A Complete Characterisation of ReLU-Invariant Distributions

Jan Macdonald[*]
Technische Universität Berlin

Stephan Wäldchen[*]
Technische Univeristät Berlin

Abstract

We give a complete characterisation of families of probability distributions that are invariant under the action of ReLU neural network layers (in the same way that the family of Gaussian distributions is invariant to affine linear transformations). The need for such families arises during the training of Bayesian networks or the analysis of trained neural networks, e.g., in the context of uncertainty quantification (UQ) or explainable artificial intelligence (XAI).

We prove that no invariant parametrised family of distributions can exist unless at least one of the following three restrictions holds: First, the network layers have a width of one, which is unreasonable for practical neural networks. Second, the probability measures in the family have finite support, which basically amounts to sampling distributions. Third, the parametrisation of the family is not locally Lipschitz continuous, which excludes all computationally feasible families.

Finally, we show that these restrictions are individually necessary. For each of the three cases we can construct an invariant family exploiting exactly one of the restrictions but not the other two.

1 Introduction

Neural networks have achieved great success in solving diverse problems ranging from image analysis (Krizhevsky et al., 2012; Szegedy et al., 2013), to natural language processing (Cho et al., 2014; Vaswani et al., 2017), to medical applications (Shen et al., 2017; McBee et al., 2018).

An important theoretical as well as computational challenge is calculating the distribution of outputs of a neural network from a given distribution of inputs. This task is a crucial ingredient for uncertainty quantification, explainable AI and Bayesian learning.

More precisely, the task is to describe the distribution μ_{out} of

$$f(\mathbf{x}) \quad \text{with} \quad \mathbf{x} \sim \mu_{\text{in}},$$

where f is a neural network function and $\mathbf{x}$ is a random input vector distributed according to some probability measure μ_{in}. In other words, we want to determine the push-forward

$$\mu_{\text{out}} = f_* \mu_{\text{in}} \tag{1}$$

of μ_{in} with respect to the transformation f. Given the fact that neural networks are highly non-linear functions this generally has no closed form solution. Obtaining approximations through numerical integration methods is expensive due to the high dimensionality and complexity of the function.

This lead to the application of other approximation schemes, such as assumed density filtering (Gast and Roth, 2018; Macdonald et al., 2019) (ADF) and in the case of Bayesian neural networks the more general framework of expectation propagation (Jylänki et al., 2014; Soudry et al., 2014) (EP) which contains ADF as a special case. Both methods were originally developed for general large-scale Bayesian inference problems (Minka, 2001; Boyen and Koller, 1998).

ADF makes use of the feed-forward network structure: If we write $f = f_L \circ f_{L-1} \circ \cdots \circ f_2 \circ f_1$ as a composition of its layers f_j, then (1) decomposes as

$$
\begin{aligned}
\mu_0 &= \mu_{\text{in}}, \\
\mu_j &= (f_j)_* \mu_{j-1} \quad \text{for} \quad j = 1, \dots, L, \\
\mu_{\text{out}} &= \mu_L.
\end{aligned}
\tag{2}
$$

The task can then be described as how to propagate distributions through neural network layers.

---

[*]Both authors contributed equally to this work.

If each μ_j was in the same family of probability distributions as the original μ_{in} then we could find efficient layer-wise propagation rules for each layer f_j to obtain μ_{out}. Such a family, if expressable with a possibly large but finite number of parameters, would be extremely useful for applications involving uncertainty quantification or the need for explainable predictions.

ADF commonly propagates Gaussians, or more generally exponential families. While the Gaussian family is invariant under purely linear layers, this is not the case for non-linear activations. Hence the method relies on a projection step back onto the Gaussian family after each network layer. In general, for any chosen family of probability distributions, (2) is replaced by

$$\mu_0 = \text{proj}(\mu_{\text{in}}),$$
$$\mu_j = \text{proj}((f_j)_* \mu_{j-1}) \quad \text{for} \quad j = 1, \dots, L,$$
$$\mu_{\text{out}} = \mu_L,$$

where $\text{proj}(\cdot)$ is a suitable projection. In this case only $\mu_{\text{out}} \approx f_* \mu_{\text{in}}$ holds and there is no guarantee how good the approximation is after multiple layers. This heuristic is justified by the implicit assumption that no invariant families exist.

