linear regions’, over which the function is affine. These regions are separated by boundaries that are made of pieces of hyperplanes and that correspond to the non differentiabilities of the function. One crosses such a boundary when the pre-activation value of a neuron (before applying the ReLU function) changes sign. By observing the boundary, one can infer information about the weights and bias of the said neuron.

Other articles adopt similar strategies for shallow [43] or deep networks [48, 52, 53]. The specificity of our proof is to proceed by induction, identifying the weights and bias layer after layer. We discuss the differences between our condition $\mathbf{P}$ and the sufficient conditions given in [43, 48, 53] in detail in Section 4.2.

In the case of shallow ReLU networks, [43] establishes a sufficient condition on the parameters for identifiability. If the condition is satisfied by two two-layer fully-connected feedforward ReLU networks whose functions coincide on all the input space, then the parameters of one network can be obtained from the parameters of the other network by permutation and positive rescaling.

In the case of deep ReLU networks, [48] gives a sufficient condition to be able to reconstruct the architecture, weights and biases of a deep ReLU network by knowing its input-output map on all the input space. The condition concerns the boundaries mentioned above: for each neuron in a hidden layer, the authors define the boundary associated to the neuron as the points at which the pre-activation value of the neuron is zero. Then, the condition requires each boundary associated to a neuron in a layer k to intersect the boundaries associated to all the neurons in layer $k+1$ and $k-1$ (see Section 4.2 for more details).

Another kind of property is local identifiability, which is identifiability of a parameter $(\mathbf{M}, \mathbf{b})$ amongst a set of parameters that are close to $(\mathbf{M}, \mathbf{b})$. [53] studies this property for shallow and deep networks. For a deep ReLU network, it first shows that under a trivial assumption, general identifiability up to permutation and positive rescaling implies local identifiability up to positive rescaling, and that the non-existence of ‘twin’ neurons is necessary to identifiability and local identifiability. Then, [53] makes a breakthrough by giving an abstract necessary and sufficient condition on $(\mathbf{M}, \mathbf{b})$ such that there exists a well-chosen *finite* set Ω from which local identifiability holds up to positive rescalings, and it gives a bound on the size of the set.

2.1.2 Inverse stability and stable recovery

Establishing identifiability properties is a first step towards establishing inverse stability properties and studying stable recovery algorithms. Given a norm between functions, we say that inverse stability holds when the proximity of the functions implemented by two networks with the same architecture implies the proximity of the corresponding parameters -up to equivalences of parameters, for instance permutations and positive rescalings in the case of ReLU networks. Inverse stability is a stronger property than identifiability, and is necessary for stable recovery algorithms, which goal is to practically recover the parameters of a network from its function.

Inverse stability does not hold in general with the uniform norm for fully-connected feedforward neural networks. Indeed, [42] shows that for any depth, for any architecture with at least 3 neurons in the first hidden layer and any practically used activation function, there exists a sequence of networks whose function tends uniformly to 0 while any parameterization of these networks tends to infinity.

Many inverse stability and stable recovery results already exist for shallow networks. [15] studies inverse stability directly up to functional equivalence classes, without specifying the nature of these classes in terms of parameters -which interests us in this paper. The authors show that inverse stability has interesting implications in terms of optimization, allowing to link the minima in the parameter space to minima in the realization space (the space of all the functions that can be implemented by a network) and to estimate the quality of local minima in the parameter space based on their radii. Referring to the counter-example given by [42], the authors of [15] argue that the Sobolev norm is more suited than the uniform norm to the problem of inverse stability. With this norm, they concretely establish an inverse stability result on shallow ReLU networks without bias, under a few conditions on the parameters.

When it comes to stable recovery algorithms, [18] provides a sample complexity under which one can recover the parameters of a shallow network with sigmoid activation function using cross-entropy as a loss. For shallow fully-connected ReLU networks, without bias and with Gaussian input, [19, 59, 60, 61] study the stable recovery of the parameters of a teacher network. They give a sample complexity under which minimizing the empirical risk allows to recover the parameters of the network. [31] studies the same configuration but with an identity mapping that skips one layer. ReLU networks can also be used to recover a network with absolute value as activation function [30]. In fact, a neuron with absolute value can be seen as a sum of two ReLU neurons.

Some results also exist in the case of shallow convolutional networks. [7, 58, 57, 14] establish stable recovery results for convolutional ReLU networks with no overlapping. [24] gives a result in the case of a sigmoidal activation function. The case of convolutional ReLU networks with overlapping is studied in [20].

Stability and stable recovery for *deep* networks is a more complicated question. A few results exist on the subject, but it stays mostly unexplored.

Among them, for deep structured linear networks, [33, 34, 35] use a tensorial lifting technique to establish inverse stability properties. [33, 34] establish necessary and sufficient conditions of inverse stability for a general constraint on the parameters defining the network. [35] specializes the analysis to the sparsity constraint on the parameters, and obtains necessary and sufficient conditions of inverse stability.

The authors of [4] consider deep feed-forward networks with Heavyside activation function which are very sparse and randomly generated. They show that these can be learned with high probability one layer after another.

The authors of [51] consider a deep feed-forward neural network, with an activation function that can be, inter alia, ReLU, sigmoid or softmax. They show that, if the input is Gaussian or its distribution is known, and if the weight matrix of the first layer is sparse, then a method based on moments and sparse dictionary learning can retrieve it exactly. Nothing is said about the stability or the estimation of the other layers.

For deep ReLU networks, in the case where one has full access to the function implemented by the network [48] provides a practical algorithm able to approximately recover the parameters modulo permutation and rescaling, and [8] reconstructs a functionally equivalent network, formulating it as a cryptanalytic problem.

Further inverse stability and stable recovery results for deep ReLU networks are still to be established. Studying identifiability for these networks, as we do in this article, is a first step towards this goal.

2.2 Motivations: privacy, robustness and interpretability

The generalization of deep networks in various applications such as life style choices or medical diagnosis has raised new concerns about privacy and security. Indeed, to perform well, neural networks need to be trained with many examples. The training of some models can take up to several weeks, and need huge datasets such as ImageNet, which contains millions of images. For instance, the training of the giant GPT-3 neural network costed an estimated 12 millions of dollars [6]. For this reason, trained models are valuable and their owners may want to protect them from replication.

In many applications, the training dataset also contains sensitive information that could be uncovered [40, 32, 9, 17, 12]. It is crucial, to the deployment of the solutions relying on deep networks, to guarantee that this cannot occur. For example, when the system returns a confidence indicator in the prediction or a notion of margin, the *Model Inversion Attack* described in [17] uncovers learning examples x by maximizing the confidence/margin, under a constraint that $\|f_{\mathbf{M},\mathbf{b}}(x) - y\| \leq \varepsilon$, where y is a target output. In moderate dimension, this can be achieved by simply applying $f_{\mathbf{M},\mathbf{b}}$ several times. In large dimension, the complexity of the computation is too large unless the adversary can compute $\nabla f_{\mathbf{M},\mathbf{b}}(x)$, for any x. To perform this computation, the adversary needs to know $(\mathbf{M}, \mathbf{b})$. Guaranteeing that the parameters cannot be recovered prevents this. With a slightly different objective, (differential) *privacy deep learning* also assumes that the adversary has the knowledge of the network parameters [1].

Furthermore, knowing the architecture and parameters of a network could make easier for a malicious user to attack it, for instance with adversarial attacks. Indeed, if some black-box adversarial attacks do exist [54, 50, 13], many of them use the knowledge of the parameters of the network, at least to compute the gradients [56, 21, 28, 41, 10, 39, 38, 5].

For all these reasons, the authors of [11] developed a method of preventing parameters extraction by artificially complexifying the network without changing its global behavior. This method builds on previous works on stable recovery of the parameters of the ReLU networks, and in particular on the fact that the piecewise-linear structure of the functions implemented by such networks can be used to recover the parameters. Further understanding of stable recovery for deep networks could help improve protection methods.

Another interest of our work is interpretability of deep neural networks. In some uses of deep networks we want to understand what happens at a layer level and how we can interpret the feature spaces defined by the different layers. But such an interpretation is more meaningful if we know that, for a given function implemented by the network, the parameterization is unique -up to elementary operations such as permutations and positive rescalings for ReLU networks.

3 Neural networks

In this section, we provide known definitions and properties of neural networks with ReLU activation functions. For a self-contained reading, all the corresponding proofs are provided in the appendix.

3.1 Parameterization of neural networks

We consider deep feedforward ReLU networks with $K \geq 2$ layers. To clarify any ambiguity, note that the input layer is not actually counted, as it does not gather any weights. As evoked in the introduction, we index the layers of a deep neural network in reverse order, from K to 0, for some $K \geq 2$. The input layer is the layer K, the output layer is the layer 0, and between them are $K-1$ hidden layers. We denote by $n_k \in \mathbb{N}^*$ the number of neurons of the layer k. The information contained at the layer k is a n_k-dimensional vector.

Let $k \in [\![0, K-1]\!]$. We denote the weights between the layer $k+1$ and the layer k with a matrix $M^k \in \mathbb{R}^{n_k \times n_{k+1}}$. We also consider a bias vector $b^k \in \mathbb{R}^{n_k}$ at the layer k, and the ReLU activation function, that is $\sigma(x) = \max(x, 0)$. By extension, for a vector $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$ we also write $\sigma(x) = (\sigma(x_1), \dots, \sigma(x_p))^T$. We denote by h_k the action of the network between the layer $k+1$ and the layer k. If $x \in \mathbb{R}^{n_{k+1}}$ is the information contained at the layer $k+1$, the information contained at the layer k is:

$$h_k(x) = \begin{cases} \sigma(M^k x + b^k) & \text{if } k \neq 0 \\ M^k x + b^k & \text{if } k = 0. \end{cases} \tag{1}$$

The parameters of the network can be summarized in the couple $(\mathbf{M}, \mathbf{b})$, where

$$\mathbf{M} = (M^0, M^1, \dots, M^{K-1}) \in \mathbb{R}^{n_0 \times n_1} \times \cdots \times \mathbb{R}^{n_{K-1} \times n_K}$$

and

$$\mathbf{b} = (b^0, b^1, \dots, b^{K-1}) \in \mathbb{R}^{n_0} \times \cdots \times \mathbb{R}^{n_{K-1}}.$$

The function implemented by the network is then

$$f_{\mathbf{M},\mathbf{b}} = h_0 \circ h_1 \circ \cdots \circ h_{K-1},$$

from $\mathbb{R}^{n_K}$ to $\mathbb{R}^{n_0}$. We refer to Figure 1 for a representation of a neural network and its parameters.

3.2 Continuous piecewise linear functions and neural networks

We will actively use the fact that the function implemented by a deep ReLU network as well as the intermediate functions between layers are continuous piecewise linear, which means that we can partition their domain of definition in closed polyhedral subsets such that they are linear on each subset. In this paper we use indifferently ‘linear’ or ‘affine’ to describe functions of the form $x \mapsto Ax + b$, with $A \in \mathbb{R}^{n \times m}$ some matrix and $b \in \mathbb{R}^n$ some vector.

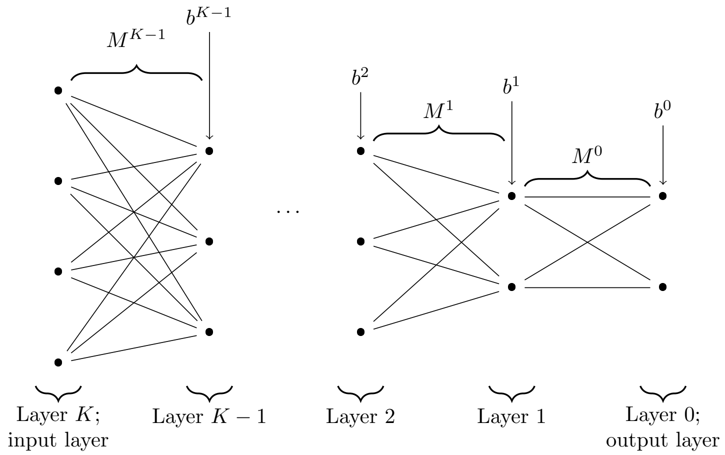

Figure 1: The parameters $\mathbf{M}$ and $\mathbf{b}$ of a neural network.

More precisely, for $m \in \mathbb{N}$, a subset $D \subset \mathbb{R}^m$ is a *closed polyhedron* iif there exist $q \in \mathbb{N}$, $a_1, \ldots, a_q \in \mathbb{R}^m$ and $b_1, \ldots b_q \in \mathbb{R}$ such that for all $x \in \mathbb{R}^m$,

$$x \in D \quad \Longleftrightarrow \quad \begin{cases} a_1^T x + b_1 \leq 0 \\ \vdots \\ a_q^T x + b_q \leq 0. \end{cases} \tag{2}$$

By convention, if $q = 0$, we obtain an empty system of equations which is satisfied for any $x \in \mathbb{R}^m$, meaning the set $\mathbb{R}^m$ is a closed polyhedron.

We say that a function $g : \mathbb{R}^m \to \mathbb{R}^n$ is *continuous piecewise linear* if there exists a finite set of closed polyhedra whose union is $\mathbb{R}^m$ and such that g is linear over each polyhedron.

It is easy to show (see Proposition 14 in the appendix) that this definition implies the continuity of the function, hence the 'continuous' in the name. We do not require here the polyhedra to be disjoint and in fact, there are always some overlaps between the borders of adjacent polyhedra. For a given continuous piecewise linear function g, there are infinitely many possible sets of closed polyhedra that match the definition. Among them, we can always find one such that all the polyhedra D have nonempty interior $\mathring{D}$ (see Proposition 17 in the appendix). We call such a set admissible, as in the following definition.

Definition 1. Let $g : \mathbb{R}^m \to \mathbb{R}^n$ be a continuous piecewise linear function. Let Π be a set of closed polyhedra of $\mathbb{R}^m$. We say that Π is *admissible* with respect to g if and only if:

$$\begin{cases} \bigcup_{D \in \Pi} D = \mathbb{R}^m, \\ \text{for all } D \in \Pi, \ g \text{ is linear on } D, \\ \text{for all } D \in \Pi, \ \mathring{D} \neq \emptyset. \end{cases} \tag{3}$$

We now define additional functions associated to a network. Recall the layer functions h_k defined in (1), that represent the actions of the network between successive layers. Let $k \in [\![0, K]\!]$. We define the following functions:

$$\begin{aligned} f_k &= h_k \circ h_{k+1} \circ \cdots \circ h_{K-1}; \\ g_k &= h_0 \circ h_1 \circ \cdots \circ h_{k-1}. \end{aligned} \tag{4}$$

Above, by convention, we let $f_K = id_{\mathbb{R}^{n_K}}$ and $g_0 = id_{\mathbb{R}^{n_0}}$, where $id_{\mathbb{R}^m}$ denotes the identity function on $\mathbb{R}^m$. The function $f_k : \mathbb{R}^{n_K} \mapsto \mathbb{R}^{n_k}$ represents the action of the network between the input layer and the layer k. The function $g_k : \mathbb{R}^{n_k} \mapsto \mathbb{R}^{n_0}$ represents the action of the network between the layer k and the output layer. Hence, for all $k \in [\![0, K]\!]$ we have $g_k \circ f_k = f_{\mathbf{M},\mathbf{b}}$, and in particular $f_0 = g_K = f_{\mathbf{M},\mathbf{b}}$.

The following proposition is easy to show by induction and using the fact that the composition of two continuous piecewise linear functions is also continuous piecewise linear (see Proposition 27 in the appendix).

Proposition 2. *For all $k \in [\![0, K]\!]$, f_k and g_k are continuous piecewise linear.*

In particular, $f_{\mathbf{M},\mathbf{b}}$ is continuous piecewise linear.

We say that a list of sets of closed polyhedra $\mathbf{\Pi} = (\Pi_1, \dots, \Pi_{K-1})$ is *admissible* with respect to $(\mathbf{M}, \mathbf{b})$ iif for all $k \in [\![1, K-1]\!]$, the set of closed polyhedra Π_k is admissible with respect to g_k. Since there always exist such Π_k (from Proposition 2 and Proposition 17 in Appendix A), there always exists an admissible list $\mathbf{\Pi}$.

3.3 Equivalence between two parameterizations

We are interested in sufficient conditions to identify the parameters of a network from its function. As discussed previously, some elementary operations on the parameters are well known to preserve the function of a network, so what we shall actually identify is the equivalence class of the parameters modulo these operations. There are two such operations:

- the permutation of neurons of a layer;
- the positive rescalings, that is multiplying all the outward weights of a neuron by a strictly positive number and dividing the inward weights by the same number.

We define below the corresponding equivalence relation, after introducing some notations. For all $m \in \mathbb{N}^*$, we denote by $\mathfrak{S}_m$ the set of all permutations of $[\![1, m]\!]$. For any permutation $\varphi \in \mathfrak{S}_m$, we denote by P_φ the $m \times m$ permutation matrix associated to φ, whose coefficients are defined as

$$(P_\varphi)_{i,j} = \begin{cases} 1 & \text{if } \varphi(j) = i \\ 0 & \text{otherwise.} \end{cases}$$

We also denote by $\mathbb{1}_m$ the vector $(1, 1, \dots, 1)^T \in \mathbb{R}^m$, by $\mathbb{R}_+^*$ the set of strictly positive real numbers and by Id_m the $m \times m$ identity matrix.

Definition 3 (Equivalence between parameters)**.** If $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ are two parameterizations of a network, we say that $(\mathbf{M}, \mathbf{b})$ is equivalent to $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, and we write $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, if and only if there exist:

- a family of permutations $\boldsymbol{\varphi} = (\varphi_0, \dots, \varphi_K) \in \mathfrak{S}_{n_0} \times \dots \times \mathfrak{S}_{n_K}$, with $P_{\varphi_0} = \mathrm{Id}_{n_0}$ and $P_{\varphi_K} = \mathrm{Id}_{n_K}$,
- a family of vectors $\boldsymbol{\lambda} = (\lambda^0, \lambda^1, \dots, \lambda^K) \in (\mathbb{R}_+^*)^{n_0} \times \dots \times (\mathbb{R}_+^*)^{n_K}$, with $\lambda^0 = \mathbb{1}_{n_0}$ and $\lambda^K = \mathbb{1}_{n_K}$,

such that for all $k \in [\![0, K-1]\!]$,

$$\begin{cases} \tilde{M}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} \\ \tilde{b}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) b^k. \end{cases} \tag{5}$$

This relation $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ is an equivalence relation, as proved in the appendix (Proposition 33). We denote by $[\mathbf{M}, \mathbf{b}]$ the equivalence class of $(\mathbf{M}, \mathbf{b})$.

The following proposition states the well known fact that equivalent parameterizations lead to the same function. For a proof, see Proposition 34 and Corollary 35 in the appendix.

Proposition 4. *If* $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$*, then* $f_{\mathbf{M},\mathbf{b}} = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}$.

In this article we give a set of conditions under which we have a reciprocal, i.e. if two parameterizations $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ satisfying the conditions lead to the same function on a set Ω, i.e. $f_{\mathbf{M},\mathbf{b}}(x) = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(x)$ for all $x \in \Omega$, then they are equivalent: $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$.

4 Main result

4.1 Conditions

We expose in this section the conditions under which the main theorem holds. They are formalized in Definition 5 and named **P**. First, we introduce a few notations.

We consider a network with $K \geq 2$ layers and with parameters $(\mathbf{M}, \mathbf{b})$, a list of sets of closed polyhedra $\mathbf{\Pi} = (\Pi_1, \dots, \Pi_{K-1})$ admissible with respect to $(\mathbf{M}, \mathbf{b})$ and a domain $\Omega \subset \mathbb{R}^{n_K}$. Recall the definitions (1) and (4) of the functions h_k, f_k and g_k associated to the network. For all $k \in [\![1, K-1]\!]$, g_k is continuous piecewise linear, and since $\mathbf{\Pi}$ is admissible with respect to $(\mathbf{M}, \mathbf{b})$, by definition, the set of closed polyhedra Π_k is admissible with respect to g_k. For all $D \in \Pi_k$, the function g_k thus coincides with a linear function on D. Since by definition the interior of D is nonempty, we define $V^k(D) \in \mathbb{R}^{n_0 \times n_k}$ and $c^k(D) \in \mathbb{R}^{n_0}$ as the unique couple satisfying, for all $x \in D$:

$$g_k(x) = V^k(D)x + c^k(D). \tag{6}$$

For any $m, n \in \mathbb{N}^*$, for any $m \times n$ matrix Σ, for any $i \in [\![1, m]\!], j \in [\![1, n]\!]$, we denote by $\Sigma_{i,.}$ the i^{th} row vector of Σ and by $\Sigma_{.,j}$ the j^{th} column vector of Σ. We denote $E_i^k = \{x \in \mathbb{R}^{n_k},\ x_i = 0\}$, and $h_k^{lin}(x) = M^k x + b^k$. For any $m \in \mathbb{N}^*$ and any subset $A \subset \mathbb{R}^m$, we denote by ∂A the topological boundary with respect to the standard topology of $\mathbb{R}^m$.

Definition 5. We say that $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ satisfies the conditions $\mathbf{P}$ iif for all $k \in [\![1, K-1]\!]$:

$\mathbf{P}.a)$ M^k is full row rank;

$\mathbf{P}.b)$ for all $i \in [\![1, n_k]\!]$, there exists $x \in \mathring{\Omega}_{k+1}$ such that

$$M_{i,.}^k x + b_i^k = 0,$$

or equivalently

$$E_i^k \cap h_k^{lin}(\mathring{\Omega}_{k+1}) \neq \emptyset;$$

$\mathbf{P}.c)$ for all $D \in \Pi_k$, for all $i \in [\![1, n_k]\!]$, if $E_i^k \cap D \cap \Omega_k \neq \emptyset$ then $V_{.,i}^k(D) \neq 0$;

$\mathbf{P}.d)$ for any affine hyperplane $H \subset \mathbb{R}^{n_{k+1}}$,

$$H \cap \mathring{\Omega}_{k+1} \not\subset \bigcup_{D \in \Pi_k} \partial h_k^{-1}(D).$$

These conditions are invariant modulo equivalences of parameters. Indeed, we show in the appendix (Proposition 45) that if some parameters $(\mathbf{M}, \mathbf{b})$ satisfy the conditions $\mathbf{P}$, then all the parameters in their equivalence class satisfy them too.