Due to the flexibility and richness of the class of neural network functions, it seems intuitively clear that such an invariant family of distributions cannot be realised in a meaningful way—a fact that, to the best of our knowledge, remained unproven. In this paper we will fill this gap.

1.1 Contribution

We give a complete characterisation of families of probability distributions that are invariant under the action of ReLU neural network layers. In fact, we prove that all possible invariant families belong to a set of degenerate cases that are either practically irrelevant or amount to sampling. Hence, the implicit assumption made by the ADF and EP frameworks is justified. For each of the degenerate cases we give an explicit example of an invariant family of distributions.

We show that there is no need for practitioners to search for "smarter" distribution families that can be propagated exactly. Instead, the prime consideration can be numerical convenience. The next step is of course to see if our arguments can be extended to approximately invariant families.

Our novel proof technique is based on very intuitive geometric constructions and ideas from metric dimension theory. This intuitive argument is made rigorous via the Hausdorff dimension of metric spaces. As a result, we provide a theoretical underpinning for statistical inference schemes such as ADF for neural networks.

1.2 Applications

Calculating output distributions of neural networks is a basic and central problem to make neural networks more applicable. Good performance on test data alone is not sufficient to ensure reliability. High-stakes applications such as, e.g., medical imaging or autonomous driving, make questions regarding uncertainty quantification (UQ), robustness, and explainability of network decisions inescapable (Oala et al., 2020; Begoli et al., 2019; Zhang et al., 2018; Holzinger et al., 2017; Michelmore et al., 2018; Kim and Canny, 2017).

Uncertainty Quantification: Studying the propagation of input uncertainties through an already trained (deterministic) network is used to analyse the reliability of their predictions and their sensitivity to changes in the data. ADF is used towards this goal, e.g., for sigmoid networks for automatic speech recognition (Astudillo and Neto, 2011; Astudillo et al., 2014) and ReLU networks for image classification and optical flow (Gast and Roth, 2018).

Explainable Artificial Intelligence: So far, neural networks are mostly considered opaque "black-box" models. Recent approaches to make them human-interpretable have focused on the crucial subtask of relevance attribution, i.e., finding the most relevant input variables for a given prediction. A probabilistic formulation of this task was recently proposed for classification networks by Fong and Vedaldi (2017); Wäldchen et al. (2021); Macdonald et al. (2019). Macdonald et al. (2019) also considered a heuristic solution strategy for deep ReLU networks that is based on ADF with Gaussian distributions.

Bayesian Networks: Instead of point estimates, Bayesian neural networks (BNN) learn a probability distribution over the network weights through Bayesian inference. BNNs provide an inbuilt regularisation by choosing appropriate priors. They naturally lead to network compression when using sparsity-promoting priors or encoding weights of high variance with less precision. ADF with Gaussian distributions has been employed for the fast calculation of posteriors during the training of BNNs with sigmoidal (Opper, 1999) and ReLU (Hernández-Lobato and Adams, 2015; Sun et al., 2017) activations. Wu et al. (2019) use a related moment-propagation approach for deterministic variational inference in BNNs with ReLU or Heaviside activations. Another approach was taken by Wang and Manning (2013) which relies on layer-wise Gaussian approximations to yield a fast alternative to dropout regularised training.

1.3 Outline of the Paper

In Section 2 we state our two main results and provide a high-level description of the proof ideas. In Section 3 we show that no (non-degenerate) network invariant families of distributions exist. Two constructive steps of the proof are deferred to Appendices A and B respectively. Finally, in Section 4 we prove our second main result by constructing several degenerate invariant families. We conclude with a short discussion in Section 5.

2 Characterisation of Invariant Families of Distributions

Before we rigorously state and prove our main result, we start by introducing some notation and terminology and precisely specify what we mean by *parametrised families* of probability distributions and by their *invariance* with respect to layers of neural networks.