Let us explain here the conditions $\mathbf{P}$. We will compare them to the state of the art in Section 4.2. The first condition, $\mathbf{P}.a)$, implies that for all $k \in [\![1, K-1]\!]$, the layer k has no more neurons than its predecessor, the layer $k+1$:

$$n_k \leq n_{k+1}.$$

Once this is satisfied, the condition is mild in the sense that it is satisfied for all matrices except a set of matrices of empty Lebesgue measure.

As a first remark about $\mathbf{P}.b)$, note that by taking $k = K-1$, we see that $\mathring{\Omega} = \mathring{\Omega}_K \neq \emptyset$. Thus, in the main result, the set Ω over which the function implemented by the network is assumed to be known needs to have nonempty interior. In particular, Ω cannot be a finite sample set. This limitation is already present in [48], which assumes an access to the function on the whole input space and [44] which considers the function of the network on a bounded open nonempty domain. However, as we discuss in the conclusion, it seems possible to establish a result for a finite Ω, and the conditions formulated here should be a basis for future work.

Let us see, for one fixed $k \in [\![1, K-1]\!]$ what the conditions $\mathbf{P}.b), \mathbf{P}.c)$ and $\mathbf{P}.d)$ mean. To understand them, see that they apply to the network made of the last $k+1$ layers of the initial network, having the layer $k+1$ as input layer and the layer 0 as output layer. The function implemented by such a network is

$$g_{k+1} : \mathbb{R}^{n_{k+1}} \longrightarrow \mathbb{R}^{n_0}.$$

As we explained in Section 3.2, the function implemented by a ReLU network is continuous piecewise linear: we can divide the input space $\mathbb{R}^{n_{k+1}}$ into polyhedral regions, over each of which the function is linear. We take advantage of this structure to acquire information about the parameters of the network. In particular, the boundaries of these regions convey important information. They are made of portions of hyperplanes, over which the function of the network is generally not differentiable. We use this non differentiability property to spot them. We go from one linear region to another when there is a change of sign in the pre-activation value (input of σ) of one neuron in a hidden layer. The

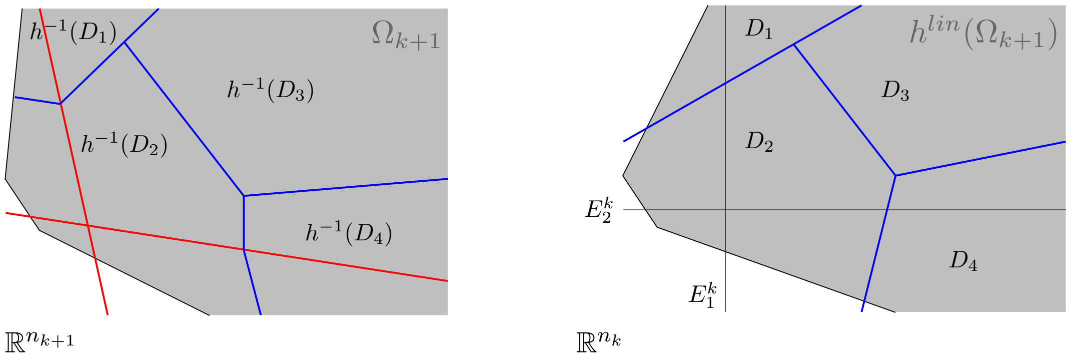

Figure 2: The admissible polyhedra with respect to g_k (right) and their inverse image by h_k (left). To make the figure lighter we write h instead of h_k. In blue, the portions of hyperplanes that separate the admissible polyhedra D_j in $\mathbb{R}^{n_k}$ (right), or their inverse images $h^{-1}(D_j)$ in $\mathbb{R}^{n_{k+1}}$ (left). The grey zone on the left side represents Ω_{k+1}, and the corresponding grey zone on the right side represents its image $h^{lin}(\Omega_{k+1})$. In red, for $i \in \{1,2\}$, H_i is the hyperplane defined by the equation $M^k_{i,.}x + b^k_i = 0$. We have $h(H_i) \subset E^k_i$.

boundary between two linear regions is thus associated to a particular neuron of a particular hidden layer.

For the partial network function g_{k+1}, we separate the action of the first layer from the action of the rest of the network thanks to the functions defined in (1) and (4), writing

$$g_{k+1} = g_k \circ h_k,$$

$$\mathbb{R}^{n_{k+1}} \xrightarrow{h_k} \mathbb{R}^{n_k} \xrightarrow{g_k} \mathbb{R}^{n_0}.$$

The goal is to identify the weights and bias of the first layer, M^k and b^k. To do so, we focus on the boundaries associated to the neurons in the first hidden layer. These 'first-order' boundaries are hyperplanes defined by the equations $M^k_{i,.}x + b^k_i = 0$, for all $i \in [\![1, n_k]\!]$. The conditions $\mathbf{P}.b)$, $\mathbf{P}.c)$ and $\mathbf{P}.d)$ are made to ensure that we are able to identify them, and consequently, the parameters M^k and b^k. The two relevant spaces are the input space of h_k, $\mathbb{R}^{n_{k+1}}$, and the input space of g_k, $\mathbb{R}^{n_k}$, which are represented in Figure 2.

The condition $\mathbf{P}.b)$ requires the hyperplane defined by the equation $M^k_{i,.}x + b^k = 0$ to intersect $\mathring{\Omega}_{k+1}$. Indeed, we only know the function on Ω_{k+1}, so this hyperplane must intersect Ω_{k+1} in order to be detectable. In the example of Figure 2, we see that the two such boundaries, which are represented as red lines, intersect Ω_{k+1}, so the condition is satisfied.

If E^k_i intersects a polyhedron $D \in \Pi_k$, then for all $x \in \mathbb{R}^{n_{k+1}}$ such that $h_k(x) \in D$, we have using (6) that the function of the partial network is $g_{k+1}(x) = \sum_{i=1}^{n_k} V^k_{.,i}(D)\sigma\left(M^k_{i,.}x + b^k_i\right) + c^k(D)$. In particular, for the points x such that $M^k_{i,.}x + b^k_i = 0$, the function $\sigma(M^k_{i,.}x + b^k_i)$ is not differentiable, and this non differentiability is reflected in the global function g_{k+1} if and only if $V^k_{.,i}(D) \neq 0$. The condition $\mathbf{P}.c)$ ensures that. In the example of Figure 2 (right part), we see that D_1 intersects E^k_1 so $\mathbf{P}.c)$ implies $V^k_{.,1}(D_1) \neq 0$. Similarly, D_2 intersects E^k_1 and E^k_2 so we must have $V^k_{.,1}(D_2) \neq 0$ and $V^k_{.,2}(D_2) \neq 0$, and the polyhedron D_4 intersects E^k_2 so we must have $V^k_{.,2}(D_4) \neq 0$.

For the last condition, $\mathbf{P}.d)$, we consider the inverse images $h_k^{-1}(D)$, for all $D \in \Pi_k$. Since h_k is piecewise linear and D is a polyhedron, $h_k^{-1}(D)$ is a finite union of polyhedra (see the first point of Proposition 18 in the appendix). In particular, its boundary $\partial h_k^{-1}(D)$ is made of pieces of hyperplanes (in blue in the left side of Figure 2). We require the union of these boundaries not to contain any full hyperplane (restricted to the domain Ω_{k+1}). In the example of Figure 2, the condition is satisfied.

4.2 Comparison with the existing work

To our knowledge, there are only two existing results on global identifiability of deep ReLU networks (with bias), as we consider here, exposed in the recent contributions [44] and [48]. Let us compare our hypotheses with theirs.

The authors of [44] introduce two notions: the notion of *general* network and the notion of *transparent network*. They note the fact that some boundaries of non differentiablity bend over some others to build a graph of dependency. The main result in [44] applies to networks whose number of neurons per layer n_k is non-increasing, as is the case in the present paper, that are transparent and general, and for which the graphs of dependency of the functions g_k satisfy additional technical conditions.

It can be verified that these hypotheses imply our conditions $\mathbf{P}.a)$, $\mathbf{P}.b)$ and $\mathbf{P}.c)$, which makes $\mathbf{P}.a)$, $\mathbf{P}.b)$ and $\mathbf{P}.c)$ more applicable.

When it comes to our last condition $\mathbf{P}.d)$, it can be compared to the technical conditions on the graph of dependency. These conditions address the way the boundaries associated to some neurons bend over the boundaries associated to neurons in previous layers. $\mathbf{P}.d)$ and this set of conditions are different, and neither implies the other.

The result exposed in [48] has a main strength compared to [44] and to us: it does not require the number of neurons per layer to be non-increasing. However, when it comes to the intersection of boundaries of linear regions, it requires each boundary, associated to some neuron, to intersect the boundaries associated to all the neurons in the previous layer, which appears to be a strong hypothesis to us. In comparison, we ask each boundary to intersect at least one of the boundaries associated to a neuron in a previous layer. Also, in [48], the function is supposed to be known on the whole input space, while [44] as well as us propose conditions on a domain Ω such that the knowledge of the function on Ω is enough. In both cases Ω has nonempty interior. [53] opens the way for considering a finite Ω by giving an abstract condition of local identifiability in that case. To our knowledge global identifiability from a finite set has not been tackled yet for deep ReLU networks.

4.3 Main theorems

We have now introduced all the necessary material to expose our main result, in Section 4.3.1, as well as an application in terms of risk minimization in Section 4.3.2.

4.3.1 Identifiability statement

Our main theorem is the following. For the proof, see Theorem 46 in Appendix *B* and its proof in Section B.4.

Theorem 6. *Let $K \in \mathbb{N}$, $K \geq 2$. Suppose we are given two networks with K layers, identical number of neurons per layer, and with respective parameters $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$. Assume $\mathbf{\Pi}$ and $\tilde{\mathbf{\Pi}}$ are two lists of sets of closed polyhedra that are admissible with respect to $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ respectively. Denote by n_K the number of neurons of the input layer, and suppose we are given a set $\Omega \subset \mathbb{R}^{n_K}$ such that $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$ satisfy the conditions $\mathbf{P}$, and such that, for all $x \in \Omega$:*

$$f_{\mathbf{M},\mathbf{b}}(x) = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(x).$$

Then:

$$(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}}).$$

As mentioned before, this theorem can be seen as a partial reciprocal to Proposition 4. Indeed, the latter shows that two networks with equivalent parameters modulo permutation and positive rescaling implement the same function. In other words, we can say that parameter equivalence implies functional equivalence of the networks. Here, we give a set of conditions, $\mathbf{P}$, under which two networks that are functionally equivalent (on a given domain Ω) have equivalent parameters modulo permutation and positive rescaling.

4.3.2 An application to risk minimization

Assume we are given a couple of input-output variables (X, Y) generated by a ground truth network with parameters $(\mathbf{M}, \mathbf{b})$:

$$Y = f_{\mathbf{M},\mathbf{b}}(X).$$

We can use Theorem 6 to show that the only way to bring the population risk to 0 is to find the ground truth parameters -modulo permutation and positive rescaling.

Indeed, let $\Omega \subset \mathbb{R}^{n_K}$ be a domain that is contained in the support of X, and suppose $L : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}_+$ is a loss function such that $L(y, y') = 0 \Rightarrow y = y'$.

Consider the population risk:

$$R(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}) = \mathbb{E}[L(f_{\tilde{\mathbf{M}}, \tilde{\mathbf{b}}}(X), Y)].$$

We have the following result. For the proof, see Corollary 47 in Appendix B and its proof in Section B.5.

Corollary 7. *Suppose there exists a list of sets of closed polyhedra* $\mathbf{\Pi}$ *admissible with respect to* $(\mathbf{M}, \mathbf{b})$ *such that* $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ *satisfies the conditions* $\mathbf{P}$*.*

If $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ *is such that there exists a list of sets of closed polyhedra* $\tilde{\mathbf{\Pi}}$ *admissible with respect to* $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ *such that* $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$ *satisfies the conditions* $\mathbf{P}$*, and if* $[\mathbf{M}, \mathbf{b}] \neq [\tilde{\mathbf{M}}, \tilde{\mathbf{b}}]$*, then:*

$$R(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}) > 0.$$

5 Sketch of proof of Theorem 6

Our main result, Theorem 6, is proven in details in Appendix B.4, and we give a sketch of the proof in this section. It is proven by induction. We are given two parameterizations $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, two lists $\mathbf{\Pi}$ and $\tilde{\mathbf{\Pi}}$ that are admissible with respect to $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ respectively, and a domain Ω that satisfy the hypotheses of Theorem 6 and we want to show that the two parameterizations are equivalent. For this, we identify the layers one after the other. To facilitate identification at a layer level, we begin with a normalisation step.

5.1 Normalisation step

Two equivalent parameterizations do not necessarily have equal weights on their layers. Indeed, the neuron permutations but more importantly the rescalings can change the structure of the intermediate layers. We are going to assume the following normalisation property: for all $k \in [\![1, K-1]\!]$, for all $i \in [\![1, n_k]\!]$, we have

$$\begin{aligned} \|M^k_{i,.}\| &= 1; \\ \|\tilde{M}^k_{i,.}\| &= 1. \end{aligned} \tag{7}$$

Indeed, we show in the appendix that for a parameterization satisfying the conditions $\mathbf{P}$, there always exists an equivalent parameterization that is normalised and that satisfies the conditions $\mathbf{P}$ (see Propositions 37 and 45). We can thus replace each parameterization $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ by an equivalent normalised parameterization. If we are able to show the normalised parameterizations are equivalent, then the original parameterizations are equivalent too.

5.2 Induction

The induction proof relies on Lemma 9 below. Let K be the number of layers of the network, and suppose the theorem is true for the networks with $K-1$ layers. As explained in section 4.1, to identify the parameters M^{K-1} and b^{K-1}, we separate the action of the first layer and of the rest of the network. For each network:

$$g_K = g_{K-1} \circ h_{K-1},$$

$$\tilde{g}_K = \tilde{g}_{K-1} \circ \tilde{h}_{K-1}.$$

We know that g_{K-1} and $\tilde{g}_{K-1}$ are continuous piecewise linear, and this will allow us to apply Lemma 9. Before stating it, we introduce a set of conditions, called $\mathbf{C}$, that need to be satisfied in order to apply it. These conditions come immediately from $\mathbf{P}$, and one can easily check that $(g_{K-1}, M^{K-1}, b^{K-1}, \Omega_K, \Pi_{K-1})$ and $(\tilde{g}_{K-1}, \tilde{M}^{K-1}, \tilde{b}^{K-1}, \Omega_K, \tilde{\Pi}_{K-1})$ satisfy $\mathbf{C}$, as a direct consequence of the conditions $\mathbf{P}$ being satisfied by $(\mathbf{M}, \mathbf{b}, \mathbf{\Omega}, \mathbf{\Pi})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \mathbf{\Omega}, \tilde{\mathbf{\Pi}})$.

Definition 8. Let l, m, n be integers, $M \in \mathbb{R}^{m\times l}$, $b \in \mathbb{R}^m$, $\Omega \subset \mathbb{R}^l$ be an open domain, let $g : \mathbb{R}^m \to \mathbb{R}^n$ a continuous piecewise linear function, and let Π be an admissible set of polyhedra with respect to g.

Let $D \in \Pi$. The function g coincides with a linear function on D. Since the interior of D is nonempty, we define $V(D) \in \mathbb{R}^{n\times m}$ and $c(D) \in \mathbb{R}^n$ as the unique couple satisfying, for all $x \in D$:

$$g(x) = V(D)x + c(D).$$

We denote $E_i = \{x \in \mathbb{R}^m, x_i = 0\}$.

We say that (g, M, b, Ω, Π) satisfies the conditions **C** iff

C.*a*) M is full row rank;

C.*b*) for all $i \in [\![1, m]\!]$, there exists $x \in \mathring{\Omega}$ such that

$$M_{i,.}x + b_i = 0,$$

or equivalently, if we denote by h^{lin} the function $x \mapsto Mx + b$, then

$$E_i \cap h^{lin}(\mathring{\Omega}) \neq \emptyset;$$

C.*c*) for all $D \in \Pi$, for all $i \in [\![1, m]\!]$, if $E_i \cap D \cap h(\Omega) \neq \emptyset$ then $V_{.,i}(D) \neq 0$;

C.*d*) for any affine hyperplane $H \subset \mathbb{R}^l$,

$$H \cap \mathring{\Omega} \not\subset \bigcup_{D\in\Pi} \partial h^{-1}(D).$$

We can now state the lemma.

Lemma 9. *Let $l, m, n \in \mathbb{N}^*$. Suppose $g, \tilde{g} : \mathbb{R}^m \to \mathbb{R}^n$ are continuous piecewise linear functions, $\Omega \subset \mathbb{R}^l$ is a subset and let $M, \tilde{M} \in \mathbb{R}^{m\times l}$, $b, \tilde{b} \in \mathbb{R}^m$. Denote $h : x \mapsto \sigma(Mx + b)$ and $\tilde{h} : x \mapsto \sigma(\tilde{M}x + \tilde{b})$. Assume Π and $\tilde{\Pi}$ are two sets of polyhedra admissible with respect to g and $\tilde{g}$.*

Suppose (g, M, b, Ω, Π) and $(\tilde{g}, \tilde{M}, \tilde{b}, \Omega, \tilde{\Pi})$ satisfy the conditions **C**, *and for all $i \in [\![1, m]\!]$, $\|M_{i,.}\| = \|\tilde{M}_{i,.}\| = 1$.*

Suppose for all $x \in \Omega$:

$$g \circ h(x) = \tilde{g} \circ \tilde{h}(x).$$

Then, there exists a permutation $\varphi \in \mathfrak{S}_m$, such that:

- *$\tilde{M} = P_\varphi M$;*
- *$\tilde{b} = P_\varphi b$;*
- *g and $y \mapsto \tilde{g}(P_\varphi y)$ coincide on $h(\Omega)$.*

Lemma 9 is restated in Appendix B as Lemma 48 and proven in Appendix C.

Applying this lemma to $(g_{K-1}, M^{K-1}, b^{K-1}, \Omega_K, \Pi_{K-1})$ and $(\tilde{g}_{K-1}, \tilde{M}^{K-1}, \tilde{b}^{K-1}, \Omega_K, \tilde{\Pi}_{K-1})$, we conclude that there exists a permutation φ_{K-1} such that

$$\begin{cases} \tilde{M}^{K-1} = P_{\varphi_{K-1}} M^{K-1} \\ \tilde{b}^{K-1} = P_{\varphi_{K-1}} b^{K-1}, \end{cases} \tag{8}$$

and that g_{K-1} and $y \mapsto \tilde{g}(P_{\varphi_{K-1}} y)$ coincide on $h_{K-1}(\Omega)$.

The functions g_{K-1} and $y \mapsto \tilde{g}(P_{\varphi_{K-1}} y)$ are the functions implemented by the networks $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ once we have removed the first layer, with a permutation of the input for the second one. Since they coincide on $\Omega_{K-1} = h_{K-1}(\Omega)$ and they satisfy the conditions **P**, we can apply the induction hypothesis to conclude the proof of Theorem 6. The complete proof is detailed in the appendices, as discussed above.

6 Conclusion

We established a set of conditions **P** under which the function implemented by a deep feedforward ReLU neural network on a subset Ω of the input space uniquely characterizes its parameters, up to permutation and positive rescaling. This contributes to the understanding of identifiability and stable recovery for deep ReLU networks, which is still largely unexplored. The conditions under which our result holds differ from the conditions of the results established in [44] and [48], which allows us to cover new situations. To be satisfied the conditions **P** need Ω to have nonempty interior, which prevents it from being a sample set. The authors of [53] are able to give a result with a finite set Ω, but for local identifiability only. Obtaining the best of both worlds, that is establishing a global identifiability result for deep ReLU networks with a finite set Ω, would be a major step forward.

Acknowledgements

Our work has benefited from the AI Interdisciplinary Institute ANITI. ANITI is funded by the French "Investing for the Future – PIA3" program under the Grant agreement n°ANR-19-PI3A-0004.

In the appendices, we restate all the notations, definitions and results of the main text, for clarity of reading. The appendices are then organized as follows. In Appendix A, we give the complete definitions and basic properties necessary to state and prove the main theorem. In Appendix B, we state the main result, Theorem 46 (Theorem 6 in the main text), and we prove it. Finally, we prove the fundamental lemma used in the proof of the main theorem, Lemma 48 (Lemma 9 in the main text), in Appendix C.

A Definitions, notations and preliminary results

Appendix A is structured as follows: after giving some notations in Section A.1, we recall the definition of a continuous piecewise linear function and some corresponding basic properties in Section A.2 and we give our formalization of deep ReLU networks as well as some well-known properties in Section A.3.

A.1 Basic notations and definitions

We denote by

$$\begin{array}{rccc} \sigma: & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & t & \longmapsto & \max(t,0) \end{array}$$

the ReLU activation function. If $x = (x_1, \dots, x_m)^T \in \mathbb{R}^m$ is a vector, we denote

$$\sigma(x) = (\sigma(x_1), \dots, \sigma(x_m))^T.$$

If $A \subset \mathbb{R}^m$, we denote by $\mathring{A}$ the interior of A and $\overline{A}$ the closure of A with respect to the standard topology of $\mathbb{R}^m$. We denote by $\partial A = \overline{A} \backslash \mathring{A}$ the topological boundary of A.

For $m, n \in \mathbb{N}^*$, we denote by $\mathbb{R}^n$ the vector space of n-dimensional real vectors and $\mathbb{R}^{m\times n}$ the vector space of real matrices with m lines and n columns. On the space of vectors, we use the norm $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$. For $x \in \mathbb{R}^n$ and $r > 0$, we denote $B(x,r) = \{y \in \mathbb{R}^n, \|y - x\| < r\}$.

For any vector $x \in \mathbb{R}^n$ whose coefficients x_i are all different from zero, we denote by x^{-1} or $\frac{1}{x}$ the vector $\left(\frac{1}{x_1}, \frac{1}{x_2}, \dots, \frac{1}{x_n}\right)^T$.

For any matrix $M \in \mathbb{R}^{m\times n}$, for all $i \in [\![1, m]\!]$, we denote by $M_{i,\cdot}$ the i^{th} line of M. The vector $M_{i,\cdot}$ is a line vector whose j^{th} component is $M_{i,j}$. Similarly, for $j \in [\![1, n]\!]$, we denote by $M_{\cdot,j}$ the j^{th} column of M, which is the column vector whose i^{th} component is $M_{i,j}$. For any matrix $M \in \mathbb{R}^{m\times n}$, we denote by $M^T \in \mathbb{R}^{n\times m}$ the transpose matrix of M.

To avoid any confusion, we will denote by $(M^T)_{i,\cdot}$ the i^{th} line of the matrix M^T and by ${M_{i,\cdot}}^T$ the transpose of the line vector $M_{i,\cdot}$, which is a column vector. Similarly, we will denote by $(M^T)_{\cdot,j}$ the j^{th} column of M^T and ${M_{\cdot,j}}^T$ the transpose of the column vector $M_{\cdot,j}$.

For $n \in \mathbb{N}^*$, we denote by Id_n the $n \times n$ identity matrix and by $\mathbb{1}_n$ the vector $(1, 1, \dots, 1)^T \in \mathbb{R}^n$.

If $\lambda \in \mathbb{R}^n$ is a vector of size n, for some $n \in \mathbb{N}^*$, we denote by $\mathrm{Diag}(\lambda)$ the $n \times n$ matrix defined by:

$$\mathrm{Diag}(\lambda)_{i,j} = \begin{cases} \lambda_i & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

For any integer $m \in \mathbb{N}^*$, we denote by $\mathfrak{S}_m$ the set of all permutations of $[\![1, m]\!]$. We denote by $id_{[\![1,m]\!]}$ and $id_{\mathbb{R}^m}$ the identity functions on $[\![1, m]\!]$ and $\mathbb{R}^m$ respectively.

For any permutation $\varphi \in \mathfrak{S}_m$, we denote by P_φ the $m \times m$ permutation matrix associated to φ:

$$\forall i, j \in [\![1, m]\!], \quad (P_\varphi)_{i,j} = \begin{cases} 1 & \text{if } \varphi(j) = i \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

For all $x \in \mathbb{R}^m$, we have:

$$(P_\varphi x)_i = x_{\varphi^{-1}(i)}. \tag{10}$$

Using (10) we see that $P_{\varphi^{-1}} P_\varphi x = x$, which shows, since P_φ is orthogonal, that we have

$$P_\varphi^{-1} = P_{\varphi^{-1}} = P_\varphi^T. \tag{11}$$

Let $l, m, n \in \mathbb{N}^*$. For any matrix $M \in \mathbb{R}^{m\times l}$ and any function $f : \mathbb{R}^m \to \mathbb{R}^n$, we denote with a slight abuse of notation $f \circ M$ the function $x \mapsto f(Mx)$.

If X and Y are two sets and $h : X \to Y$ is a function, for a subset $A \subset Y$, we denote by $h^{-1}(A)$ the following set:

$$\{x \in X, h(x) \in A\}.$$

Note that this does not require the function h to be injective.

A.2 Continuous piecewise linear functions

We now introduce a few definitions and properties around the notion of continuous piecewise linear function.

Definition 10. Let $m \in \mathbb{N}^*$. A subset $D \subset \mathbb{R}^m$ is a closed polyhedron iif there exist $q \in \mathbb{N}^*$, $a_1, \ldots, a_q \in \mathbb{R}^m$ and $b_1, \ldots b_q \in \mathbb{R}$ such that for all $x \in \mathbb{R}^m$,

$$x \in D \iff \begin{cases} a_1^T x + b_1 \leq 0 \\ \vdots \\ a_q^T x + b_q \leq 0. \end{cases}$$

Remarks.

- A closed polyhedron is convex as an intersection of convex sets.
- Since we can fuse the inequation systems of several closed polyhedrons into one system, we see that an intersection of closed polyhedrons is a closed polyhedron.
- For $q = 1$ and $a_1 = 0$, taking $b_1 > 0$ and $b_1 \leq 0$ respectively we can show that $\emptyset$ and $\mathbb{R}^m$ are both closed polyhedra.

Proposition 11. *Let $m, l \in \mathbb{N}^*$. If $h : \mathbb{R}^l \to \mathbb{R}^m$ is linear and C is a closed polyhedron of $\mathbb{R}^m$, then $h^{-1}(C)$ is a closed polyhedron of $\mathbb{R}^l$.*

Proof. The function h is linear so there exist $M \in \mathbb{R}^{m\times l}$ and $b \in \mathbb{R}^m$ such that for all $x \in \mathbb{R}^l$,

$$h(x) = Mx + b.$$

The set C is a closed polyhedron so there exist $a_1, \ldots, a_q \in \mathbb{R}^m$ and $b_1, \ldots b_q \in \mathbb{R}$ such that $y \in C$ if and only if

$$\begin{cases} a_1^T y + b_1 \leq 0 \\ \vdots \\ a_q^T y + b_q \leq 0. \end{cases}$$

For all $x \in \mathbb{R}^l$,

$$\begin{aligned} x \in h^{-1}(C) \quad &\iff \quad h(x) \in C \\ &\iff \quad \begin{cases} a_1^T (Mx + b) + b_1 \leq 0 \\ \vdots \\ a_q^T (Mx + b) + b_q \leq 0 \end{cases} \\ &\iff \quad \begin{cases} (a_1^T M)x + (a_1^T b + b_1) \leq 0 \\ \vdots \\ (a_q^T M)x + (a_q^T b + b_q) \leq 0. \end{cases} \end{aligned}$$

This shows that $h^{-1}(C)$ is a closed polyhedron. □

Definition 12. We say that a function $g : \mathbb{R}^m \to \mathbb{R}^n$ is continuous piecewise linear if there exists a finite set of closed polyhedra whose union is $\mathbb{R}^m$ and such that g is linear over each polyhedron.

Example. Since $\mathbb{R}^m$ is a closed polyhedron, we see in particular that an affine function $x \mapsto Ax + b$, with $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$, is continuous piecewise linear from $\mathbb{R}^m$ to $\mathbb{R}^n$.

Example 13. The vectorial ReLU function $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ is continuous piecewise linear. Indeed, each of the 2^m closed orthants is a closed polyhedron, defined by a system of the form

$$\begin{cases} \epsilon_1 x_1 \geq 0 \\ \vdots \\ \epsilon_m x_m \geq 0, \end{cases}$$

with $\epsilon_i \in \{-1, 1\}$, and over such an orthant, the ReLU coincides with the affine function

$$(x_1, \ldots, x_m) \mapsto \left(\frac{1+\epsilon_1}{2} x_1, \ldots, \frac{1+\epsilon_m}{2} x_m \right).$$

In this definition the continuity is not obvious. We show it in the following proposition.

Proposition 14. *A continuous piecewise linear function is continuous.*

Proof. Let $g : \mathbb{R}^m \to \mathbb{R}^n$ be a continuous piecewise linear function. There exists a finite family of closed polyhedra $C_1, \ldots, C_r$ such that $\bigcup_{i=1}^r C_i = \mathbb{R}^m$ and g is linear on each closed polyhedron C_i.

Let $x \in \mathbb{R}^m$. Let $\epsilon > 0$.

Let us denote $I = \{i \in [\![1, n]\!],\ x \in C_i\}$. Since the polyhedrons are closed, there exists $r_0 > 0$ such that for all $i \notin I$, $B(x, r_0) \cap C_i = \emptyset$. We thus have

$$B(x, r_0) = \bigcup_{i=1}^{m} (B(x, r_0) \cap C_i) = \bigcup_{i \in I} (B(x, r_0) \cap C_i).$$

For all $i \in I$, g is linear -therefore continuous- on C_i so there exists $r_i > 0$, such that

$$y \in C_i \cap B(x, r_i) \ \Rightarrow\ \|g(y) - g(x)\| \leq \epsilon.$$

Let $r = \min(r_0, \min_{i \in I}(r_i))$. For all $y \in B(x, r)$ there exists $i \in I$ such that $y \in C_i$, and since $r \leq r_i$, we have

$$\|g(y) - g(x)\| \leq \epsilon.$$

Summarizing, for any $x \in \mathbb{R}^n$ and for any $\epsilon > 0$, there exists $r > 0$ such that

$$y \in B(x, r) \ \Rightarrow\ \|g(y) - g(x)\| \leq \epsilon.$$

This shows g is continuous. □

Proposition 15. *If $h : \mathbb{R}^l \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^n$ are two continuous piecewise linear functions, then $g \circ h$ is continuous piecewise linear.*

Proof. By definition there exist a family $C_1, \ldots, C_r$ of closed polyhedra of $\mathbb{R}^l$ such that $\bigcup_{i=1}^r C_i = \mathbb{R}^l$ and h is linear on each C_i and a family $D_1, \ldots, D_s$ of closed polyhedra of $\mathbb{R}^m$ such that $\bigcup_{i=1}^s D_i = \mathbb{R}^m$ and g is linear on each D_i. Let $i \in [\![1, r]\!]$ and $j \in [\![1, s]\!]$. The function h coincides with a linear map $\tilde{h} : \mathbb{R}^l \to \mathbb{R}^m$ on C_i and the inverse image of a closed polyhedron by a linear map is a closed polyhedron (Proposition 11) so $\tilde{h}^{-1}(D_j)$ is a closed polyhedron. Thus $h^{-1}(D_j) \cap C_i = \tilde{h}^{-1}(D_j) \cap C_i$ is a closed polyhedron as an intersection of closed polyhedra. The function h is linear on C_i and g is linear on D_j so $g \circ h$ is linear on $h^{-1}(D_j) \cap C_i$. We have a family of closed polyhedra,

$$\left(h^{-1}(D_j) \cap C_i\right)_{\substack{i \in [\![1, r]\!] \\ j \in [\![1, s]\!]}},$$

each of which $g \circ h$ is linear over. Given that

$$\bigcup_{i=1}^{r} \bigcup_{j=1}^{s} h^{-1}(D_j) \cap C_i = \bigcup_{i=1}^{r} C_i = \mathbb{R}^l,$$

we can conclude that $g \circ h$ is continuous piecewise linear. □

Definition 16. Let $g : \mathbb{R}^m \to \mathbb{R}^n$ be a continuous piecewise linear function. Let Π be a set of closed polyhedra of $\mathbb{R}^m$. We say that Π is admissible with respect to the function g if and only if:

- $\bigcup_{D\in\Pi} D = \mathbb{R}^m$,
- for all $D \in \Pi$, g is linear on D,
- for all $D \in \Pi$, $\mathring{D} \neq \emptyset$.

Proposition 17. *For all $g : \mathbb{R}^m \to \mathbb{R}^n$ continuous piecewise linear, there exists a set of closed polyhedra Π admissible with respect to g.*

Proof. Let $g : \mathbb{R}^m \to \mathbb{R}^n$ be a continuous piecewise linear function. By definition there exists a finite set of closed polyhedra $D_1, \dots, D_s$ such that $\bigcup_{i=1}^{s} D_i = \mathbb{R}^m$ and g is linear on each D_i.

Let $I = \{i \in [\![1, s]\!], \mathring{D}_i \neq \emptyset\}$. Let us show that $\bigcup_{i\in I} D_i = \mathbb{R}^m$.

We first show that if a polyhedron D_i has empty interior, then it is contained in an affine hyperplane. Indeed, if it is not contained in an affine hyperplane, then there exist $m + 1$ affinely independent points $x_1, \dots, x_{m+1} \in D_i$. Since a closed polyhedron is convex, the convex hull of the points $\mathrm{Conv}(x_1, \dots, x_{m+1})$, which is a m-simplex, is contained in D_i, and thus D_i has nonempty interior.

Let $x \in \mathbb{R}^m$. For all $i \notin I$, D_i is contained in an affine hyperplane, and a finite union of affine hyperplanes does not contain any nontrivial ball. As a consequence, for all $n \in \mathbb{N}$, the ball $B(x, \frac{1}{n})$ is not contained in $\bigcup_{i\notin I} D_i$ and thus there exists $i_n \in I$ such that $D_{i_n} \cap B(x, \frac{1}{n}) \neq \emptyset$. Since I is finite, there exists $i \in I$ such that $i_n = i$ for infinitely many n, and thus $x \in \overline{D_i}$.

We have shown that for all $x \in \mathbb{R}^m$ there exists $i \in I$ such that $x \in \overline{D_i} = D_i$, which means that

$$\bigcup_{i\in I} D_i = \mathbb{R}^m.$$

Hence, the set $\Pi := \{D_i, i \in I\}$ is admissible with respect to g. □

Proposition 18. *Let $h : \mathbb{R}^l \to \mathbb{R}^m$ be a continuous piecewise linear function and let $\mathcal{P}$ be a finite set of closed polyhedra of $\mathbb{R}^m$. Then*

- *for all $D \in \mathcal{P}$, $h^{-1}(D)$ is a finite union of closed polyhedra;*
- *$\bigcup_{D\in\mathcal{P}} \partial h^{-1}(D)$ is contained in a finite union of hyperplanes $\bigcup_{k=1}^{s} A_k$.*

Proof. Consider Π an admissible set of closed polyhedra with respect to h. Let $D \in \mathcal{P}$. Since $\bigcup_{C\in\Pi} C = \mathbb{R}^l$, we can write

$$h^{-1}(D) = h^{-1}(D) \cap \left(\bigcup_{C\in\Pi} C \right) = \bigcup_{C\in\Pi} \left(h^{-1}(D) \cap C \right).$$

For all $C \in \Pi$, h is linear over C, so $h^{-1}(D) \cap C$ is a polyhedron (see Proposition 11). This shows the first point of the proposition.

Since $h^{-1}(D) \cap C$ is a polyhedron, $\partial \left(h^{-1}(D) \cap C \right)$ is contained in a finite union of hyperplanes. In topology, we have

$$\partial \left[\bigcup_{C\in\Pi} \left(h^{-1}(D) \cap C \right) \right] \subset \bigcup_{C\in\Pi} \partial \left(h^{-1}(D) \cap C \right),$$

which shows that $\partial \left[\bigcup_{C\in\Pi} \left(h^{-1}(D) \cap C \right) \right]$ i.e. $\partial h^{-1}(D)$ is contained in a finite union of hyperplanes too. This is true for any $D \in \mathcal{P}$, and since $\mathcal{P}$ is finite, this is also true of the union $\bigcup_{D\in\mathcal{P}} \partial h^{-1}(D)$. □

A.3 Neural networks

We consider fully connected feedforward neural networks, with ReLU activation function. We index the layers in reverse order, from K to 0, for some $K \geq 2$. The input layer is the layer K, the output layer is the layer 0, and between them are $K - 1$ *hidden* layers. For $k \in [\![0, K]\!]$, we denote by $n_k \in \mathbb{N}$ the number of neurons of the layer k. This means the information contained at the layer k is a n_k-dimensional vector.

Let $k \in [\![0, K-1]\!]$. We denote the weights between the layer $k+1$ and the layer k with a matrix $M^k \in \mathbb{R}^{n_k \times n_{k+1}}$, and we consider a bias $b^k \in \mathbb{R}^{n_k}$ in the layer k. If $k \neq 0$, we add a ReLU activation function. If $x \in \mathbb{R}^{n_{k+1}}$ is the information contained at the layer $k+1$, the layer k contains:

$$\begin{cases} \sigma(M^k x + b^k) & \text{if } k \neq 0 \\ M^0 x + b^0 & \text{if } k = 0. \end{cases}$$

The parameters of the network can be summarized in the couple $(\mathbf{M}, \mathbf{b})$, where

$$\mathbf{M} = (M^0, M^1, \dots, M^{K-1}) \in \mathbb{R}^{n_0 \times n_1} \times \cdots \times \mathbb{R}^{n_{K-1} \times n_K}$$

and

$$\mathbf{b} = (b^0, b^1, \dots, b^{K-1}) \in \mathbb{R}^{n_0} \times \cdots \times \mathbb{R}^{n_{K-1}}.$$

We formalize the action of one layer of the network with the following definition.

Definition 19. For a network with parameters $(\mathbf{M}, \mathbf{b})$, we define the family of functions $(h_0, \dots, h_{K-1})$ such that for all $k \in [\![0, K-1]\!]$, $h_k : \mathbb{R}^{n_{k+1}} \to \mathbb{R}^{n_k}$ and for all $x \in \mathbb{R}^{n_k}$,

$$h_k(x) = \begin{cases} \sigma(M^k x + b^k) & \text{if } k \neq 0 \\ M^0 x + b^0 & \text{if } k = 0. \end{cases}$$

The function implemented by the network is then

$$f_{\mathbf{M},\mathbf{b}} = h_0 \circ h_1 \circ \cdots \circ h_{K-1} : \mathbb{R}^{n_K} \longrightarrow \mathbb{R}^{n_0}. \tag{12}$$

The network with its parameters are represented in Figure 1 in the main part.

For all $l \in [\![0, K-1]\!]$, we denote $\mathbf{M}^{\leq l} = (M^0, M^1, \dots, M^l)$ and $\mathbf{b}^{\leq l} = (b^0, b^1, \dots, b^l)$.

Remark 20. Since the vectorial ReLU function is continuous piecewise linear, Proposition 15 guarantees that the functions h_k are continuous piecewise linear.

We now define a few more functions associated to a network.

Definition 21. For a network with parameters $(\mathbf{M}, \mathbf{b})$, we define the family of functions $(h_0^{lin}, \dots, h_{K-1}^{lin})$ such that for all $k \in [\![0, K-1]\!]$, $h_k^{lin} : \mathbb{R}^{n_{k+1}} \to \mathbb{R}^{n_k}$ and for all $x \in \mathbb{R}^{n_{k+1}}$,

$$h_k^{lin}(x) = M^k x + b^k.$$

The functions h_k^{lin} correspond to the linear action of the network between two layers, before applying σ.

Definition 22. For a network with parameters $(\mathbf{M}, \mathbf{b})$, we define the family of functions $(f_K, f_{K-1}, \dots, f_0)$ as follows:

- $f_K = id_{\mathbb{R}^{n_K}}$,
- for all $k \in [\![0, K-1]\!]$, $\quad f_k = h_k \circ h_{k+1} \circ \cdots \circ h_{K-1}$.

Remark. In particular we have $f_0 = f_{\mathbf{M},\mathbf{b}}$.

The function $f_k : \mathbb{R}^{n_K} \mapsto \mathbb{R}^{n_k}$ represents the action of the network between the input layer and the layer k.

Definition 23. For a network with parameters $(\mathbf{M}, \mathbf{b})$, we define the sequence $(g_0, \dots, g_K)$ as follows:

- $g_0 = id_{\mathbb{R}^{n_0}}$,
- for all $k \in [\![1, K]\!]$, $\quad g_k = h_0 \circ h_1 \circ \cdots \circ h_{k-1}$.

Remark. We have in particular

- $g_K = f_{\mathbf{M},\mathbf{b}}$;
- for all $k \in [\![0, K]\!]$, $f_{\mathbf{M},\mathbf{b}} = g_k \circ f_k$.

The function $g_k : \mathbb{R}^{n_k} \mapsto \mathbb{R}^{n_0}$ represents the action of the network between the layer k and the output layer.

In this paper the functions implemented by the networks are considered on a subset $\Omega \subset \mathbb{R}^{n_K}$. The successive layers of a network project this subset onto the spaces $\mathbb{R}^{n_k}$, inducing a subset Ω_k of $\mathbb{R}^{n_k}$ for all k, as in the following definition.

Definition 24. For a network with parameters $(\mathbf{M}, \mathbf{b})$, for any $\Omega \subset \mathbb{R}^{n_K}$, we denote for all $k \in [\![0, K]\!]$,

$$\Omega_k = f_k(\Omega).$$

Definition 25. For a network with parameters $(\mathbf{M}, \mathbf{b})$, for all $k \in [\![2, K]\!]$, for all $i \in [\![1, n_{k-1}]\!]$, we define

$$H_i^k = \{x \in \mathbb{R}^{n_k},\ M_{i,.}^{k-1}x + b_i^{k-1} = 0\}.$$

Remark. When $M_{i,.}^{k-1} \neq 0$, the set H_i^k is a hyperplane.

Remark 26. The objects defined in Definitions 19, 21, 22, 23, 24 and 25 all depend on $(\mathbf{M}, \mathbf{b})$, but to simplify the notation we do not write it explicitly. To disambiguate when manipulating a second network, whose parameters we will denote by $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, we will denote by $\tilde{h}_k$, $\tilde{h}_k^{lin}$, $\tilde{f}_k$, $\tilde{g}_k$, $\tilde{\Omega}_k$ and $\tilde{H}_i^k$ the corresponding objects.

Proposition 27. *For all $k \in [\![0, K]\!]$, f_k and g_k are continuous piecewise linear.*

Proof. We show this by induction: for the initialisation we have $f_K = id_{\mathbb{R}^{n_K}}$ which is continuous piecewise linear. Now let $k \in [\![0, K-1]\!]$ and assume f_{k+1} is continuous piecewise linear. By definition, we have $f_k = h_k \circ f_{k+1}$. The function h_k is continuous piecewise linear as noted in Remark 20. By Proposition 15, the composition of two continuous piecewise linear functions is continuous piecewise linear, so f_k is continuous piecewise linear. The conclusion follows by induction.

We do the same for $(g_0, \dots, g_K)$ starting with g_0: first we have $g_0 = id_{\mathbb{R}^{n_0}}$ which is continuous piecewise linear, then for all $k \in [\![1, K]\!]$, we have $g_k = g_{k-1} \circ h_{k-1}$, and we conclude by composition of two continuous piecewise linear functions. □

Corollary 28. *The function $f_{\mathbf{M},\mathbf{b}}$ is continuous piecewise linear.*

Proof. It comes immediately from $f_{\mathbf{M},\mathbf{b}} = f_0$ and Proposition 27. □

Recall the definition of an admissible set with respect to a continuous piecewise linear function (Definition 16). Proposition 27 allows the following definition.

Definition 29. Consider a network parameterization $(\mathbf{M}, \mathbf{b})$, and the functions g_k associated to it. We say that a list of sets of closed polyhedra $\mathbf{\Pi} = (\Pi_1, \dots, \Pi_{K-1})$ is *admissible* with respect to $(\mathbf{M}, \mathbf{b})$ iif for all $k \in [\![1, K-1]\!]$, the set Π_k is admissible with respect to g_k.