2.1 Preliminaries

We denote by $\mathcal{D}(\mathbb{R}^d)$ the set of Radon[1] Borel probability measures on $\mathbb{R}^d$. We equip $\mathcal{D}(\mathbb{R}^d)$ with the Prokhorov metric (Prokhorov, 1956).

Definition 2.1 (Family of Distributions). Let $\Omega \subseteq \mathbb{R}^n$ and $p\colon \Omega \to \mathcal{D}(\mathbb{R}^d)$. We call p an n-parameter family of probability distributions. It is called a continuous family, if p is continuous.

It might seem more intuitive to refer to the set $\{p(\theta)\}_{\theta \in \Omega}$ as the family of distributions. However, a set of distributions can be parametrised in multiple ways and even the number n of parameters describing such a set is not unique. Whenever we speak about a family of distributions, we never think of it as a mere set of distributions but always attach it to a fixed chosen parametrisation p. All our results are stated in terms of this parametrisation.

Example 2.2 (2-dimensional Gaussian Family). Consider the parameter space $\Omega = \mathbb{R}^2 \times \widetilde{\Omega}$ where

$$\widetilde{\Omega} = \left\{ (\sigma_1, \sigma_2, \sigma_3) \in \mathbb{R}^3 : \sigma_1, \sigma_2 \geq 0,\ \sigma_1\sigma_2 - \sigma_3^2 \geq 0 \right\},$$

and $p\colon \Omega \to \mathcal{D}(\mathbb{R}^2)$ with

$$p(\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_3) = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1 & \sigma_3 \\ \sigma_3 & \sigma_2 \end{bmatrix} \right),$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then p is a continuous 5-parameter family of 2-dimensional Gaussian distributions.

---

[1]In fact $\mathbb{R}^d$ is a separable complete metric space, thus every Borel probability measure is automatically Radon.

Definition 2.3 (Invariance). Let $p\colon \Omega \to \mathcal{D}(\mathbb{R}^d)$ be a family of probability distributions and let $f\colon \mathbb{R}^d \to \mathbb{R}^d$ be any measurable function. Then the family is called f-invariant, if for any $\theta \in \Omega$ the pushforward of the measure $p(\theta)$ under f is again in the family, i.e., there exists $\omega \in \Omega$ such that $p(\omega) = f_* p(\theta)$. For a collection $\mathcal{F} \subseteq \{ f\colon \mathbb{R}^d \to \mathbb{R}^d\ :\ f \text{ measurable} \}$ of measurable functions it is called $\mathcal{F}$-invariant, if it is f-invariant for all $f \in \mathcal{F}$.

Example 2.4. (Invariance to Linear Transformations) The family p from Example 2.2 is $\mathcal{F}$-invariant with respect to affine linear transformations

$$\mathcal{F} = \{ f\colon \mathbb{R}^2 \to \mathbb{R}^2\ : f(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{b},$$
$$\mathbf{W} \in \mathbb{R}^{2 \times 2},\ \mathbf{b} \in \mathbb{R}^2 \}.$$

We are interested in the special case of invariance with respect to ReLU layers of neural networks. The rectified linear unit (ReLU) (Nair and Hinton, 2010; Glorot et al., 2011), defined as $\varrho(x) = \max\{0, x\}$, has emerged as the dominant choice for activation functions as of today (Ramachandran et al., 2017).

Definition 2.5 (ReLU-Invariance). Let $p\colon \Omega \to \mathcal{D}(\mathbb{R}^d)$ be a family of distributions. The family is called ReLU-invariant, if it is $\mathcal{F}$-invariant for the collection

$$\mathcal{F} = \{ f\colon \mathbb{R}^d \to \mathbb{R}^d\ : f(\mathbf{x}) = \varrho(\mathbf{W}\mathbf{x} + \mathbf{b}),$$
$$\mathbf{W} \in \mathbb{R}^{d \times d},\ \mathbf{b} \in \mathbb{R}^d \},$$

where $\varrho(x) = \max\{0, x\}$ is applied componentwise.