Remark. For a list $\mathbf{\Pi} = (\Pi_1, \dots, \Pi_{K-1})$, for all $l \in [\![1, K-1]\!]$, we denote $\mathbf{\Pi}^{\leq l} = (\Pi_1, \dots, \Pi_l)$. If $\mathbf{\Pi}$ is admissible with respect to $(\mathbf{M}, \mathbf{b})$, then $\mathbf{\Pi}^{\leq l}$ is admissible with respect to $(\mathbf{M}^{\leq l}, \mathbf{b}^{\leq l})$.

Proposition 30. *For any network parameterization* $(\mathbf{M}, \mathbf{b})$*, there always exists a list of sets of closed polyhedra* $\mathbf{\Pi}$ *that is admissible with respect to* $(\mathbf{M}, \mathbf{b})$*.*

Proof. For all $k \in [\![1, K-1]\!]$, since g_k is continuous piecewise linear, Proposition 17 guarantees that there exists an admissible set of polyhedra Π_k with respect to g_k. We simply define $\mathbf{\Pi} = (\Pi_1, \dots, \Pi_{K-1})$. □

Definition 31. For a parameterization $(\mathbf{M}, \mathbf{b})$ and a list $\mathbf{\Pi}$ admissible with respect to $(\mathbf{M}, \mathbf{b})$, for all $k \in [\![1, K-1]\!]$, for all $D \in \Pi_k$, since g_k is linear over D and D has nonempty interior, we can define $V^k(D) \in \mathbb{R}^{n_0 \times n_k}$ and $c^k(D) \in \mathbb{R}^{n_0}$ as the unique couple that satisfies:

$$\forall x \in D, \quad g_k(x) = V^k(D)x + c^k(D).$$

We now introduce the equivalence relation between parameterizations, often referred to as *equivalence modulo permutation and positive rescaling.*

Definition 32 (Equivalent parameterizations)**.** If $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ are two network parameterizations, we say that $(\mathbf{M}, \mathbf{b})$ is *equivalent modulo permutation and positive rescaling*, or simply *equivalent*, to $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, and we write $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, if and only if there exist:

- a family of permutations $\boldsymbol{\varphi} = (\varphi_0, \dots, \varphi_K) \in \mathfrak{S}_{n_0} \times \cdots \times \mathfrak{S}_{n_K}$, with $\varphi_0 = id_{\llbracket 1, n_0 \rrbracket}$ and $\varphi_K = id_{\llbracket 1, n_K \rrbracket}$,
- a family of vectors $\boldsymbol{\lambda} = (\lambda^0, \lambda^1, \dots, \lambda^K) \in (\mathbb{R}_+^*)^{n_0} \times \cdots \times (\mathbb{R}_+^*)^{n_K}$, with $\lambda^0 = \mathbb{1}_{n_0}$ and $\lambda^K = \mathbb{1}_{n_K}$,

such that for all $k \in \llbracket 0, K-1 \rrbracket$,

$$\begin{cases} \tilde{M}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} \\ \tilde{b}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) b^k. \end{cases} \tag{13}$$

Remarks.

1. Recall that we denote by $\frac{1}{\lambda^{k+1}}$ the vector whose components are $\frac{1}{\lambda_i^{k+1}}$. Note that $\operatorname{Diag}(\lambda^{k+1})^{-1} = \operatorname{Diag}(\frac{1}{\lambda^{k+1}})$. Using (10), for all $k \in \llbracket 0, K-1 \rrbracket$, (13) means that for all $(i, j) \in \llbracket 1, n_k \rrbracket \times \llbracket 1, n_{k+1} \rrbracket$,
$$\tilde{M}_{i,j}^k = \frac{\lambda^k_{\varphi_k^{-1}(i)}}{\lambda^{k+1}_{\varphi_{k+1}^{-1}(j)}} M^k_{\varphi_k^{-1}(i), \varphi_{k+1}^{-1}(j)}$$
and
$$\tilde{b}_i^k = \lambda^k_{\varphi_k^{-1}(i)} b^k_{\varphi_k^{-1}(i)}.$$
2. We go from a parameterization to an equivalent one by:
   - permuting the neurons of each hidden layer k with a permutation φ_k;
   - for each hidden layer k, multiplying all the weights of the edges arriving (from the layer $k+1$) to the neuron j, as well as the bias b_j^k, by some positive number λ_j^k, and multiplying all the weights of the edges leaving (towards the layer $k-1$) the neuron j by $\frac{1}{\lambda_j^k}$.

Proposition 33. *The relation $\sim$ is an equivalence relation.*

Proof. Let us first show the following equality, that we are going to use in the proof. For any $n \in \mathbb{N}^*$, $\lambda \in \mathbb{R}^n$ and $\varphi \in \mathfrak{S}_n$,

$$\operatorname{Diag}(\lambda) P_\varphi = P_\varphi \operatorname{Diag}(P_\varphi^{-1} \lambda). \tag{14}$$

Indeed, $\operatorname{Diag}(\lambda) P_\varphi$ is the matrix obtained by multiplying each line i of P_φ by λ_i, so recalling (9), for all $i, j \in \llbracket 1, m \rrbracket$, we have

$$(\operatorname{Diag}(\lambda) P_\varphi)_{i,j} = \begin{cases} \lambda_i & \text{if } \varphi(j) = i \\ 0 & \text{otherwise.} \end{cases}$$

At the same time, $P_\varphi \operatorname{Diag}(P_\varphi^{-1} \lambda)$ is the matrix obtained by multiplying each column j of P_φ by $(P_\varphi^{-1} \lambda)_j = \lambda_{\varphi(j)}$ (see (10) and (11)), so for all $i, j \in \llbracket 1, m \rrbracket$, we have

$$(P_\varphi \operatorname{Diag}(P_\varphi^{-1} \lambda))_{i,j} = \begin{cases} \lambda_{\varphi(j)} & \text{if } \varphi(j) = i \\ 0 & \text{otherwise.} \end{cases}$$

The two matrices are clearly equal.

We can now show the proposition.

- To show reflexivity we can take $\lambda^k = \mathbb{1}_{n_k}$ and $\varphi_k = id_{\llbracket 1, n_k \rrbracket}$ for all $k \in \llbracket 0, K \rrbracket$.
- Let us show symmetry. Assume a parameterization $(\mathbf{M}, \mathbf{b})$ is equivalent to another parameterization $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$. Let us denote by $\boldsymbol{\varphi}$ and $\boldsymbol{\lambda}$ the corresponding families of permutations and vectors, as in Definition 32. Inverting the expression of $\tilde{M}^k$ in Definition 32 and using (14) twice, we have for all $k \in \llbracket 0, K-1 \rrbracket$:
$$\begin{aligned} & \tilde{M}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} \\ \iff \quad & \operatorname{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1} \tilde{M}^k P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) = M^k \\ \iff \quad & P_{\varphi_k}^{-1} \operatorname{Diag}(P_{\varphi_k} \lambda^k)^{-1} \tilde{M}^k \operatorname{Diag}(P_{\varphi_{k+1}} \lambda^{k+1}) P_{\varphi_{k+1}} = M^k, \end{aligned}$$

so denoting $\tilde{\varphi}_k = \varphi_k^{-1}$ and $\tilde{\lambda}^k = (P_{\varphi_k}\lambda^k)^{-1}$, and recalling that $P_{\varphi_k^{-1}} = P_{\varphi_k}^{-1}$, we have, for all $k \in [\![0, K-1]\!]$,

$$M^k = P_{\tilde{\varphi}_k} \operatorname{Diag}(\tilde{\lambda}^k) \tilde{M}^k \operatorname{Diag}(\tilde{\lambda}^{k+1})^{-1} P^{-1}_{\tilde{\varphi}_{k+1}}.$$

We show similarly that for all $k \in [\![0, K-1]\!]$,

$$b^k = P_{\tilde{\varphi}_k} \operatorname{Diag}(\tilde{\lambda}_k) \tilde{b}^k.$$

We naturally have $\tilde{\varphi}_0 = id_{[\![1,n_0]\!]}$ and $\tilde{\varphi}_K = id_{[\![1,n_K]\!]}$, as well as $\tilde{\lambda}^0 = \mathbb{1}_{n_0}$ and $\tilde{\lambda}^K = \mathbb{1}_{n_K}$. This proves the symmetry of the relation.

- Let us show transitivity. Assume $(\mathbf{M}, \mathbf{b})$, $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ and $(\check{\mathbf{M}}, \check{\mathbf{b}})$ are three parameterizations such that $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}) \sim (\check{\mathbf{M}}, \check{\mathbf{b}})$.

  As in Definition 32, we denote by $\boldsymbol{\varphi}$, $\tilde{\boldsymbol{\varphi}}$, $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$ the families of permutations and vectors such that, for all $k \in [\![0, K-1]\!]$,

$$\begin{cases} \tilde{M}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}} \\ \tilde{b}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) b^k, \end{cases}$$

  and

$$\begin{cases} \check{M}^k = P_{\tilde{\varphi}_k} \operatorname{Diag}(\tilde{\lambda}^k) \tilde{M}^k \operatorname{Diag}(\tilde{\lambda}^{k+1})^{-1} P^{-1}_{\tilde{\varphi}_{k+1}} \\ \check{b}^k = P_{\tilde{\varphi}_k} \operatorname{Diag}(\tilde{\lambda}^k) \tilde{b}^k. \end{cases}$$

  Combining these and using (14), we have

$$\begin{aligned} \check{M}^k &= P_{\tilde{\varphi}_k} \operatorname{Diag}(\tilde{\lambda}^k) P_{\varphi_k} \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}} \operatorname{Diag}(\tilde{\lambda}^{k+1})^{-1} P^{-1}_{\tilde{\varphi}_{k+1}} \\ &= P_{\tilde{\varphi}_k} \left(\operatorname{Diag}(\tilde{\lambda}^k) P_{\varphi_k} \right) \operatorname{Diag}(\lambda^k) M^k \\ &\qquad\qquad \cdot \operatorname{Diag}(\lambda^{k+1})^{-1} \left(\operatorname{Diag}(\tilde{\lambda}^{k+1}) P_{\varphi_{k+1}} \right)^{-1} P^{-1}_{\tilde{\varphi}_{k+1}} \\ &= P_{\tilde{\varphi}_k} \left(P_{\varphi_k} \operatorname{Diag}(P^{-1}_{\varphi_k} \tilde{\lambda}^k) \right) \operatorname{Diag}(\lambda^k) M^k \\ &\qquad\qquad \cdot \operatorname{Diag}(\lambda^{k+1})^{-1} \left(P_{\varphi_{k+1}} \operatorname{Diag}(P^{-1}_{\varphi_{k+1}} \tilde{\lambda}^{k+1}) \right)^{-1} P^{-1}_{\tilde{\varphi}_{k+1}} \\ &= P_{\tilde{\varphi}_k} P_{\varphi_k} \operatorname{Diag}(P^{-1}_{\varphi_k} \tilde{\lambda}^k) \operatorname{Diag}(\lambda^k) M^k \\ &\qquad\qquad \cdot \operatorname{Diag}(\lambda^{k+1})^{-1} \operatorname{Diag}(P^{-1}_{\varphi_{k+1}} \tilde{\lambda}^{k+1})^{-1} P^{-1}_{\varphi_{k+1}} P^{-1}_{\tilde{\varphi}_{k+1}}, \end{aligned}$$

  and

$$\begin{aligned} \check{b}^k &= P_{\tilde{\varphi}_k} \operatorname{Diag}(\tilde{\lambda}^k) P_{\varphi_k} \operatorname{Diag}(\lambda^k) b^k \\ &= P_{\tilde{\varphi}_k} P_{\varphi_k} \operatorname{Diag}(P^{-1}_{\varphi_k} \tilde{\lambda}^k) \operatorname{Diag}(\lambda^k) b^k. \end{aligned}$$

  Hence denoting $\check{\varphi}_k = \tilde{\varphi}_k \circ \varphi_k$ and $\check{\lambda}^k = \operatorname{Diag}(P^{-1}_{\varphi_k} \tilde{\lambda}^k) \lambda^k$, for all $k \in [\![0, K]\!]$, we see that, for $k \in [\![0, K-1]\!]$,

$$\check{M}^k = P_{\check{\varphi}_k} \operatorname{Diag}(\check{\lambda}^k) M^k \operatorname{Diag}(\check{\lambda}^{k+1})^{-1} P^{-1}_{\check{\varphi}_{k+1}}$$

  and

$$\check{b}^k = P_{\check{\varphi}_k} \operatorname{Diag}(\check{\lambda}^k) b^k.$$

  Naturally, we also have $\check{\varphi}_0 = id_{[\![1,n_0]\!]}$ and $\check{\varphi}_K = id_{[\![1,n_K]\!]}$, as well as $\check{\lambda}^0 = \mathbb{1}_{n_0}$ and $\check{\lambda}^K = \mathbb{1}_{n_K}$, which shows that $(\mathbf{M}, \mathbf{b}) \sim (\check{\mathbf{M}}, \check{\mathbf{b}})$.

□

Recall the objects $h_k, f_k, g_k, \Omega_k, H_i^k$ associated to a parameterization $(\mathbf{M}, \mathbf{b})$, defined in Definitions 19, 22, 23, 24 and 25, and recall that we denote by $\tilde{h}_k, \tilde{f}_k, \tilde{g}_k, \tilde{\Omega}_k$ and $\tilde{H}_i^k$ the corresponding objects with respect to another parameterization $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$. We give in the following proposition the relations that link these objects when the two parameterizations $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ are equivalent.

Proposition 34. *Assume* $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ *and consider* $\boldsymbol{\varphi}$ *and* $\boldsymbol{\lambda}$ *as in Definition 32. Let* $\mathbf{\Pi}$ *be a list of sets of closed polyhedra that is admissible with respect to* $(\mathbf{M}, \mathbf{b})$*. Then:*

1. *for all* $k \in [\![0, K-1]\!]$,

$$\tilde{h}_k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) \circ h_k \circ \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}},$$

2. *for all* $k \in [\![0, K]\!]$,

$$\begin{aligned} \tilde{f}_k &= P_{\varphi_k} \operatorname{Diag}(\lambda^k) \circ f_k, \\ \tilde{g}_k &= g_k \circ \operatorname{Diag}(\lambda^k)^{-1} P^{-1}_{\varphi_k}, \\ \tilde{\Omega}_k &= P_{\varphi_k} \operatorname{Diag}(\lambda^k) \Omega_k, \end{aligned} \tag{15}$$

3. *for all* $k \in [\![2, K]\!]$, *for all* $i \in [\![1, n_{k-1}]\!]$,

$$\tilde{H}^k_i = P_{\varphi_k} \operatorname{Diag}(\lambda^k) H^k_{\varphi^{-1}_{k-1}(i)},$$

4. *for all* $k \in [\![1, K-1]\!]$, *the set of closed polyhedra* $\tilde{\Pi}_k = \{P_{\varphi_k} \operatorname{Diag}(\lambda^k) D, D \in \Pi_k\}$ *is admissible for* $\tilde{g}_k$, *i.e. the list* $\tilde{\mathbf{\Pi}} = (\tilde{\Pi}_1, \ldots, \tilde{\Pi}_{K-1})$ *is admissible with respect to* $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$.

Proof. 1. Let $k \in [\![0, K-1]\!]$. If $k \neq 0$, we have from Definition 19:

$$\begin{aligned} \tilde{h}_k(x) &= \sigma(\tilde{M}^k x + \tilde{b}^k) \\ &= \sigma\left(P_{\varphi_k} \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}} x \right. \\ &\qquad \left. + P_{\varphi_k} \operatorname{Diag}(\lambda^k) b^k\right) \\ &= \sigma\left(P_{\varphi_k} \operatorname{Diag}(\lambda^k) \left[M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}} x + b^k\right]\right). \end{aligned}$$

Denote $y := \left[M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}} x + b^k\right]$. Let $i \in [\![1, n_k]\!]$. Using (10) and the fact that $\lambda^k_{\varphi^{-1}_k(i)}$ is nonnegative, the i^{th} coordinate of $\tilde{h}_k(x)$ is

$$\begin{aligned} \tilde{h}_k(x)_i = \left[\sigma\left(P_{\varphi_k} \operatorname{Diag}(\lambda^k) y\right)\right]_i &= \sigma\left(\left[P_{\varphi_k} \operatorname{Diag}(\lambda^k) y\right]_i\right) \\ &= \sigma\left(\lambda^k_{\varphi^{-1}_k(i)} y_{\varphi^{-1}_k(i)}\right) \\ &= \lambda^k_{\varphi^{-1}_k(i)} \sigma\left(y_{\varphi^{-1}_k(i)}\right) \\ &= \left[P_{\varphi_k} \operatorname{Diag}(\lambda^k) \sigma(y)\right]_i. \end{aligned}$$

Finally, we find the expression of $\tilde{h}_k(x)$:

$$\begin{aligned} \tilde{h}_k(x) &= P_{\varphi_k} \operatorname{Diag}(\lambda^k) \sigma(y) \\ &= P_{\varphi_k} \operatorname{Diag}(\lambda^k) \sigma\left(M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}} x + b^k\right) \\ &= P_{\varphi_k} \operatorname{Diag}(\lambda^k) h_k\left(\operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}}(x)\right). \end{aligned}$$

This concludes the proof when $k \neq 0$.

The case $k = 0$ is proven similarly but replacing the ReLU function σ by the identity.

2. - We prove by induction the expression of $\tilde{f}_k$.

   For $k = K$, we have $\tilde{f}_K = f_K = id_{\mathbb{R}^{n_K}}$, and since $P_{\varphi_K} = \operatorname{Id}_{n_K}$ and $\lambda^K = \mathbb{1}_{n_K}$ the equality $\tilde{f}_K = P_{\varphi_K} \operatorname{Diag}(\lambda^K) f_K$ holds.

   Now let $k \in [\![0, K-1]\!]$. Suppose the induction hypothesis is true for $\tilde{f}_{k+1}$. Using the expression of $\tilde{h}_k$ we just proved in 1 and the induction hypothesis, we have

$$\begin{aligned} \tilde{f}_k &= \tilde{h}_k \circ \tilde{f}_{k+1} \\ &= \left(P_{\varphi_k} \operatorname{Diag}(\lambda^k) \circ h_k \circ \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}}\right) \circ \left(P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) \circ f_{k+1}\right) \\ &= P_{\varphi_k} \operatorname{Diag}(\lambda^k) \circ h_k \circ f_{k+1} \\ &= P_{\varphi_k} \operatorname{Diag}(\lambda^k) \circ f_k. \end{aligned}$$

   This concludes the induction.

- We prove similarly the expression of $\tilde{g}_k$, but starting from $k = 0$: first we have $\tilde{g}_0 = g_0 = id_{\mathbb{R}^{n_0}}$, and then, for $k \in [\![0, K-1]\!]$, we write $\tilde{g}_{k+1} = \tilde{g}_k \circ \tilde{h}_k$ and we use the induction hypothesis and the expression of $\tilde{h}_k$.
- Using the relation (15), that we just proved, we obtain

$$\tilde{\Omega}_k = \tilde{f}_k(\Omega) = P_{\varphi_k} \operatorname{Diag}(\lambda^k) f_k(\Omega) = P_{\varphi_k} \operatorname{Diag}(\lambda^k)\Omega_k.$$

3. Let $k \in [\![2, K]\!]$ and $i \in [\![1, n_{k-1}]\!]$. For all $x \in \mathbb{R}^{n_k}$, using (13) and (10),

$$\begin{aligned}
x \in \tilde{H}^k_i \quad &\iff \quad \tilde{M}^{k-1}_{i,.} x + \tilde{b}^{k-1}_i = 0 \\
&\iff \quad \left[P_{\varphi_{k-1}} \operatorname{Diag}(\lambda^{k-1}) M^{k-1} \operatorname{Diag}(\lambda^k)^{-1} P^{-1}_{\varphi_k}\right]_{i,.} x \\
&\qquad\quad + \left[P_{\varphi_{k-1}} \operatorname{Diag}(\lambda^{k-1}) b^{k-1}\right]_i = 0 \\
&\iff \quad \lambda^{k-1}_{\varphi^{-1}_{k-1}(i)} M^{k-1}_{\varphi^{-1}_{k-1}(i),.} \operatorname{Diag}(\lambda^k)^{-1} P^{-1}_{\varphi_k} x + \lambda^{k-1}_{\varphi^{-1}_{k-1}(i)} b^{k-1}_{\varphi^{-1}_{k-1}(i)} = 0 \\
&\iff \quad \lambda^{k-1}_{\varphi^{-1}_{k-1}(i)} \left(M^{k-1}_{\varphi^{-1}_{k-1}(i),.} \operatorname{Diag}(\lambda^k)^{-1} P^{-1}_{\varphi_k} x + b^{k-1}_{\varphi^{-1}_{k-1}(i)} \right) = 0 \\
&\iff \quad M^{k-1}_{\varphi^{-1}_{k-1}(i),.} \operatorname{Diag}(\lambda^k)^{-1} P^{-1}_{\varphi_k} x + b^{k-1}_{\varphi^{-1}_{k-1}(i)} = 0 \\
&\iff \quad \operatorname{Diag}(\lambda^k)^{-1} P^{-1}_{\varphi_k} x \in H^k_{\varphi^{-1}_{k-1}(i)}.
\end{aligned}$$

Thus, $\tilde{H}^k_i = P_{\varphi_k} \operatorname{Diag}(\lambda^k) H^k_{\varphi^{-1}_{k-1}(i)}$.