This also includes invariance with respect to ReLU-layers with non-square matrices, since they could be made square by zero-padding. This is explained in more detail in Appendix A.

We observe that if a family p is both f-invariant and g-invariant for two functions f and g such that the composition $g \circ f$ is well-defined, then it is also $(g \circ f)$-invariant. In particular, a ReLU-invariant family is invariant for all ReLU networks of any depths L.

2.2 The Main Results

We can now state the main theorem of this work.

Theorem 2.6. *Let $\Omega \subseteq \mathbb{R}^n$ and $p\colon \Omega \to \mathcal{D}(\mathbb{R}^d)$ be a continuous and ReLU-invariant n-parameter family of probability distributions. Then at least one of the following restrictions has to hold:*

R1. Restricted Dimension: $d = 1$,

R2. Restricted Support: $\mathrm{supp}(p(\theta))$ *is finite for all $\theta \in \Omega$,*

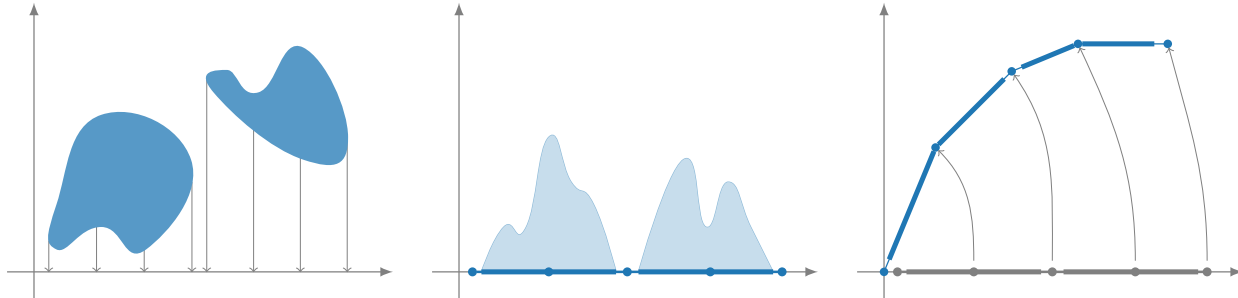**R3. Restricted Regularity:** p *is not locally Lipschitz continuous.*

Figure 1: Main steps of transforming a generic probability distribution to a distribution supported on a polygonal chain. The original distribution (left) is projected onto a single dimension and partitioned into intervals (centre) that each contain a sufficiently large portion of the probability mass (shown as the lightly shaded region). Finally it is "bent" into a polygonal chain (right). The number of segments of the chain and its segment lengths can be chosen arbitrarily.

This can be interpreted as follows: Besides some rather degenerate cases there cannot be any family of probability distributions that is invariant with respect to the layers of a ReLU neural network. The three restrictions characterise which kind of degenerate cases can occur.

Restriction (R1) Neural networks with only one neuron per layer are not really powerful function classes, so that the ReLU-invariance is not a strong limitation in this case. However, already two dimensions are enough to unlock the expressive power of neural networks.

Restriction (R2) Under mild regularity assumptions on the parametrisation the only ReLU-invariant distributions in dimensions $d \geq 2$ are finite mixtures of Dirac distributions. This amounts to sampling distributions of finite size, which of course can work in many scenarios but are often too computationally expensive in high-dimensions.

Restriction (R3) Continuity alone is not a strong enough assumptions on the parametrisation. This is due to the fact that the class of continuous functions is too flexible and includes non-intuitive examples such as space-filing curves. These can be used to construct invariant distributions. However, such parametrisations have to be be rather wild and would be impractical to work with. A slightly stronger assumption like local Lipschitz continuity, fulfilled by virtually all commonly used parametrisations of distribution, is enough to exclude these degenerate examples.

We complement Theorem 2.6 by providing a kind of reverse statement showing that the three restrictions are individually necessary.

Theorem 2.7. *For each of the restrictions (R1), (R2), and (R3) there exists a continuous ReLU-invariant n-parameter family of probability distributions that is subject to that restriction but avoids the other two.*

The proof for Theorem 2.6 will be given in Section 3 and the proof for the reverse result Theorem 2.7 in Section 4.