4. For all $D \in \Pi_k$, denote $\tilde{D} = P_{\varphi_k} \operatorname{Diag}(\lambda^k) D$. We have $\tilde{\Pi}_k = \{\tilde{D}, D \in \Pi_k\}$.
Let $D \in \Pi_k$. The matrix $P_{\varphi_k} \operatorname{Diag}(\lambda^k)$ is invertible so, according to Proposition 11, $\tilde{D} = P_{\varphi_k} \operatorname{Diag}(\lambda^k) D$ is a closed polyhedron, and since $\mathring{D} \neq \emptyset$ we also have $\mathring{\tilde{D}} \neq \emptyset$.
Now recall from Item 2 that:

$$\tilde{g}_k = g_k \circ \operatorname{Diag}(\lambda^k)^{-1} P^{-1}_{\varphi_k}.$$

For all $x \in \tilde{D}$, we have $\operatorname{Diag}(\lambda^k)^{-1} P^{-1}_{\varphi_k} x \in D$. Since Π_k is admissible with respect to g_k (by definition of $\mathbf{\Pi}$), g_k is linear on D, and thus the function $\tilde{g}_k$ is linear on $\tilde{D}$.
Again, since Π_k is admissible with respect to g_k, we have $\bigcup_{D \in \Pi_k} D = \mathbb{R}^m$, and thus

$$\begin{aligned}
\bigcup_{\tilde{D} \in \tilde{\Pi}_k} \tilde{D} = \bigcup_{D \in \Pi_k} & P_{\varphi_k} \operatorname{Diag}(\lambda^k) D \\
&= P_{\varphi_k} \operatorname{Diag}(\lambda^k) \left(\bigcup_{D \in \Pi_k} D \right) \\
&= P_{\varphi_k} \operatorname{Diag}(\lambda^k) \left(\mathbb{R}^m\right) \\
&= \mathbb{R}^m,
\end{aligned}$$

which shows that $\tilde{\Pi}_k$ is admissible with respect to $\tilde{g}_k$.
This being true for any $k \in [\![1, K-1]\!]$, we conclude that $\tilde{\mathbf{\Pi}}$ is admissible with respect to $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$.

□

Corollary 35. *If* $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$*, then* $f_{\mathbf{M},\mathbf{b}} = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}$.

Proof. Consider $\boldsymbol{\varphi}$ and $\boldsymbol{\lambda}$ as in Definition 32. Looking at (15) for $k = 0$, and using the fact that $f_0 = f_{\mathbf{M},\mathbf{b}}$ and $\tilde{f}_0 = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}$, we obtain from Proposition 34

$$f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}} = P_{\varphi_0} \operatorname{Diag}(\lambda^0) f_{\mathbf{M},\mathbf{b}}.$$

By definition of $\boldsymbol{\varphi}$ and $\boldsymbol{\lambda}$, we have $P_{\varphi_0} = \operatorname{Id}_{n_0}$ and $\lambda^0 = \mathbb{1}_{n_0}$, so we can finally conclude:

$$f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}} = f_{\mathbf{M},\mathbf{b}}.$$

□

Definition 36. We say that $(\mathbf{M}, \mathbf{b})$ is normalized if for all $k \in [\![1, K-1]\!]$, for all $i \in [\![1, n_k]\!]$, we have:

$$\|M^k_{i,\cdot}\| = 1.$$

Proposition 37. *If $(\mathbf{M}, \mathbf{b})$ satisfies, for all $k \in [\![1, K-1]\!]$, for all $i \in [\![1, n_k]\!]$, $M^k_{i,\cdot} \neq 0$, then there exists an equivalent parameterization $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ that is normalized.*

Proof. We define recursively the family $(\lambda^0, \lambda^1, \ldots, \lambda^K) \in (\mathbb{R}^*_+)^{n_0} \times \cdots \times (\mathbb{R}^*_+)^{n_K}$ by:

- $\lambda^K = \mathbb{1}_{n_K}$;
- for all $k \in [\![1, K-1]\!]$, for all $i \in [\![1, n_k]\!]$,

$$\lambda^k_i = \frac{1}{\|M^k_{i,\cdot} \operatorname{Diag}(\lambda^{k+1})^{-1}\|};$$

- $\lambda^0 = \mathbb{1}_{n_0}$.

Consider the parameterization $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ defined by, for all $k \in [\![0, K-1]\!]$:

$$\begin{cases} \tilde{M}^k = \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} \\ \tilde{b}^k = \operatorname{Diag}(\lambda^k) b^k. \end{cases}$$

The parameterization is, by definition, equivalent to $(\mathbf{M}, \mathbf{b})$, and, for all $k \in [\![1, K-1]\!]$, for all $i \in [\![1, n_k]\!]$:

$$\begin{aligned} \|\tilde{M}^k_{i,\cdot}\| &= \left\| \left[\operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1}\right]_{i,\cdot} \right\| \\ &= \left\| \lambda^k_i M^k_{i,\cdot} \operatorname{Diag}(\lambda^{k+1})^{-1} \right\| \\ &= \left\| \frac{1}{\|M^k_{i,\cdot} \operatorname{Diag}(\lambda^{k+1})^{-1}\|} M^k_{i,\cdot} \operatorname{Diag}(\lambda^{k+1})^{-1} \right\| \\ &= 1. \end{aligned}$$

□

Proposition 38. *If $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ are both normalized, then they are equivalent if and only if there exists a family of permutations $(\varphi_0, \ldots, \varphi_K) \in \mathfrak{S}_{n_0} \times \cdots \times \mathfrak{S}_{n_K}$, with $\varphi_0 = id_{[\![1,n_0]\!]}$ and $\varphi_K = id_{[\![1,n_K]\!]}$, such that for all $k \in [\![0, K-1]\!]$:*

$$\begin{cases} \tilde{M}^k = P_{\varphi_k} M^k P^{-1}_{\varphi_{k+1}} \\ \tilde{b}^k = P_{\varphi_k} b^k. \end{cases} \tag{16}$$

Proof. Assume $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ are equivalent. Then there exist a family of permutations $(\varphi_0, \ldots, \varphi_K) \in \mathfrak{S}_{n_0} \times \cdots \times \mathfrak{S}_{n_K}$ and a family $(\lambda^0, \ldots, \lambda^K) \in (\mathbb{R}^*_+)^{n_0} \times \cdots \times (\mathbb{R}^*_+)^{n_K}$ as in Definition 32.

Let us prove by induction that $\lambda^k = \mathbb{1}_{n_k}$ for all $k \in [\![0, K]\!]$.

For $k = K$ it is true by Definition 32.

Let $k \in [\![1, K-1]\!]$, and suppose $\lambda^{k+1} = \mathbb{1}_{n_{k+1}}$. This means $\operatorname{Diag}(\lambda^{k+1}) = \operatorname{Id}_{n_{k+1}}$. Let $i \in [\![1, n_k]\!]$. Since $(\mathbf{M}, \mathbf{b})$ is normalized, $\|M^k_{i,\cdot}\| = 1$. Since $P^{-1}_{\varphi_{k+1}}$ is a permutation matrix, it is orthogonal so $\|M^k_{i,\cdot} P^{-1}_{\varphi_{k+1}}\| = \|M^k_{i,\cdot}\| = 1$. Recalling (13) and using the fact that $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ is normalized, that $\operatorname{Diag}(\lambda^{k+1}) = \operatorname{Id}_{n_{k+1}}$ and that λ^k_i is positive, we have:

$$\begin{aligned} 1 = \|\tilde{M}^k_{\varphi_k(i),\cdot}\| &= \|\lambda^k_i M^k_{i,\cdot} \operatorname{Diag}(\lambda^{k+1})^{-1} P^{-1}_{\varphi_{k+1}}\| \\ &= \lambda^k_i \|M^k_{i,\cdot} P^{-1}_{\varphi_{k+1}}\| \\ &= \lambda^k_i. \end{aligned}$$

This shows $\lambda^k = \mathbb{1}_{n_k}$.

The case $k = 0$ is also true by Definition 32.

Equation (13) with $\lambda^k = \mathbb{1}_{n_k}$ for all $k \in [\![0, K]\!]$ is precisely equation (16).

The reciprocal is clear: (16) is a particular case of (13) with $\lambda^k = \mathbb{1}_{n_k}$. □

B Main theorem

In Appendix B, we prove the main theorem using the notations and results of Appendix A, and admitting Lemma 48, which is proven in Appendix C.

More precisely, we begin by stating the conditions **C** and **P** in Section B.1, we then state our main result, which is Theorem 46, in Section B.2, and we give a consequence of this result in terms of risk minimization, which is Corollary 47, in Section B.3. Finally we prove Theorem 46 and Corollary 47 in Sections B.4 and B.5 respectively.

B.1 Conditions

Assume $g : \mathbb{R}^m \to \mathbb{R}^n$ is a continuous piecewise linear function, Π is a set of closed polyhedra admissible with respect to g, and let $\Omega \subset \mathbb{R}^l$, $M \in \mathbb{R}^{m\times l}$ and $b \in \mathbb{R}^m$.

We define

$$\begin{array}{rccc} h: & \mathbb{R}^l & \longrightarrow & \mathbb{R}^m \\ & x & \longmapsto & \sigma(Mx+b) \end{array}$$

and

$$\begin{array}{rccc} h^{lin}: & \mathbb{R}^l & \longrightarrow & \mathbb{R}^m \\ & x & \longmapsto & Mx+b. \end{array}$$

Definition 39. For all $i \in [\![1, m]\!]$, we denote $E_i = \{x \in \mathbb{R}^m,\ x_i = 0\}$.

Definition 40. Let $D \in \Pi$. The function g coincides with a linear function on D. Since the interior of D is nonempty, we define $V(D) \in \mathbb{R}^{n\times m}$ and $c(D) \in \mathbb{R}^n$ as the unique couple satisfying, for all $x \in D$:

$$g(x) = V(D)x + c(D).$$

Definition 41. We say that (g, M, b, Ω, Π) satisfies the conditions **C** iif:

C.a) M is full row rank;

C.b) for all $i \in [\![1, m]\!]$, there exists $x \in \mathring{\Omega}$ such that

$$M_{i,.}x + b_i = 0,$$

or equivalently,

$$E_i \cap h^{lin}(\mathring{\Omega}) \neq \emptyset;$$

C.c) for all $D \in \Pi$, for all $i \in [\![1, m]\!]$, if $E_i \cap D \cap h(\Omega) \neq \emptyset$ then $V_{.,i}(D) \neq 0$;

C.d) for any affine hyperplane $H \subset \mathbb{R}^l$,

$$H \cap \mathring{\Omega} \not\subset \bigcup_{D\in\Pi} \partial h^{-1}(D).$$

Definition 42. For all $k \in [\![1, K-1]\!]$, for all $i \in [\![1, n_k]\!]$, we denote $E_i^k = \{x \in \mathbb{R}^{n_k}, x_i = 0\}$.

We now state the conditions **P** (already stated in the main text in Definition 5).

Definition 43. We say that $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ satisfies the conditions **P** iif for all $k \in [\![1, K-1]\!]$, $(g_k, M^k, b^k, \Omega_{k+1}, \Pi_k)$ satisfies the conditions **C**.

Explicitly, for all $k \in [\![1, K-1]\!]$, the conditions are the following:

P.a) M^k is full row rank;

P.b) for all $i \in [\![1, n_k]\!]$, there exists $x \in \mathring{\Omega}_{k+1}$ such that

$$M^k_{i,.}x + b^k_i = 0,$$

or equivalently

$$E^k_i \cap h^{lin}_k(\mathring{\Omega}_{k+1}) \neq \emptyset;$$

P.c) for all $D \in \Pi_k$, for all $i \in [\![1, n_k]\!]$, if $E^k_i \cap D \cap \Omega_k \neq \emptyset$ then $V^k_{.,i}(D) \neq 0$;

$\mathbf{P}.d)$ for any affine hyperplane $H \subset \mathbb{R}^{n_{k+1}}$,

$$H \cap \mathring{\Omega}_{k+1} \not\subset \bigcup_{D \in \Pi_k} \partial h_k^{-1}(D).$$

Remark 44. The condition $\mathbf{P}.b)$ implies that for all $k \in \llbracket 1, K-1 \rrbracket$, $\mathring{\Omega}_{k+1} \neq \emptyset$, and in particular for $k = K-1$, the set $\Omega = \Omega_K$ has nonempty interior.

The following proposition shows that the conditions $\mathbf{P}$ are stable modulo permutation and positive rescaling, as defined in Definition 32.

Proposition 45. *Suppose* $(\mathbf{M}, \mathbf{b})$ *and* $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ *are two equivalent network parameterizations, and suppose* $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ *satisfies the conditions* $\mathbf{P}$. *Then, if we define* $\tilde{\mathbf{\Pi}}$ *as in Item 4 of Proposition 34,* $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$ *satisfies the conditions* $\mathbf{P}$.

Proof. Since $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ are equivalent, by Definition 32 there exist

- a family of permutations $(\varphi_0, \dots, \varphi_K) \in \mathfrak{S}_{n_0} \times \dots \times \mathfrak{S}_{n_K}$, with $\varphi_0 = id_{\llbracket 1, n_0 \rrbracket}$ and $\varphi_K = id_{\llbracket 1, n_K \rrbracket}$,
- a family $(\lambda^0, \lambda^1, \dots, \lambda^K) \in (\mathbb{R}_+^*)^{n_0} \times \dots \times (\mathbb{R}_+^*)^{n_K}$, with $\lambda^0 = \mathbb{1}_{n_0}$ and $\lambda^K = \mathbb{1}_{n_K}$,

such that

$$\begin{cases} \tilde{M}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} \\ \tilde{b}^k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) b^k. \end{cases} \tag{17}$$

Let $k \in \llbracket 1, K-1 \rrbracket$. We know the conditions $\mathbf{P}.a) - \mathbf{P}.d)$ are satisfied by $(g_k, M^k, b^k, \Omega_{k+1}, \Pi_k)$, let us show they are satisfied by $(\tilde{g}_k, \tilde{M}^k, \tilde{b}^k, \tilde{\Omega}_{k+1}, \tilde{\Pi}_k)$.

$\mathbf{P}.a)$ Since M^k satisfies $\mathbf{P}.a)$, it is full row rank, and using (17) and the fact that the matrices $P_{\varphi_k}, \operatorname{Diag}(\lambda^k), \operatorname{Diag}(\lambda^{k+1})^{-1}$ and $P_{\varphi_{k+1}}^{-1}$ are invertible, we see that $\tilde{M}^k$ is full row rank.

$\mathbf{P}.b)$ Let $i \in \llbracket 1, n_k \rrbracket$. Since $(g_k, M^k, b^k, \Omega_{k+1}, \Pi_k)$ satisfies the condition $\mathbf{P}.b)$, we can choose $x \in \mathring{\Omega}_{k+1}$ such that

$$M^k_{\varphi_k^{-1}(i),.} x + b^k_{\varphi_k^{-1}(i)} = 0. \tag{18}$$

Recall from Proposition 34 that

$$\tilde{\Omega}_{k+1} = P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) \Omega_{k+1}.$$

Since $P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1})$ is an invertible matrix, it induces an homeomorphism on $\mathbb{R}^{n_{k+1}}$, and thus this identity also holds for the interiors:

$$\mathring{\tilde{\Omega}}_{k+1} = P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) \mathring{\Omega}_{k+1}.$$

Given that $x \in \mathring{\Omega}_{k+1}$, defining $y = P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) x$, we have $y \in \mathring{\tilde{\Omega}}_{k+1}$.
Using (17), (10) and (18), we have

$$\begin{aligned} \tilde{M}^k_{i,.} y + \tilde{b}^k_i &= [P_{\varphi_k} \operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1}]_{i,.} y + [P_{\varphi_k} \operatorname{Diag}(\lambda^k) b^k]_i \\ &= [\operatorname{Diag}(\lambda^k) M^k \operatorname{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1}]_{\varphi_k^{-1}(i),.} y + [\operatorname{Diag}(\lambda^k) b^k]_{\varphi_k^{-1}(i)} \\ &= \lambda^k_{\varphi_k^{-1}(i)} M^k_{\varphi_k^{-1}(i),.} \operatorname{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} y + \lambda^k_{\varphi_k^{-1}(i)} b^k_{\varphi_k^{-1}(i)} \\ &= \lambda^k_{\varphi_k^{-1}(i)} M^k_{\varphi_k^{-1}(i),.} x + \lambda^k_{\varphi_k^{-1}(i)} b^k_{\varphi_k^{-1}(i)} \\ &= 0. \end{aligned}$$

We showed that there exists $y \in \mathring{\tilde{\Omega}}_{k+1}$ such that

$$\tilde{M}^k_{i,.} y + \tilde{b}^k_i = 0,$$

which concludes the proof of $\mathbf{P}.b)$.

P.*c*) Let $\tilde{D} \in \tilde{\Pi}_k$ and $i \in [\![1, n_k]\!]$. Suppose $E_i^k \cap \tilde{D} \cap \tilde{h}_k(\tilde{\Omega}_{k+1}) \neq \emptyset$, and let us show $\tilde{V}_{i,.}^k(\tilde{D}) \neq 0$.

Let $x \in \tilde{\Omega}_{k+1}$ such that $\tilde{h}_k(x) \in E_i^k \cap \tilde{D}$. Inverting the equalities of Proposition 34 we get

- $h_k = \mathrm{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1} \tilde{h}_k \circ P_{\varphi_{k+1}} \mathrm{Diag}(\lambda^{k+1})$,
- $H_{\varphi_k^{-1}(i)}^{k+1} = \mathrm{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} \tilde{H}_i^{k+1}$,
- $\Omega_{k+1} = \mathrm{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} \tilde{\Omega}_{k+1}$.

Denote $D = \mathrm{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1} \tilde{D}$. Since $\tilde{\Pi}_k$ has been defined as in Item 4 of Proposition 34, we know that $D \in \Pi_k$. Let $y = \mathrm{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} x$. Let us prove that $h_k(y) \in E_{\varphi_k(i)^{-1}}^k \cap D \cap h_k(\Omega_{k+1})$.

Since $x \in \tilde{\Omega}_{k+1}$, we see that $y \in \Omega_{k+1}$, so $h_k(y) \in h_k(\Omega_{k+1})$.

We also have

$$\begin{aligned} h_k(y) &= \mathrm{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1} \tilde{h}_k \circ P_{\varphi_{k+1}} \mathrm{Diag}(\lambda^{k+1}) \left(\mathrm{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} x \right) \\ &= \mathrm{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1} \tilde{h}_k\,(x)\,, \end{aligned}$$

which shows, since $\tilde{h}_k(x) \in \tilde{D}$, that $h_k(y) \in D$.

Since, by hypothesis, $\tilde{h}_k(x) \in E_i^k$, using (10) and (11), we have

$$\begin{aligned} [h_k(y)]_{\varphi_k^{-1}(i)} &= \left[\mathrm{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1} \tilde{h}_k\,(x) \right]_{\varphi_k^{-1}(i)} \\ &= \frac{1}{\lambda_{\varphi_k^{-1}(i)}^k} \left[P_{\varphi_k}^{-1} \tilde{h}_k(x) \right]_{\varphi_k^{-1}(i)} \\ &= \frac{1}{\lambda_{\varphi_k^{-1}(i)}^k} (\tilde{h}_k(x))_i \\ &= 0. \end{aligned}$$

This proves that $h_k(y) \in E_{\varphi_k^{-1}(i)}^k$.

We proved that

$$h_k(y) \in E_{\varphi_k^{-1}(i)}^k \cap D \cap h_k(\Omega_{k+1}),$$

which shows this intersection is not empty. Since $(g_k, M^k, b^k, \Omega_{k+1}, \Pi_k)$ satisfies **P**.*c*), we have $V_{.,\varphi_k^{-1}(i)}^k(D) \neq 0$.

Since, according to proposition 34,

$$\tilde{g}_k = g_k \circ \mathrm{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1},$$

we deduce:

$$\tilde{V}^k(\tilde{D}) = V^k(D)\,\mathrm{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1}. \tag{19}$$

For a matrix A and a permutation φ, we have $[P_\varphi A]_{i,.} = A_{\varphi^{-1}(i),.}$, so by taking the transpose, we see that $[A^T P_\varphi^{-1}]_{.,i} = (A^T)_{.,\varphi^{-1}(i)}$.

Taking the i^{th} column of (19), we thus obtain

$$\tilde{V}_{.,i}^k(\tilde{D}) = \left[V^k(D)\,\mathrm{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1} \right]_{.,i} = \frac{1}{\lambda_{\varphi_k^{-1}(i)}^k} V_{.,\varphi_k^{-1}(i)}^k(D),$$

which shows that $\tilde{V}_{.,i}^k(\tilde{D}) \neq 0$.

P.*d*) Let $\tilde{H} \subset \mathbb{R}^{n_{k+1}}$ be an affine hyperplane. Denote $H = \mathrm{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1} \tilde{H}$. Since **P**.*d*) holds for $(g_k, M^k, b^k, \Omega_{k+1}, \Pi_k)$, using Item 2 of Proposition 34, we have

$$\begin{aligned} \tilde{H} \cap \mathring{\tilde{\Omega}}_{k+1} &= P_{\varphi_{k+1}} \mathrm{Diag}(\lambda^{k+1}) \left(H \cap \mathring{\Omega}_{k+1} \right) \\ &\not\subset P_{\varphi_{k+1}} \mathrm{Diag}(\lambda^{k+1}) \bigcup_{D \in \Pi_k} \partial h_k^{-1}(D) \\ &= \bigcup_{D \in \Pi_k} P_{\varphi_{k+1}} \mathrm{Diag}(\lambda^{k+1}) \partial h_k^{-1}(D). \end{aligned} \tag{20}$$

For all k, $P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1})$ is an invertible matrix, so it induces an homeomorphism of $\mathbb{R}^{n_{k+1}}$. We thus have

$$P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) \partial h_k^{-1}(D) = \partial \left(P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) h_k^{-1}(D) \right). \tag{21}$$

Furthermore, by Item 1 of Proposition 34, we have $\tilde{h}_k = P_{\varphi_k} \operatorname{Diag}(\lambda^k) h_k \circ \operatorname{Diag}(\lambda^{k+1})^{-1} P_{\varphi_{k+1}}^{-1}$, so

$$\tilde{h}_k^{-1} = P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) h_k^{-1} \circ \operatorname{Diag}(\lambda^k)^{-1} P_{\varphi_k}^{-1},$$

and since $\tilde{D} = P_{\varphi_k} \operatorname{Diag}(\lambda^k) D$,

$$\tilde{h}_k^{-1}(\tilde{D}) = P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) h_k^{-1}(D). \tag{22}$$

Combining (21) and (22), we obtain

$$P_{\varphi_{k+1}} \operatorname{Diag}(\lambda^{k+1}) \partial h_k^{-1}(D) = \partial \tilde{h}_k^{-1}(\tilde{D}),$$

and we can thus reformulate (20) as

$$\tilde{H} \cap \mathring{\tilde{\Omega}}_{k+1} \quad \not\subset \quad \bigcup_{\tilde{D} \in \tilde{\Pi}_k} \partial \tilde{h}_k^{-1}(\tilde{D}).$$

□

B.2 Identifiability statement

We restate here the main theorem, already stated as Theorem 6 in the main part of the article.