3 Proof of the Main Result (Theorem 2.6)

The main idea for proving Theorem 2.6 is to use the layers of a neural network to transform simple, more or less arbitrary probability distributions to complicated distributions that need an arbitrarily high number of parameters to be described. But a family of parametrised distributions necessarily has a finite number of parameters leading to a contradiction if the family is assumed to be invariant under neural network layers. We present a high-level construction before going into detail.

More precisely, ReLU neural network layers can be used to transform arbitrary probability distributions to distributions that are supported on certain polygonal chains, which we call arcs. This is visualised in Figure 1. These arcs can be described by the lengths of their line segments. However, the number of segments can be made larger than the number of parameters describing the family of distributions as long as the support of the initial distribution is large enough. From this a contradiction can be derived. There are only three ways to circumvent this, corresponding to the three restrictions in Theorem 2.6. Firstly, restricting the allowed neural network layers so that transformations to the arcs are not possible. Secondly, restricting the support of the distributions so that the number of line segments they can be transformed to is limited. Thirdly, using a wild parametrisation that leverages ideas similar to space-filling curves in order to use only few parameters to describe a set that effectively would require more parameters.

The idea of mapping (a subset of) the parametrisation

$$\Omega \supseteq \Omega_{\delta,m} \xrightarrow{\quad \Xi \quad} \mathbb{R}_+^{m-1} \times \{1\} \subseteq \mathbb{R}_+^m$$
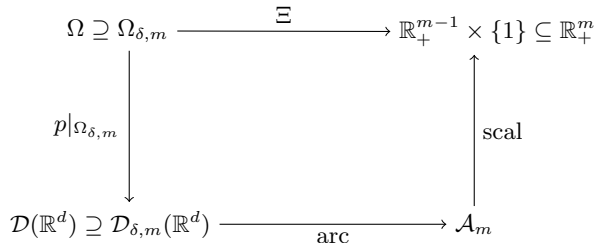
Figure 2: Schematic overview of the spaces and functions involved in our proof.

domain to distributions supported on polygonal arcs, then to the respective arcs, and finally to segment lengths is schematically shown in Figure 2. We will give exact definitions of all involved spaces and mappings below.

3.1 Details of the Proof

We will now present the main steps of the proof of Theorem 2.6. The proofs of the individual steps will be deferred in order to not distract from the main idea. We will assume towards a contradiction that all three restrictions in Theorem 2.6 are not satisfied.

Assumption 3.1. Let $d \geq 2$, $\Omega \subseteq \mathbb{R}^n$, and $p\colon \Omega \to \mathcal{D}(\mathbb{R}^d)$ be a locally Lipschitz continuous and ReLU-invariant n-parameter family of probability distributions such that there exists a $\theta \in \Omega$ for which $\mathrm{supp}(p(\theta))$ is not finite.

From this we will derive a contradiction. The proof is based on the idea of transforming arbitrary probability distributions to distributions supported on certain polygonal chains, which we call arcs. We denote the convex hull operator as $\mathrm{conv}(\cdot)$.

Definition 3.2 (Arcs)**.** An m-arc is a subset $A \subseteq \mathbb{R}^2$ defined as a polygonal chain, i.e., a connected piecewise linear curve,

$$A = \bigcup_{i=1}^m \mathrm{conv}(\{\mathbf{v}_{i-1}, \mathbf{v}_i\}),$$

determined by $m + 1$ vertices $\{\mathbf{v}_0, \dots, \mathbf{v}_m\}$, that for some length scales $r_1, \dots, r_m > 0$ and a fixed angle $\phi_m = \frac{\pi}{2m}$ satisfy

$$\mathbf{v}_i = \mathbf{v}_{i-1} + r_i \begin{bmatrix} \sin(i\phi_m) \\ \cos(i\phi_m) \end{bmatrix}, \quad i = 1, \dots, m.$$