Theorem 46. *Let $K \in \mathbb{N}$, $K \geq 2$. Suppose we are given two networks with K layers, identical number of neurons per layer, and with respective parameters $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$. Assume $\mathbf{\Pi}$ and $\tilde{\mathbf{\Pi}}$ are two lists of sets of closed polyhedra that are admissible with respect to $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ respectively. Denote by n_K the number of neurons of the input layer, and suppose we are given a set $\Omega \subset \mathbb{R}^{n_K}$ such that $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$ satisfy the conditions $\mathbf{P}$, and such that, for all $x \in \Omega$:*

$$f_{\mathbf{M},\mathbf{b}}(x) = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(x).$$

Then:

$$(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}}).$$

B.3 An application to risk minimization

We restate here the consequence of the main result in terms of minimization of the population risk, already stated as Corollary 7 in the main part.

Assume we are given a couple of input-output variables (X, Y) generated by a ground truth network with parameters $(\mathbf{M}, \mathbf{b})$:

$$Y = f_{\mathbf{M},\mathbf{b}}(X).$$

We can use Theorem 46 to show that the only way to bring the population risk to 0 is to find the ground truth parameters -modulo permutation and positive rescaling.

Indeed, let $\Omega \subset \mathbb{R}^{n_K}$ be a domain that is contained in the support of X, and suppose $L : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}_+$ is a loss function such that $L(y, y') = 0 \Rightarrow y = y'$.

Consider the population risk:

$$R(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}) = \mathbb{E}[L(f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(X), Y)].$$

We have the following result.

Corollary 47. *Suppose there exists a list of sets of closed polyhedra $\mathbf{\Pi}$ admissible with respect to $(\mathbf{M}, \mathbf{b})$ such that $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ satisfies the conditions $\mathbf{P}$.*

If $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ is also such that there exists a list of sets of closed polyhedra $\tilde{\mathbf{\Pi}}$ admissible with respect to $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ such that $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$ satisfies the conditions $\mathbf{P}$, and if $(\mathbf{M}, \mathbf{b}) \not\sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, then:

$$R(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}) > 0.$$

B.4 Proof of Theorem 46

To prove Theorem 46, we can assume the parameterizations $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ are normalized. Indeed, if they are not, by Proposition 37 there exist a normalized parameterization $(\mathbf{M'}, \mathbf{b'})$ equivalent to $(\mathbf{M}, \mathbf{b})$ and a normalized parameterization $(\tilde{\mathbf{M}}', \tilde{\mathbf{b}}')$ equivalent to $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$. Note that we can apply Proposition 37 because M^k and $\tilde{M}^k$ are full row rank (condition $\mathbf{P}.a)$) for all $k \in [\![1, K-1]\!]$ so their lines are always nonzero. We derive $\mathbf{\Pi'}$ from $\mathbf{\Pi}$ and $\tilde{\mathbf{\Pi}}'$ from $\tilde{\mathbf{\Pi}}$ as in Item 4 of Proposition 34. By Proposition 45, $(\mathbf{M'}, \mathbf{b'}, \Omega, \mathbf{\Pi'})$ and $(\tilde{\mathbf{M}}', \tilde{\mathbf{b}}', \Omega, \tilde{\mathbf{\Pi}}')$ also satisfy the conditions $\mathbf{P}$. By Corollary 35, $f_{\mathbf{M'},\mathbf{b'}} = f_{\mathbf{M},\mathbf{b}}$ and $f_{\tilde{\mathbf{M}}',\tilde{\mathbf{b}}'} = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}$, so we have, for all $x \in \Omega$:

$$f_{\mathbf{M'},\mathbf{b'}}(x) = f_{\tilde{\mathbf{M}}',\tilde{\mathbf{b}}'}(x).$$

$(\mathbf{M'}, \mathbf{b'}, \Omega, \mathbf{\Pi'})$ and $(\tilde{\mathbf{M}}', \tilde{\mathbf{b}}', \Omega, \tilde{\mathbf{\Pi}}')$ satisfy the hypotheses of Theorem 46. If we are able to show that $(\mathbf{M'}, \mathbf{b'}) \sim (\tilde{\mathbf{M}}', \tilde{\mathbf{b}}')$, then $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ follows immediately from the transitivity of the equivalence relation, proven in Proposition 33.

Thus in the proof $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ will be assumed to be normalized.

To prove the theorem, we need the following fundamental lemma (already stated as Lemma 9 in the main text), that is proven in Appendix C.

Lemma 48. *Let $l, m, n \in \mathbb{N}^*$. Suppose $g, \tilde{g} : \mathbb{R}^m \to \mathbb{R}^n$ are continuous piecewise linear functions, $\Omega \subset \mathbb{R}^l$ is a subset and let $M, \tilde{M} \in \mathbb{R}^{m \times l}$, $b, \tilde{b} \in \mathbb{R}^m$. Denote $h : x \mapsto \sigma(Mx+b)$ and $\tilde{h} : x \mapsto \sigma(\tilde{M}x+\tilde{b})$. Assume Π and $\tilde{\Pi}$ are two sets of polyhedra admissible with respect to g and $\tilde{g}$ respectively as in Definition 16.*

Suppose (g, M, b, Ω, Π) and $(\tilde{g}, \tilde{M}, \tilde{b}, \Omega, \tilde{\Pi})$ satisfy the conditions $\mathbf{C}$, and for all $i \in [\![1, m]\!]$, $\|M_{i,.}\| = \|\tilde{M}_{i,.}\| = 1$.

Suppose for all $x \in \Omega$:

$$g \circ h(x) = \tilde{g} \circ \tilde{h}(x).$$

Then, there exists a permutation $\varphi \in \mathfrak{S}_m$, such that:

- *$\tilde{M} = P_\varphi M$;*
- *$\tilde{b} = P_\varphi b$;*
- *g and $\tilde{g} \circ P_\varphi$ coincide on $h(\Omega)$.*

Proof of Theorem 46. We prove the theorem by induction on K.

Initialization. Assume here $K = 2$. We are going to apply Lemma 48. Since $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$ satisfy the conditions $\mathbf{P}$, by definition, $(g_1, M^1, b^1, \Omega_2, \Pi_1)$ and $(\tilde{g}_1, \tilde{M}^1, \tilde{b}^1, \Omega_2, \tilde{\Pi}_1)$ satisfy the conditions $\mathbf{C}$ (note that $\tilde{\Omega}_2 = \Omega_2 = \Omega$). The network is normalized, so we have, for all $i \in [\![1, n_1]\!]$,

$$\|M^1_{i,.}\| = \|\tilde{M}^1_{i,.}\| = 1.$$

By the assumptions of Theorem 46, for all $x \in \Omega$,

$$g_1 \circ h_1(x) = f_{\mathbf{M},\mathbf{b}}(x) = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(x) = \tilde{g}_1 \circ \tilde{h}_1(x).$$

We can thus apply Lemma 48, which shows that there exists a permutation $\varphi \in \mathfrak{S}_{n_1}$ such that

- $\tilde{M}^1 = P_\varphi M^1$;
- $\tilde{b}^1 = P_\varphi b^1$;
- g_1 and $\tilde{g}_1 \circ P_\varphi$ coincide on $h_1(\Omega)$.

Recall from Definition 25 that for all $i \in [\![1, n_1]\!]$, we denote $H^2_i = \{x \in \mathbb{R}^{n_2},\ M^1_{i,.}x + b^1_i = 0\}$. Let $(v_1, \dots, v_{n_1})$ be the canonical basis of $\mathbb{R}^{n_1}$. Let us show that for all $i \in [\![1, n_1]\!]$,

$$M^0 v_i = \tilde{M}^0 P_\varphi v_i.$$

Let $i \in [\![1, n_1]\!]$. By $\mathbf{P}.b)$, $H^2_i \cap \mathring{\Omega} \neq \emptyset$. Since M^1 is full row rank by $\mathbf{P}.a)$, none of the hyperplanes H^2_j, with $j \neq i$, is parallel to H^2_i. As a consequence, the intersections $H^2_i \cap H^2_j$ have Hausdorff dimension smaller than $n_2 - 2$, so there exists $x \in \mathring{\Omega} \cap H_i^2 \backslash \left(\bigcup_{j \neq i} H_j^2\right)$, and $\epsilon > 0$ such that $B(x, \epsilon) \cap H_j^2 = \emptyset$ for all $j \neq i$. Let u be a unit vector such that $M^1_{j,.}u = 0$ for all $j \neq i$ and $M^1_{i,.}u = \alpha > 0$ (this is possible again since M^1 is full row rank).

For all $j \in [\![1, n_1]\!] \backslash \{i\}$, we have

$$\sigma(M^1_{j,.}(x + \epsilon u) + b^1_j) - \sigma(M^1_{j,.}x + b^1_j) = \sigma(M^1_{j,.}x + b^1_j) - \sigma(M^1_{j,.}x + b^1_j) = 0.$$

At the same time, we have

$$\begin{aligned} \sigma(M^1_{i,.}(x + \epsilon u) + b^1_i) - \sigma(M^1_{i,.}x + b^1_i) &= M^1_{i,.}(x + \epsilon u) + b^1_i - M^1_{i,.}x + b^1_i \\ &= \epsilon M^1_{i,.}u \\ &= \epsilon\alpha. \end{aligned}$$

Summarizing,

$$\begin{aligned} h_1(x + \epsilon u) - h_1(x) &= \sigma(M^1(x + \epsilon u) + b^1) - \sigma(M^1 x + b^1) \\ &= \epsilon\alpha v_i. \end{aligned}$$

Let us denote $y_2 = h_1(x + \epsilon u) \in h_1(\Omega)$ and $y_1 = h_1(x) \in h_1(\Omega)$. We have shown $y_2 - y_1 = \epsilon\alpha v_i$, and since g_1 and $\tilde{g}_1 \circ P_\varphi$ coincide on $h_1(\Omega)$, we have

$$\begin{aligned} & g_1(y_2) - g_1(y_1) = \tilde{g}_1 \circ P_\varphi(y_2) - \tilde{g}_1 \circ P_\varphi(y_1) \\ \iff \ & M^0(y_2 - y_1) = \tilde{M}^0 P_\varphi(y_2 - y_1) \\ \iff \ & \epsilon\alpha M^0 v_i = \epsilon\alpha \tilde{M}^0 P_\varphi v_i \\ \iff \ & M^0 v_i = \tilde{M}^0 P_\varphi v_i. \end{aligned}$$

Since this last equality holds for any $i \in [\![1, n_1]\!]$, we conclude that

$$M^0 = \tilde{M}^0 P_\varphi,$$

and using one last time that g_1 and $\tilde{g}_1 \circ P_\varphi$ coincide on $h_1(\Omega)$, we obtain

$$b^0 = \tilde{b}^0,$$

i.e. we have shown

$$\begin{cases} \tilde{M}^0 = M^0 P_\varphi^{-1} \\ \tilde{b}^0 = b^0. \end{cases}$$

Defining $P_{\varphi_1} = P_\varphi$, $P_{\varphi_0} = \mathrm{Id}_{n_0}$ and $P_{\varphi_2} = \mathrm{Id}_{n_2}$, we can use Proposition 38 to conclude that

$$(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}}).$$

Induction step. Let $K \geq 3$ be an integer. Suppose Theorem 46 is true for all networks with $K - 1$ layers.

Consider two networks with parameters $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, with K layers and, for all $k \in [\![0, K]\!]$, same number n_k of neurons per layer. Let $\mathbf{\Pi}$ and $\tilde{\mathbf{\Pi}}$ be two list of sets of closed polyhedra that are admissible with respect to $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ respectively (Definition 29), and let $\Omega \subset \mathbb{R}^{n_K}$ such that $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$ satisfy the conditions $\mathbf{P}$ and $f_{\mathbf{M},\mathbf{b}}$ and $f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}$ coincide on Ω.

Recall the functions h_k and g_k associated to $(\mathbf{M}, \mathbf{b})$, defined in Definition 19 and Definition 23 respectively, and the corresponding functions $\tilde{h}_k$ and $\tilde{g}_k$ associated to $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$.

We have two matrices M^{K-1} and $\tilde{M}^{K-1} \in \mathbb{R}^{n_{K-1} \times n_K}$, two vectors b^{K-1} and $\tilde{b}^{K-1} \in \mathbb{R}^{n_{K-1}}$, two functions g_{K-1} and $\tilde{g}_{K-1} : \mathbb{R}^{n_{K-1}} \to \mathbb{R}^{n_0}$, two sets Π_{K-1} and $\tilde{\Pi}_{K-1}$ such that:

- $\forall x \in \Omega, \quad g_{K-1} \circ h_{K-1}(x) = g_K(x) = f_{\mathbf{M},\mathbf{b}}(x) = f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(x) = \tilde{g}_K(x) = \tilde{g}_{K-1} \circ \tilde{h}_{K-1}(x)$,
- g_{K-1} and $\tilde{g}_{K-1}$ are continuous piecewise linear, and Π_{K-1} and $\tilde{\Pi}_{K-1}$ are admissible with respect to g_{K-1} and $\tilde{g}_{K-1}$ respectively,
- $(g_{K-1}, M^{K-1}, b^{K-1}, \Omega, \Pi_{K-1})$ and $(\tilde{g}_{K-1}, \tilde{M}^{K-1}, \tilde{b}^{K-1}, \Omega, \tilde{\Pi}_{K-1})$ satisfy the conditions $\mathbf{C}$,

- $\forall i \in [\![1, n_{K-1}]\!], \quad \|M^{K-1}_{i,.}\| = \|\tilde{M}^{K-1}_{i,.}\| = 1.$

The third point comes from the fact that the conditions $\mathbf{P}$ hold for $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$, and the fourth point comes from the fact that $(\mathbf{M}, \mathbf{b})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$ are normalized.

Thus, the objects $g_{K-1}, \tilde{g}_{K-1}, M^{K-1}, b^{K-1}, \tilde{M}^{K-1}, \tilde{b}^{K-1}, \Pi_{K-1}$ and $\tilde{\Pi}_{K-1}$ satisfy the hypotheses of Lemma 48 and hence there exists $\varphi \in \mathfrak{S}_{n_{K-1}}$ such that

$$\begin{cases} \tilde{M}^{K-1} = P_\varphi M^{K-1}, \\ \tilde{b}^{K-1} = P_\varphi b^{K-1}, \end{cases} \tag{23}$$

and g_{K-1} and $\tilde{g}_{K-1} \circ P_\varphi$ coincide on Ω_{K-1}.

Let us denote $\mathbf{M}^* = (M^0, \dots, M^{K-3}, M^{K-2}P_\varphi^{-1})$. The functions $g_{K-1} \circ P_\varphi^{-1}$ and $\tilde{g}_{K-1}$ are implemented by two networks with $K-1$ layers, indexed from $K-1$ up to 0, with parameters $(\mathbf{M}^*, \mathbf{b}^{\leq K-2})$ and $(\tilde{\mathbf{M}}^{\leq K-2}, \tilde{\mathbf{b}}^{\leq K-2})$ respectively. The previous paragraph shows these functions coincide on $P_\varphi \Omega_{K-1}$. Recalling the definition of $\tilde{\Omega}_{K-1}$ and since, by (23), $\tilde{f}_{K-1} = \tilde{h}_{K-1} = P_\varphi h_{K-1}$, we have

$$\tilde{\Omega}_{K-1} = \tilde{f}_{K-1}(\Omega) = P_\varphi h_{K-1}(\Omega) = P_\varphi \Omega_{K-1},$$

i.e. the functions $g_{K-1} \circ P_\varphi^{-1} = f_{\mathbf{M^*, b^{\leq K-2}}}$ and $\tilde{g}_{K-1} = f_{\mathbf{\tilde{M}^{\leq K-2}, \tilde{b}^{\leq K-2}}}$ coincide on $\tilde{\Omega}_{K-1}$.

Since $(\mathbf{M}, \mathbf{b}, \Omega, \mathbf{\Pi})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\mathbf{\Pi}})$ satisfy the conditions $\mathbf{P}$, we see that $(g_k, M^k, b^k, \Omega_{k+1}, \Pi_k)$ and $(\tilde{g}_k, \tilde{M}^k, \tilde{b}^k, \tilde{\Omega}_{k+1}, \tilde{\Pi}_k)$ satisfy the conditions $\mathbf{C}$ for all $k \in [\![1, K-1]\!]$ so in particular these conditions are satisfied for $k \in [\![1, K-2]\!]$, so $(\mathbf{M}^{\leq K-2}, \mathbf{b}^{\leq K-2}, \Omega_{K-1}, \mathbf{\Pi}^{\leq K-2})$ and $(\tilde{\mathbf{M}}^{\leq K-2}, \tilde{\mathbf{b}}^{\leq K-2}, \tilde{\Omega}_{K-1}, \tilde{\mathbf{\Pi}}^{\leq K-2)})$ satisfy the conditions $\mathbf{P}$.

Let us verify that $(\mathbf{M}^*, \mathbf{b}^{\leq K-2}, \tilde{\Omega}_{K-1}, \mathbf{\Pi}^{\leq K-2})$ also satisfies the conditions $\mathbf{P}$. Indeed, the only thing that differs from $(\mathbf{M}^{\leq K-2}, \mathbf{b}^{\leq K-2}, \Omega_{K-1}, \mathbf{\Pi}^{\leq K-2})$ is $\tilde{\Omega}_{K-1}$ and the weights M^{*K-2} between the layer $K-1$ and the layer $K-2$. Writing that $M^{*K-2} = M^{K-2}P_\varphi^{-1}$, $h^*_{K-2} = h_{K-2} \circ P_\varphi^{-1}$, $\tilde{\Omega}_{K-1} = P_\varphi \Omega_{K-1}$ and $H_i^{*K-1} = P_\varphi H_i^{K-1}$, let us check that the conditions $\mathbf{C}$ also hold for $(g_{K-2}, M^{*K-2}, b^{K-2}, \tilde{\Omega}_{K-1}, \Pi_{K-2})$.

Indeed P_φ^{-1} is invertible, so M^{*K-2} is full row rank and $\mathbf{C}.a)$ holds.

If $x \in \mathring{\Omega}$ satisfies $M^{K-2}_{i,.}x + b^{K-2}_i = 0$, we define $h^{*lin}_{K-2}(x) = M^{*K-2}x + b^{K-2}$, we have $h^{*lin}_{K-2} = h^{lin}_{K-2} \circ P_\varphi^{-1}$, so

$$E_i \cap h^{*lin}_{K-2}(\mathring{\tilde{\Omega}}_{K-1}) = E_i \cap h^{lin}_{K-2}(\mathring{\Omega}_{K-1}) \neq \emptyset,$$

and $\mathbf{C}.b)$ is satisfied.

Similarly, the observation $h^*_{K-2}(\tilde{\Omega}_{K-1}) = h_{K-2}(\Omega_{K-1})$ yields $\mathbf{C}.c)$.

Finally, assume $H^* \subset \mathbb{R}^{n_{K-1}}$ is an affine hyperplane. Let $H = P_\varphi^{-1}H^*$. We have by hypothesis

$$H \cap \mathring{\Omega}_{K-1} \not\subset \bigcup_{D \in \Pi_{K-2}} \partial h^{-1}_{K-2}(D),$$

thus

$$\begin{aligned} H^* \cap \mathring{\tilde{\Omega}}_{K-1} &= P_\varphi \left(H \cap \mathring{\Omega}_{K-1} \right) \\ &\not\subset P_\varphi \bigcup_{D \in \Pi_{K-2}} \partial h^{-1}_{K-2}(D) \\ &= \bigcup_{D \in \Pi_{K-2}} \partial (P_\varphi h^{-1}_{K-2}(D)). \end{aligned}$$

For all $D \in \Pi_{K-2}$ we have

$$\begin{aligned} P_\varphi h^{-1}_{K-2}(D) &= P_\varphi \{y,\ h_{K-2}(y) \in D\} \\ &= P_\varphi \{P_\varphi^{-1}x,\ h_{K-2} \circ P_\varphi^{-1}(x) \in D\} \\ &= \{x,\ h^*_{K-2}(x) \in D\} \\ &= h^{*-1}_{K-2}(D). \end{aligned}$$

Therefore,

$$H^* \cap \mathring{\tilde{\Omega}}_{K-1} = \bigcup_{D \in \Pi_{K-2}} \partial h^{*-1}_{K-2}(D),$$

which proves $\mathbf{C}.d)$.

Since the rest stays unchanged, we can conclude.

The induction hypothesis can thus be applied to $(\mathbf{M}^*, \mathbf{b}^{\leq K-2}, \tilde{\Omega}_{K-1}, \mathbf{\Pi}^{\leq K-2})$ and $(\tilde{\mathbf{M}}^{\leq K-2}, \tilde{\mathbf{b}}^{\leq K-2}, \tilde{\Omega}_{K-1}, \tilde{\mathbf{\Pi}}^{\leq K-2})$, to obtain:

$$(\mathbf{M}^*, \mathbf{b}^{\leq K-2}) \sim (\tilde{\mathbf{M}}^{\leq K-2}, \tilde{\mathbf{b}}^{\leq K-2}).$$

Since we also have

$$\forall k \in [\![1, K-3]\!],\ \forall i \in [\![1, n_k]\!], \qquad \|M^{*k}_{i,\cdot}\| = \|M^k_{i,\cdot}\| = 1 \quad \text{and} \quad \|\tilde{M}^k_{i,\cdot}\| = 1,$$

$$\forall i \in [\![1, n_{K-2}]\!], \quad \|M^{*K-2}_{i,\cdot}\| = \|M^{K-2}_{i,\cdot} P_\varphi^{-1}\| = \|M^{K-2}_{i,\cdot}\| = 1 \quad \text{and} \quad \|\tilde{M}^{K-2}_{i,\cdot}\| = 1,$$

Proposition 38 shows that there exists a family of permutations $(\varphi_0, \dots, \varphi_{K-1}) \in \mathfrak{S}_{n_0} \times \cdots \times \mathfrak{S}_{n_{K-1}}$, with $\varphi_0 = id_{[\![1,n_0]\!]}$ and $\varphi_{K-1} = id_{[\![1,n_{K-1}]\!]}$, such that:

$$\forall k \in [\![0, K-3]\!], \quad \begin{cases} \tilde{M}^k = P_{\varphi_k} M^{*k} P^{-1}_{\varphi_{k+1}} = P_{\varphi_k} M^k P^{-1}_{\varphi_{k+1}} \\ \tilde{b}^k = P_{\varphi_k} b^k, \end{cases} \tag{24}$$

and:

$$\begin{cases} \tilde{M}^{K-2} = P_{\varphi_{K-2}} M^{*K-2} P^{-1}_{\varphi_{K-1}} = P_{\varphi_{K-2}} (M^{K-2} P_\varphi^{-1}) P^{-1}_{\varphi_{K-1}} = P_{\varphi_{K-2}} M^{K-2} P_\varphi^{-1} \\ \tilde{b}^{K-2} = P_{\varphi_{K-2}} b^{K-2}. \end{cases} \tag{25}$$

We can define $(\psi_0, \dots, \psi_K) \in \mathfrak{S}_{n_0} \times \cdots \times \mathfrak{S}_{n_K}$ by:

- $\psi_0 = id_{[\![1,n_0]\!]}$, $\psi_K = id_{[\![1,n_K]\!]}$;
- $\forall k \in [\![1, K-2]\!]$, $\psi_k = \varphi_k$;
- $\psi_{K-1} = \varphi$;

and using (24), (25) and (23) altogether, we then have, for all $k \in [\![0, K-1]\!]$:

$$\begin{cases} \tilde{M}^k = P_{\psi_k} M^k P^{-1}_{\psi_{k+1}} \\ \tilde{b}^k = P_{\psi_k} b^k. \end{cases}$$

It follows from Proposition 38 that $(\mathbf{M}, \mathbf{b}) \sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$.

□

B.5 Proof of Corollary 47

Theorem 47 is an immediate consequence of Theorem 46.

Since $(\mathbf{M}, \mathbf{b}, \Omega, \Pi)$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}, \Omega, \tilde{\Pi})$ satisfy the conditions $\mathbf{P}$ and $(\mathbf{M}, \mathbf{b}) \not\sim (\tilde{\mathbf{M}}, \tilde{\mathbf{b}})$, the contrapositive of Theorem 46 shows that there exists $x \in \Omega$ such that $f_{\mathbf{M},\mathbf{b}}(x) \neq f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(x)$. The function $f_{\mathbf{M},\mathbf{b}} - f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}$ is continuous so there exists $r > 0$ such that for all $u \in B(x, r)$, $f_{\mathbf{M},\mathbf{b}}(u) \neq f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(u)$ so $L(f_{\mathbf{M},\mathbf{b}}(u), f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(u)) > 0$. Since Ω is included in the support of X and $x \in \Omega$, denoting $\mathbb{P}_X$ the law of X we have $\mathbb{P}_X(B(x, r)) > 0$ and thus

$$\begin{aligned} R(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}) &= \mathbb{E}[L(f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(X), f_{\mathbf{M},\mathbf{b}}(X))] \\ &\geq \int_{B(x,r)} L(f_{\mathbf{M},\mathbf{b}}(u), f_{\tilde{\mathbf{M}},\tilde{\mathbf{b}}}(u)) \mathrm{d}\mathbb{P}_X(u) \\ &> 0. \end{aligned}$$

C Proof of Lemma 48

In this section we prove Lemma 48.

Let (g, M, b, Ω, Π) and $(\tilde{g}, \tilde{M}, \tilde{b}, \Omega, \tilde{\Pi})$ be as in the lemma. In particular, we assume they satisfy the conditions $\mathbf{C}$ all along Appendix C.

We denote, for all $x \in \mathbb{R}^l$:

$$f(x) = g\left(\sigma(Mx + b)\right).$$

Recall that, for all $x \in \mathbb{R}^l$, $h(x) = \sigma(Mx + b)$ and $\tilde{h}(x) = \sigma(\tilde{M}x + \tilde{b})$.

Recall that, as in Definition 25, we define for all $i \in [\![1, m]\!]$ the sets $H_i = \{x \in \mathbb{R}^l\ ,\ M_{i,.}x + b_i = 0\}$ and $\tilde{H}_i = \{x \in \mathbb{R}^l\ ,\ \tilde{M}_{i,.}x + \tilde{b}_i = 0\}$. By condition $C.a)$, for all $i \in [\![1, m]\!]$, $M_{i,.} \neq 0$ and $\tilde{M}_{i,.} \neq 0$ so H_i and $\tilde{H}_i$ are hyperplanes.

Recall that for all $D \in \Pi$, we define $V(D) \in \mathbb{R}^{n \times m}$ and $c(D) \in \mathbb{R}^n$ as in Definition 40, and similarly for all $\tilde{D} \in \tilde{\Pi}$, we define $\tilde{V}(\tilde{D}) \in \mathbb{R}^{n \times m}$ and $\tilde{c}(\tilde{D}) \in \mathbb{R}^n$ associated to $\tilde{g}$.

We now define $s : \mathbb{R}^l \to \{0, 1\}^m$ as follows:

$$\forall i \in [\![1, m]\!], \quad s_i(x) := \begin{cases} 1 & \text{if } M_{i,.}x + b_i \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

We define similarly $\tilde{s}$ for $(\tilde{M}, \tilde{b})$. We thus have, for all $i \in [\![1, m]\!]$,

$$\sigma(M_{i,.}x + b_i) = s_i(x)(M_{i,.}x + b_i)$$

and

$$\sigma(\tilde{M}_{i,.}x + \tilde{b}_i) = \tilde{s}_i(x)(\tilde{M}_{i,.}x + \tilde{b}_i).$$

Let $D \in \Pi$. For all $y \in D$, we have, by definition,

$$g(y) = V(D)y + c(D),$$

thus, for all $x \in h^{-1}(D)$,

$$\begin{aligned} f(x) &= V(D)h(x) + c(D) \\ &= V(D)\sigma(Mx + b) + c(D) \\ &= \sum_{k=1}^{m} V_{.,k}(D)s_k(x)(M_{k,.}x + b_k) + c(D). \end{aligned} \tag{27}$$

Similarly, for all $\tilde{D} \in \tilde{\Pi}$, for all $x \in \tilde{h}^{-1}(\tilde{D})$,

$$f(x) = \sum_{k=1}^{m} \tilde{V}_{.,k}(\tilde{D})\tilde{s}_k(x)(\tilde{M}_{k,.}x + \tilde{b}_k) + \tilde{c}(\tilde{D}). \tag{28}$$

Proposition 49. *Let $D \in \Pi$. For all $i \in [\![1, m]\!]$, for all $x \in H_i \cap \overbrace{h^{-1}(D)}^{\circ} \cap \mathring{\Omega} \backslash \left(\bigcup_{k \neq i} H_k\right)$, f is not differentiable at the point x.*

Proof. Let $i \in [\![1, m]\!]$ and suppose $x \in H_i \cap \overbrace{h^{-1}(D)}^{\circ} \cap \mathring{\Omega} \backslash \left(\bigcup_{k \neq i} H_k\right)$. Let us consider the function $t \mapsto f(x + tM_{i,.}{}^T)$. Since $x \in H_i$ and $\|M_{i,.}\| = 1$ by hypothesis,

$$M_{i,.}(x + tM_{i,.}{}^T) + b_i = tM_{i,.}M_{i,.}{}^T + M_{i,.}x + b_i = t\|M_{i,.}\|^2 = t. \tag{29}$$

Given the definition of s in (26), we thus have

$$s_i(x + tM_{i,.}{}^T) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{if } t < 0. \end{cases}$$

Since $x \in \overbrace{h^{-1}(D)}^{\circ}$ which is an open set, for t small enough we have $x + tM_{i,.}{}^T \in \overbrace{h^{-1}(D)}^{\circ}$ and thus, using (27) and (29),

$$\begin{aligned} f(x + tM_{i,.}{}^T) &= \sum_{k=1}^{m} V_{.,k}(D)s_k(x + tM_{i,.}{}^T)\left(M_{k,.}(x + tM_{i,.}{}^T) + b_k\right) + c(D) \\ &= \begin{cases} \sum_{k \neq i} V_{.,k}(D)s_k(x + tM_{i,.}{}^T)\left(M_{k,.}(x + tM_{i,.}{}^T) + b_k\right) & \\ +c(D) + tV_{.,i}(D) & \text{if } t \geq 0 \\ \sum_{k \neq i} V_{.,k}(D)s_k(x + tM_{i,.}{}^T)\left(M_{k,.}(x + tM_{i,.}{}^T) + b_k\right) & \\ +c(D) & \text{if } t < 0. \end{cases} \end{aligned}$$

Since x does not belong to any of the hyperplanes H_k for $k \neq i$, which are closed, there exists $\epsilon > 0$ such that for all $t \in \,]-\epsilon, \epsilon[$ and for all $k \neq i$, $x + tM_{i,.}{}^T \notin H_k$. Therefore, for all $t \in \,]-\epsilon, \epsilon[$, for all $k \in [\![1, m]\!] \backslash \{i\}$, $s_k(x + tM_{i,.}{}^T) = s_k(x)$ and

$$f(x + tM_{i,.}{}^T) = \begin{cases} \sum_{k \neq i} V_{.,k}(D) s_k(x)(M_{k,.}(x + tM_{i,.}{}^T) + b_k) + c(D) & \\ +tV_{.,i}(D) & \text{if } t \geq 0 \\ \sum_{k \neq i} V_{.,k}(D) s_k(x)(M_{k,.}(x + tM_{i,.}{}^T) + b_k) + c(D) & \text{if } t < 0. \end{cases}$$

The right derivative of $t \mapsto f(x + tM_{i,.}{}^T)$ at 0 is:

$$\sum_{k \neq i} V_{.,k}(D) s_k(x) M_{k,.} M_{i,.}{}^T + V_{.,i}(D).$$

The left derivative of $t \mapsto f(x + tM_{i,.}{}^T)$ at 0 is:

$$\sum_{k \neq i} V_{.,k}(D) s_k(x) M_{k,.} M_{i,.}{}^T.$$

Since $x \in H_i \cap h^{-1}(D) \cap \Omega$, we have $h(x) \in E_i \cap D \cap h(\Omega)$ so the condition $\mathbf{C}.c)$ implies that $V_{.,i}(D) \neq 0$. We conclude that the left and right derivatives at x do not coincide and thus f is not differentiable at x. □

Lemma 50. *Let $D \in \Pi$. For all $x \in \overbrace{h^{-1}(D)}^{\circ} \setminus (\bigcup_{i=1}^m H_i)$, there exists $r > 0$ such that f is differentiable on $B(x, r)$.*

Proof. Consider $x \in \overbrace{h^{-1}(D)}^{\circ} \setminus (\bigcup_{i=1}^m H_i)$. Since the hyperplanes H_i are closed, there exists a ball $B(x, r) \subset \overbrace{h^{-1}(D)}^{\circ}$ such that for all $i \in [\![1, m]\!]$, $B(x, r) \cap H_i = \emptyset$. As a consequence, for all $y \in B(x, r)$, $s(y) = s(x)$. Using (27) we get, for all $y \in B(x, r)$,

$$f(y) = \sum_{i=1}^m V_{.,i}(D) s_i(x) \left(M_{i,.} y + b_i \right) + c(D).$$

The right side of this equality is affine in the variable y, so f is differentiable on $B(x, r)$. □

Lemma 51. *Let $\gamma : \mathbb{R}^l \to \mathbb{R}^m$ be a continuous piecewise linear function. Let $\mathcal{P}$ be a finite set of polyhedra of $\mathbb{R}^m$ such that $\bigcup_{D \in \mathcal{P}} D = \mathbb{R}^m$. Let $A_1, \ldots A_s$ be a set of hyperplanes such that $\bigcup_{D \in \mathcal{P}} \partial \gamma^{-1}(D) \subset \bigcup_{k=1}^s A_k$ (Proposition 18 shows the existence of such hyperplanes). Let H be an affine hyperplane and $a \in \mathbb{R}^l, b \in \mathbb{R}$ such that $H = \{x \in \mathbb{R}^l, a^T x + b = 0\}$. Denote $I = \{k \in [\![1, s]\!], A_k = H\}$. Let $x \in H$ such that for all $k \in [\![1, s]\!] \backslash I$, $x \notin A_k$. Then there exists $r > 0$, D_- and $D_+ \in \mathcal{P}$ (not necessarily distinct) such that*

$$\begin{array}{rcl} B(x, r) \cap \{y \in \mathbb{R}^l, a^T y + b < 0\} & \subset & \gamma^{-1}(D_-) \\ B(x, r) \cap \{y \in \mathbb{R}^l, a^T y + b > 0\} & \subset & \gamma^{-1}(D_+). \end{array}$$

Proof. Let $r > 0$ such that

$$B(x, r) \cap \left(\bigcup_{k \notin I} A_k \right) = \emptyset.$$

$B(x, r) \backslash H$ has two connected components: $B_- = B(x, r) \cap \{y \in \mathbb{R}^l, a^T y + b < 0\}$ and $B_+ = B(x, r) \cap \{y \in \mathbb{R}^l, a^T y + b > 0\}$. The set B_- (resp. B_+) is convex as an intersection of two convex sets.

Since $\bigcup_{D \in \mathcal{P}} D = \mathbb{R}^m$, there exists $D_- \in \mathcal{P}$ such that $\gamma^{-1}(D_-) \cap B_- \neq \emptyset$. Let us show that

$$B_- \subset \gamma^{-1}(D_-).$$

Indeed, $B_- \cap \left(\bigcup_{k \notin I} A_k\right) = \emptyset$ and $B_- \cap H = \emptyset$ so $B_- \cap \left(\bigcup_{k \in I} A_k\right) = \emptyset$, therefore we have

$$B_- \cap \left(\bigcup_{D \in \mathcal{P}} \partial \gamma^{-1}(D) \right) \ \subset \ B_- \cap \left(\bigcup_{k=1}^{s} A_k \right) = \emptyset.$$

In particular, $B_- \cap \partial \gamma^{-1}(D_-) = \emptyset$. Let $Y = \gamma^{-1}(D_-) \cap B_-$. Let us denote by $\partial_{B_-} Y$ the topological boundary of Y with respect to the topology of B_-. Let us show the following inclusion:

$$\partial_{B_-} Y \ \subset \ \partial \gamma^{-1}(D_-) \cap B_-.$$

Indeed, let $y \in \partial_{B_-} Y$. By definition, there exist two sequences (u_n) and (v_n) such that $u_n \in Y$, $v_n \in B_- \backslash Y$, and both u_n and v_n tend to y. In particular, $u_n \in \gamma^{-1}(D_-)$ and $v_n \in \mathbb{R}^l \backslash \gamma^{-1}(D_-)$, so $y \in \partial \gamma^{-1}(D_-)$. Since $y \in B_-$, we have $y \in \partial \gamma^{-1}(D_-) \cap B_-$.

This shows $\partial_{B_-} Y = \emptyset$, and as a consequence Y is open and closed in B_-. Since B_- is connex and Y is not empty, we conclude that $Y = B_-$, i.e. $B_- \ \subset \ \gamma^{-1}(D_-)$.

We show similarly that there exists $D_+ \in \Pi$ such that $B_+ \subset \gamma^{-1}(D_+)$. □

Proposition 52. *There exists a bijection $\varphi \in \mathfrak{S}_m$ such that for all $i \in [\![1, m]\!]$, $\tilde{H}_i = H_{\varphi^{-1}(i)}$.*

Proof. We denote by X the set of all points of $\mathring{\Omega}$ at which f is not differentiable. We denote by $\mathcal{G}$ the set of all hyperplanes of $\mathbb{R}^l$. We denote $\mathcal{H} = \{H \in \mathcal{G} \ , \ H \cap \mathring{\Omega} \neq \emptyset \text{ and } H \cap \mathring{\Omega} \subset \overline{X}\}$. We want to show $\mathcal{H} = \{H_i, i \in [\![1, m]\!]\}$.

Indeed, once this established, since $\mathcal{H}$ only depends on Ω and f, we also have $\mathcal{H} = \{\tilde{H}_i, i \in [\![1, m]\!]\}$, and thus $\{H_i, i \in [\![1, m]\!]\} = \{\tilde{H}_i, i \in [\![1, m]\!]\}$. Since, using *C.a*), for all i, j, $i \neq j$, we have $H_i \neq H_j$ and $\tilde{H}_i \neq \tilde{H}_j$, we can conclude that there exists a permutation $\varphi \in \mathfrak{S}_m$ such that, for all $i \in [\![1, m]\!]$, $\tilde{H}_i = H_{\varphi^{-1}(i)}$.

– **Let us show $\mathcal{H} \subset \{H_i, i \in [\![1, m]\!]\}$.**

To begin, let us show that $\overline{X} \cap \mathring{\Omega} \ \subset \ \bigcup_{D \in \Pi} \partial h^{-1}(D) \cup \bigcup_{i=1}^m H_i$. Let $x \in \overline{X} \cap \mathring{\Omega}$. Let $D \in \Pi$ such that $h(x) \in D$. Since $x \in \overline{X}$, there does not exist any $r > 0$ such that f is differentiable on $B(x, r)$. The contrapositive of Lemma 50 shows that $x \notin \overbrace{h^{-1}(D)}^{\circ} \backslash \left(\bigcup_{i=1}^m H_i\right)$, so either $x \in \bigcup_{i=1}^m H_i$ or $x \notin \overbrace{h^{-1}(D)}^{\circ}$. In the latter case, since $x \in h^{-1}(D)$ by definition of D, we have $x \in h^{-1}(D) \backslash \overbrace{h^{-1}(D)}^{\circ} \subset \partial h^{-1}(D)$.

This shows:

$$\overline{X} \cap \mathring{\Omega} \quad \subset \quad \bigcup_{D \in \Pi} \partial h^{-1}(D) \cup \bigcup_{i=1}^{m} H_i. \tag{30}$$

Let $H \in \mathcal{H}$. We are going to show that there exists $i \in [\![1, m]\!]$ such that $H = H_i$.

We know by condition *C.d* that $H \cap \mathring{\Omega} \not\subset \bigcup_{D \in \Pi} \partial h^{-1}(D)$. Let $x \in (H \cap \mathring{\Omega}) \backslash \left(\bigcup_{D \in \Pi} \partial h^{-1}(D)\right)$. The set $\bigcup_{D \in \Pi} \partial h^{-1}(D)$ is closed, so there exists a ball

$$B(x, r) \subset \mathring{\Omega} \backslash \left(\bigcup_{D \in \Pi} \partial h^{-1}(D) \right). \tag{31}$$

By definition of $\mathcal{H}$,

$$H \cap \mathring{\Omega} \ \subset \ \overline{X} \cap \mathring{\Omega},$$

so using the fact that $B(x, r) \subset \mathring{\Omega}$ we have:

$$B(x, r) \cap H \ = \ B(x, r) \cap H \cap \mathring{\Omega} \ \subset \ B(x, r) \cap \overline{X} \cap \mathring{\Omega}.$$

Thus, using (30),

$$\begin{aligned} B(x, r) \cap H \ &\subset \ B(x, r) \cap \overline{X} \cap \mathring{\Omega} \\ &\subset \ B(x, r) \cap \left(\bigcup_{D \in \Pi} \partial h^{-1}(D) \cup \bigcup_{i=1}^{m} H_i \right) \\ &= \ \left(B(x, r) \cap \bigcup_{D \in \Pi} \partial h^{-1}(D) \right) \cup \left(B(x, r) \cap \bigcup_{i=1}^{m} H_i \right), \end{aligned}$$

and since by (31) the first set of the last equality is empty, we have

$$B(x,r) \cap H \ \subset \ B(x,r) \cap \bigcup_{i=1}^{m} H_i.$$

Therefore,

$$\begin{aligned} B(x,r) \cap H &= (B(x,r) \cap H) \cap \left(B(x,r) \cap \bigcup_{i=1}^{m} H_i \right) \\ &= B(x,r) \cap H \cap \bigcup_{i=1}^{m} H_i \\ &= B(x,r) \cap \bigcup_{i=1}^{m} (H \cap H_i). \end{aligned}$$

Assume, by contradiction, that for all $i \in \llbracket 1, m \rrbracket$ we have $H \neq H_i$. Then $H \cap H_i$ is an affine space of dimension less or equal to $l-2$ so it has Hausdorff dimension smaller or equal to $l-2$. A finite union of sets of Hausdorff dimension smaller or equal to $l-2$ has Hausdorff dimension smaller or equal to $l-2$. Thus, $B(x,r) \cap H = B(x,r) \cap \bigcup_{i=1}^{m} (H \cap H_i)$ has Hausdorff dimension smaller or equal to $l-2$, which is absurd since $x \in H$ so $B(x,r) \cap H$ has Hausdorff dimension $l-1$. Hence there exists $i \in \llbracket 1, m \rrbracket$ such that $H = H_i$.

We have shown

$$\mathcal{H} \subset \{H_i, i \in \llbracket 1, m \rrbracket\}. \tag{32}$$

– **Let us show $\{H_i, i \in \llbracket 1, m \rrbracket\} \subset \mathcal{H}$.**

Let $i \in \llbracket 1, m \rrbracket$. Let us prove $H_i \in \mathcal{H}$.

First, by condition $C.b)$ we know that $E_i \cap h^{lin}(\mathring{\Omega}) \neq \emptyset$, so there exists $x \in \mathring{\Omega}$ such that $h^{lin}(x) \in E_i$. Since $h^{lin}(x) = Mx + b$ and E_i is the space of vectors whose i^{th} coordinate is 0, this is equivalent to

$$M_{i,.}x + b_i = 0,$$

or said otherwise $x \in H_i$. This proves that $H_i \cap \mathring{\Omega} \neq \emptyset$. We still need to prove $H_i \cap \mathring{\Omega} \subset \overline{X}$.

Let $x \in H_i \cap \mathring{\Omega}$. Let us prove $x \in \overline{X}$.

Since M is full row rank, the line vectors $M_{1,.}, \dots, M_{m,.}$ are linearly independent, and thus for all $k \in \llbracket 1, m \rrbracket \backslash \{i\}$, $H_k \cap H_i$ has Hausdorff dimension smaller or equal to $l-2$.

Proposition 18 shows that $\bigcup_{D \in \Pi} \partial h^{-1}(D)$ is contained in a finite union of hyperplanes $\bigcup_{k=1}^{s} A_k$. Let $I = \{k \in \llbracket 1, s \rrbracket \ , \ A_k = H_i\}$. For all $k \in \llbracket 1, s \rrbracket \backslash I$, $A_k \cap H_i$ is either empty, or an intersection of two non parallel hyperplanes, in both cases it is an affine space of dimension smaller than $l-2$.

Thus,

$$H_i \cap \left((\bigcup_{k \neq i} H_k) \cup (\bigcup_{k \notin I} A_k) \right)$$

has Hausdorff dimension strictly smaller than $l-1$, so for any $r > 0$ there exists

$$y \in B(x,r) \cap H_i \cap \mathring{\Omega} \backslash \left((\bigcup_{k \neq i} H_k) \cup (\bigcup_{k \notin I} A_k) \right). \tag{33}$$

In the rest of the proof, we show that such a y is an element of X. Once this is established, since it is true for all $r > 0$, we conclude that $x \in \overline{X}$ and therefore $H_i \in \mathcal{H}$.

If there exists $D \in \Pi$ such that $y \in \overbrace{h^{-1}(D)}^{\circ}$, then

$$y \in H_i \cap \overbrace{h^{-1}(D)}^{\circ} \cap \mathring{\Omega} \backslash \left(\bigcup_{k \neq i} H_k \right)$$

therefore we can use Proposition 49 to conclude that f is not differentiable at y.

Otherwise we can use Lemma 51 to find $R_1 > 0$, D_- and $D_+ \in \Pi$ such that

$$\begin{aligned} B(y, R_1) \cap \{z \in \mathbb{R}^l, M_{i,.}z + b_i < 0\} &\subset h^{-1}(D_-) \\ B(y, R_1) \cap \{z \in \mathbb{R}^l, M_{i,.}z + b_i > 0\} &\subset h^{-1}(D_+). \end{aligned}$$

Since for all $j \neq i$, $y \notin H_j$ and since these hyperplanes are closed, there exists $R_2 > 0$ such that for all $j \neq i$, $B(y, R_2) \cap H_j = \emptyset$. Let $R = \min(R_1, R_2)$ and denote $B_- = B(y, R) \cap \{z \in \mathbb{R}^l, M_{i,.}z + b_i < 0\}$ and $B_+ = B(y, R) \cap \{z \in \mathbb{R}^l, M_{i,.}z + b_i > 0\}$.

For all $z \in B_-$, using (27) with the fact that $s_i(z) = 0$ and $s_k(z) = s_k(y)$ for all $k \neq i$, we have

$$f(z) = \sum_{k \neq i} V_{.,k}(D_-) s_k(y)(M_{k,.}z + b_k) + c(D_-). \tag{34}$$

For all $z \in B_+$, using this time that $s_i(z) = 1$, we have

$$f(z) = \sum_{k \neq i} V_{.,k}(D_+) s_k(y)(M_{k,.}z + b_k) + c(D_+) + V_{.,i}(D_+)(M_{i,.}z + b_i). \tag{35}$$

If f was differentiable at y, we would derive from (34) the expression of the Jacobian matrix

$$J_f(y) = \sum_{k \neq i} V_{.,k}(D_-) s_k(y) M_{k,.}, \tag{36}$$

but we would also derive from (35) the expression

$$J_f(y) = \sum_{k \neq i} V_{.,k}(D_+) s_k(y) M_{k,.} + V_{.,i}(D_+) M_{i,.}, \tag{37}$$

hence subtracting (36) to (37) we would find

$$\sum_{k \neq i} (V_{.,k}(D_+) - V_{.,k}(D_-)) s_k(y) M_{k,.} + V_{.,i}(D_+) M_{i,.} = 0.$$

Since M is full row rank, this would imply that $V_{.,i}(D_+) = 0$.

However since $h^{-1}(D_+)$ is closed and contains B_+, we have $y \in \overline{B_+} \subset h^{-1}(D_+)$. Recalling (33) we thus have

$$y \in H_i \cap h^{-1}(D_+) \cap \mathring{\Omega},$$

thus

$$h(y) \in E_i \cap D_+ \cap h(\mathring{\Omega}),$$

which shows the latter intersection is not empty. By assumption $C.c)$ this implies that $V_{.,i}(D_+) \neq 0$, which is a contradiction. Therefore f is not differentiable at y.

As a conclusion, we have showed that for all $r > 0$, there exists $y \in B(x, r)$ such that f is not differentiable at y and $y \in \mathring{\Omega}$. In other words, $x \in \overline{X}$.

Since x is arbitrary in $H_i \cap \mathring{\Omega}$, we have shown that for all $i \in [\![1, m]\!]$,

$$H_i \cap \mathring{\Omega} \subset \overline{X},$$

i.e., since we have already shown that $H_i \cap \mathring{\Omega} \neq \emptyset$,

$$H_i \in \mathcal{H}.$$

Finally $\{H_i, i \in [\![1, m]\!]\} \subset \mathcal{H}$, and, using (32),

$$\mathcal{H} = \{H_i, i \in [\![1, m]\!]\}.$$

□

Proposition 53. *For all $i \in [\![1, m]\!]$, there exists $\epsilon_{\varphi^{-1}(i)} \in \{-1, 1\}$ such that*

$$\tilde{M}_{i,.} = \epsilon_{\varphi^{-1}(i)} M_{\varphi^{-1}(i),.} \qquad \text{and} \qquad \tilde{b}_i = \epsilon_{\varphi^{-1}(i)} b_{\varphi^{-1}(i)}.$$

Proof. Let $i \in [\![1, m]\!]$. We know that $\tilde{H}_i = H_{\varphi^{-1}(i)}$, so the equations $\tilde{M}_{i,.}x + \tilde{b}_i = 0$ and $M_{\varphi^{-1}(i),.}x + b_{\varphi^{-1}(i)} = 0$ define the same hyperplanes. This is only possible if the parameters of the equation are proportional (but nonzero): there exists $\epsilon_{\varphi^{-1}(i)} \in \mathbb{R}^*$ such that $\tilde{M}_{i,.} = \epsilon_{\varphi^{-1}(i)} M_{\varphi^{-1}(i),.}$ and $\tilde{b}_i = \epsilon_{\varphi^{-1}(i)} b_{\varphi^{-1}(i)}$. But since $\|\tilde{M}_{i,.}\| = \|M_{\varphi^{-1}(i),.}\| = 1$ by hypothesis, we necessarily have $\epsilon_{\varphi^{-1}(i)} \in \{-1, 1\}$. □

Proposition 54. *For all $i \in [\![1, m]\!]$,*

- $\tilde{M}_{i,.} = M_{\varphi^{-1}(i),.}$*;*
- $\tilde{b}_i = b_{\varphi^{-1}(i)}$.

Proof. By Proposition 53, we know that there exists $(\epsilon_i)_{1 \le i \le m} \in \{-1, 1\}^m$ such that for all $i \in [\![1, m]\!]$,

$$\tilde{M}_{i,.} = \epsilon_{\varphi^{-1}(i)} M_{\varphi^{-1}(i),.} \qquad \text{and} \qquad \tilde{b}_i = \epsilon_{\varphi^{-1}(i)} b_{\varphi^{-1}(i)}. \tag{38}$$

We need to prove that for all $i \in [\![1, m]\!]$, $\epsilon_{\varphi^{-1}(i)} = 1$.

Let $i \in [\![1, m]\!]$.

Applying Proposition 18 to h and Π, we see that $\bigcup_{D \in \Pi} \partial h^{-1}(D)$ is contained in a finite union of hyperplanes $\bigcup_{k=1}^{s} A_k$. Applying it to $\tilde{h}$ and $\tilde{\Pi}$, we see similarly that $\bigcup_{\tilde{D} \in \tilde{\Pi}} \partial \tilde{h}^{-1}(\tilde{D})$ is contained in a finite union of hyperplanes $\bigcup_{k=1}^{r} B_k$.

Let $I = \{k \in [\![1, s]\!]\ ,\ A_k = H_i\}$ and $J = \{k \in [\![1, r]\!]\ ,\ B_k = H_i\}$. For all $k \in [\![1, s]\!] \backslash I$, since $A_k \neq H_i$, $A_k \cap H_i$ is either empty, or an intersection of two non parallel hyperplanes, in both cases it is an affine space of dimension smaller than $l - 2$. The same applies for all $k \in [\![1, r]\!] \backslash J$ to $B_k \cap H_i$. For all $j \neq i$, $H_j \neq H_i$ so $H_j \cap H_i$ is also an affine space of dimension smaller than $l - 2$. Since $H_i \cap \mathring{\Omega}$ is nonempty by **C**.b), we can thus find a vector

$$x \quad \in \quad \mathring{\Omega} \ \cap \ H_i \setminus \left((\bigcup_{k \notin I} A_k) \cup (\bigcup_{k \notin J} B_k) \cup (\bigcup_{j \neq i} H_j) \right).$$

Applying Lemma 51 with Π, h, H_i and $(M_{i,.}, b_i)$, we find $r_1 > 0$, D_- and $D_+ \in \Pi$ such that

$$\begin{aligned} B(x, r_1) \cap \{y \in \mathbb{R}^l, M_{i,.}y + b_i < 0\} &\quad \subset \quad h^{-1}(D_-) \\ B(x, r_1) \cap \{y \in \mathbb{R}^l, M_{i,.}y + b_i > 0\} &\quad \subset \quad h^{-1}(D_+). \end{aligned} \tag{39}$$

Applying the same lemma with $\tilde{\Pi}$, $\tilde{h}$, H_i and $(M_{i,.}, b_i)$ we find $r_2 > 0$, $\tilde{D}_-$ and $\tilde{D}_+ \in \tilde{\Pi}$ such that

$$\begin{aligned} B(x, r_2) \cap \{y \in \mathbb{R}^l, M_{i,.}y + b_i < 0\} &\quad \subset \quad \tilde{h}^{-1}(\tilde{D}_-) \\ B(x, r_2) \cap \{y \in \mathbb{R}^l, M_{i,.}y + b_i > 0\} &\quad \subset \quad \tilde{h}^{-1}(\tilde{D}_+). \end{aligned} \tag{40}$$

Since the hyperplanes H_j are closed, we can also find $r_3 > 0$ such that for all $j \neq i$, $B(x, r_3) \cap H_j = \emptyset$. Taking $r = \min(r_1, r_2, r_3)$ and denoting $B_+ = B(x, r) \cap \{y \in \mathbb{R}^l, M_{i,.}y + b_i > 0\}$, we derive from (39) and (40) that

$$B_+ \quad \subset \quad h^{-1}(D_+) \cap \tilde{h}^{-1}(\tilde{D}_+).$$

Since $r \le r_3$, we have $B_+ \cap \left(\bigcup_{j \neq i} H_j \right) = \emptyset$, and by definition $B_+ \cap \{y \in \mathbb{R}^l, M_{i,.}y + b_i = 0\} = \emptyset$, so $B_+ \cap H_i = \emptyset$. We have $B_+ \cap \left(\bigcup_{j=1}^{m} H_j \right) = \emptyset$, so for all $j \in [\![1, m]\!]$, there exist $\delta_j \in \{0, 1\}$ such that for all $y \in B_+$, $s_j(y) = \delta_j$. We have $\bigcup_{j=1}^{m} \tilde{H}_j = \bigcup_{j=1}^{m} H_j$ so similarly, $B_+ \cap \bigcup_{j=1}^{m} \tilde{H}_j = \emptyset$ and there exists $\tilde{\delta}_j \in \{0, 1\}$ such that for all $j \in [\![1, m]\!]$, for all $y \in B_+$, $\tilde{s}_j(y) = \tilde{\delta}_j$.

For all $y \in B_+$, we thus have, using (27),

$$\sum_{j=1}^{m} V_{.,j}(D_+) \delta_j \left(M_{j,.}y + b_j \right) + c(D_+) = \sum_{j=1}^{m} \tilde{V}_{.,j}(\tilde{D}_+) \tilde{\delta}_j \left(\tilde{M}_{j,.}y + \tilde{b}_j \right) + \tilde{c}(\tilde{D}_+).$$

B_+ is a nonempty open set so we have the equality

$$\begin{aligned}\sum_{j=1}^{m} V_{.,j}(D_+)\delta_j M_{j,.} &= \sum_{j=1}^{m} \tilde{V}_{.,j}(\tilde{D}_+)\tilde{\delta}_j \tilde{M}_{j,.} \\ &= \sum_{j=1}^{m} \tilde{V}_{.,j}(\tilde{D}_+)\tilde{\delta}_j \epsilon_{\varphi^{-1}(j)} M_{\varphi^{-1}(j),.} \\ &= \sum_{j=1}^{m} \tilde{V}_{.,\varphi(j)}(\tilde{D}_+)\tilde{\delta}_{\varphi(j)} \epsilon_j M_{j,.}. \qquad (41)\end{aligned}$$

The condition $\mathbf{C}.a)$ states that M is full row rank, so the vectors $M_{j,.}$ are linearly independent. Applied to (41), this information yields, for all $j \in [\![1, m]\!]$,

$$V_{.,j}(D_+)\delta_j = \tilde{V}_{.,\varphi(j)}(\tilde{D}_+)\tilde{\delta}_{\varphi(j)}\epsilon_j,$$

and in particular,

$$V_{.,i}(D_+)\delta_i = \tilde{V}_{.,\varphi(i)}(\tilde{D}_+)\tilde{\delta}_{\varphi(i)}\epsilon_i. \qquad (42)$$

Since $h^{-1}(D_+)$ and $\tilde{h}^{-1}(\tilde{D}_+)$ are closed, we have

$$\overline{B_+} \subset h^{-1}(D_+) \cap \tilde{h}^{-1}(\tilde{D}_+),$$

and since $x \in \overline{B_+}$, we have $h^{-1}(D_+) \cap H_i \neq \emptyset$ and $\tilde{h}^{-1}(\tilde{D}_+) \cap H_i \neq \emptyset$. The condition $C.c)$ implies that $V_{.,i}(D_+) \neq 0$ and $\tilde{V}_{.,\varphi(i)}(\tilde{D}_+) \neq 0$ (recall that $H_i = \tilde{H}_{\varphi(i)}$). We also have $\epsilon_i \neq 0$, so from (42) we obtain

$$\delta_i = 0 \Leftrightarrow \tilde{\delta}_{\varphi(i)} = 0.$$

By definition, the coefficient δ_i depends on the sign of $M_{i,.}y + b_i$: if $M_{i,.}y + b_i$ is positive, $\delta_i = 1$ and if $M_{i,.}y + b_i$ is negative then $\delta_i = 0$ ($M_{i,.}y + b_i$ cannot be equal to zero since $y \notin H_i$). The coefficient $\tilde{\delta}_{\varphi(i)}$ depends similarly on the sign of $\tilde{M}_{\varphi(i),.}y + \tilde{b}_{\varphi(i)}$. Thus, $M_{i,.}y + b_i$ and $\tilde{M}_{\varphi(i),.}y + \tilde{b}_{\varphi(i)}$ have same sign.

Since $\epsilon_i \in \{-1, 1\}$ and

$$\tilde{M}_{\varphi(i),.}y + \tilde{b}_{\varphi(i)} = \epsilon_i M_{i,.}y + \epsilon_i b_i = \epsilon_i \left(M_{i,.}y + b_i\right),$$

we conclude that $\epsilon_i = 1$.

□

We can now finish the proof of Lemma 48. It results from the above that:

$$\tilde{M} = P_\varphi M$$

$$\tilde{b} = P_\varphi b.$$

We have by hypothesis, for all $x \in \Omega$,

$$\tilde{g}(\sigma(\tilde{M}x + \tilde{b})) = g(\sigma(Mx + b)),$$

but since $\tilde{M} = P_\varphi M$ and $\tilde{b} = P_\varphi b$ we also have:

$$\tilde{g}(\sigma(\tilde{M}x + \tilde{b})) = \tilde{g}(\sigma(P_\varphi Mx + P_\varphi b)) = \tilde{g}(P_\varphi \sigma(Mx + b)).$$

Combining these, we have for all $x \in \Omega$,

$$\tilde{g} \circ P_\varphi(h(x)) = g(h(x)),$$

i.e. $\tilde{g} \circ P_\varphi$ and g coincide on $h(\Omega)$.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Rilwan A Adewoyin, Peter Dueben, Peter Watson, Yulan He, and Ritabrata Dutta. Tru-net: a deep learning approach to high resolution prediction of rainfall. *Machine Learning*, 110(8):2035–2062, 2021.

[3] Francesca Albertini, Eduardo D Sontag, and Vincent Maillot. Uniqueness of weights for neural networks. *Artificial Neural Networks for Speech and Vision*, pages 115–125, 1993.

[4] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 584–592, Bejing, China, 22–24 Jun 2014. PMLR.

[5] Shumeet Baluja and Ian Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *arXiv preprint arXiv:1703.09387*, 2017.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[7] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a ConvNet with Gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 605–614, 2017.

[8] Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. Cryptanalytic extraction of neural network models. In *Annual International Cryptology Conference*, pages 189–218. Springer, 2020.

[9] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 267–284, 2019.

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[11] Hervé Chabanne, Vincent Despiegel, and Linda Guiga. A protection against the extraction of neural network models. *arXiv preprint arXiv:2005.12782*, 2020.

[12] Jiyu Chen, Yiwen Guo, Qianjun Zheng, and Hao Chen. Protect privacy of deep classification networks by exploiting their generative power. *Machine Learning*, 110(4):651–674, 2021.

[13] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

[14] Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? In *International Conference on Learning Representations*, 2018.

[15] Dennis Maximilian Elbrächter, Julius Berner, and Philipp Grohs. How degenerate is the parametrization of neural networks with the ReLU activation function? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[16] Charles Fefferman. Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 10(3):507–555, 1994.

[17] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[18] Haoyu Fu, Yuejie Chi, and Yingbin Liang. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE Transactions on Signal Processing*, 68:3225–3235, 2020.

[19] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[20] Surbhi Goel, Adam Klivans, and Raghu Meka. Learning one convolutional layer with overlapping patches. In *International Conference on Machine Learning*, pages 1783–1791, 2018.

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

[22] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[23] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[24] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[25] Paul C Kainen, Věra Kůrková, Vladik Kreinovich, and Ongard Sirisaengtaksin. Uniqueness of network parametrization and faster learning. *Neural, Parallel & Scientific Computations*, 2(4):459–466, 1994.

[26] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[28] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *International Conference on Learning Representations*, 2017.

[29] Věra Kůrková and Paul C Kainen. Functionally equivalent feedforward neural networks. *Neural Computation*, 6(3):543–558, 1994.

[30] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond NTK. In *Conference on Learning Theory*, pages 2613–2682. PMLR, 2020.

[31] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in neural information processing systems*, pages 597–607, 2017.

[32] Grigorios Loukides, Joshua C Denny, and Bradley Malin. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*, 17(3):322–327, 2010.

[33] François Malgouyres and Joseph Landsberg. On the identifiability and stable recovery of deep/multi-layer structured matrix factorization. In *IEEE, Info. Theory Workshop*, Sept. 2016.

[34] François Malgouyres and Joseph Landsberg. Multilinear compressive sensing and an application to convolutional linear networks. *SIAM Journal on Mathematics of Data Science*, 1(3):446–475, 2019.

[35] Francois Malgouyres. On the stable recovery of deep structured linear networks under sparsity constraints. In *Mathematical and Scientific Machine Learning*, pages 107–127. PMLR, 2020.

[36] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[37] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048, 2010.

[38] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[39] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[40] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[41] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[42] Philipp Petersen, Mones Raslan, and Felix Voigtlaender. Topological properties of the set of functions generated by neural networks of fixed size. *Foundations of Computational Mathematics*, 21:375–444, 2021.

[43] Henning Petzka, Martin Trimmel, and Cristian Sminchisescu. Notes on the symmetries of 2-layer ReLU-networks. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 1, pages 6–6, 2020.

[44] Mary Phuong and Christoph H. Lampert. Functional vs. parametric equivalence of ReLU networks. In *International Conference on Learning Representations*, 2020.

[45] José Pedro Pinto, André Pimenta, and Paulo Novais. Deep learning and multivariate time series for cheat detection in video games. *Machine Learning*, 110(11):3037–3057, 2021.

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[47] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[48] David Rolnick and Konrad Kording. Reverse-engineering deep ReLU networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8178–8187, 13–18 Jul 2020.

[49] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[50] Sayantan Sarkar, Ankan Bansal, Upal Mahbub, and Rama Chellappa. UPSET and ANGRI: Breaking high performance image classifiers. *arXiv preprint arXiv:1707.01159*, 2017.

[51] Hanie Sedghi and Anima Anandkumar. Provable methods for training neural networks with sparse connectivity. In *Deep Learning and representation learning workshop: NIPS*, 2014.

[52] Pierre Stock. *Efficiency and Redundancy in Deep Learning Models : Theoretical Considerations and Practical Applications*. PhD thesis, Université de Lyon, April 2021.

[53] Pierre Stock and Rémi Gribonval. An embedding of ReLU networks and an analysis of their identifiability. *arXiv preprint arXiv:2107.09370, forthcoming in Constructive Approximation*, 2021.

[54] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[55] Héctor J Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural networks*, 5(4):589–593, 1992.

[56] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[57] Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Guaranteed convergence of training convolutional neural networks via accelerated gradient descent. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2020.

[58] Shuai Zhang, Meng Wang, Jinjun Xiong, Sijia Liu, and Pin-Yu Chen. Improved linear convergence of training CNNs with generalizability guarantees: A one-hidden-layer case. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2622–2635, 2020.

[59] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer ReLU networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1524–1534. PMLR, 2019.

[60] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149, 2017.

[61] Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. *arXiv preprint arXiv:2102.02410*, 2021.