We denote the set of vertices of an m-arc A as $\mathrm{vert}(A) = \{\mathbf{v}_0, \dots, \mathbf{v}_m\}$, the set of its line segments as $\mathrm{segm}(A) = \{\ell_1, \dots, \ell_m\}$, where $\ell_i = \mathrm{conv}(\{\mathbf{v}_{i-1}, \mathbf{v}_i\})$, and the set of its scaling factors as $\mathrm{scal}(A) = \{r_1, \dots, r_m\}$.
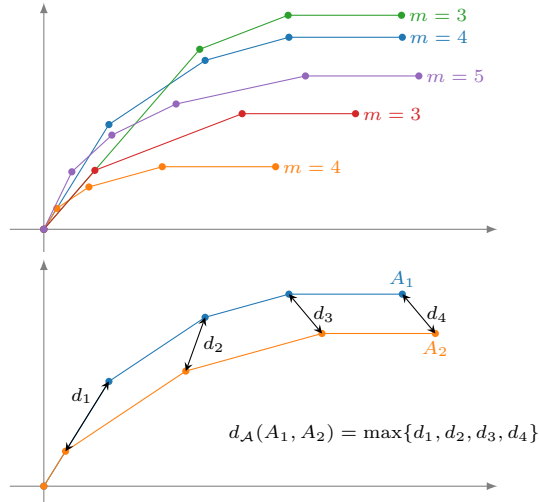

Figure 3: Examples of several standard m-arcs for varying m (top). Standardised arcs are always contained within the non-negative orthant and end on a horizontal line segment. The distance between two m-arcs is the maximal Euclidean distance of corresponding vertices (bottom).

It will turn out to be useful to remove some ambiguity from the set of m-arcs by standardising their behaviour at the start and end vertices. Some examples can be seen in Figure 3.

Definition 3.3 (Standard Arcs)**.** An m-arc A with vertices $\{\mathbf{v}_0, \dots, \mathbf{v}_m\}$ and length scales $\{r_1, \dots, r_m\}$ is called standardised (or a standard m-arc), if it starts at the origin and has a normalised last line segment, i.e., if $\mathbf{v}_0 = \mathbf{0}$ and $r_m = 1$. The set of all standard m-arcs is denoted $\mathcal{A}_m$ and equipped with the metric

$$d_{\mathcal{A}}(A_1, A_2) = \max_{i=1,\dots,m} \| \mathrm{vert}(A_1)_i - \mathrm{vert}(A_2)_i \|_2.$$

The metric $d_{\mathcal{A}}$ is induced by a mixed (ℓ_2, ℓ_∞)-norm and thus indeed a proper metric. There is a one-to-one correspondence between $\mathcal{A}_m$ and $\mathbb{R}_+^{m-1} \times \{1\} \cong \mathbb{R}_+^{m-1}$ via the scaling factors.

Definition 3.4 (Arc-Supported Measures)**.** For any $\delta > 0$, a probability measure $\mu \in \mathcal{D}(\mathbb{R}^2)$ is said to be δ-distributed on a standard m-arc $A \in \mathcal{A}_m$ if it satisfies $\mathrm{supp}(\mu) \subseteq A$ and for each line segment $\ell \in \mathrm{segm}(A)$ we have $\mu(\ell) \geq \delta$.

Accordingly, for $d > 2$ we say that a probability measure $\mu \in \mathcal{D}(\mathbb{R}^d)$ is δ-distributed on a standard m-arc $A \in \mathcal{A}_m$ if $\mathrm{supp}(\mu) \subseteq \mathrm{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ and by identifying $\mathrm{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ with $\mathbb{R}^2$ it is δ-distributed on a standard m-arc in the above sense.[2]

---

[2]There is no particular significance to choosing $\mathrm{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ here. This is done for ease of notation, but any two-dimensional subspace would work as well.

We denote the set of all probability measures that are δ-distributed on standard m-arcs by $\mathcal{D}_{\delta,m}(\mathbb{R}^d) \subseteq \mathcal{D}(\mathbb{R}^d)$.

Lemma 3.5. *Let $\mu \in \mathcal{D}_{\delta,m}(\mathbb{R}^d)$ be δ-distributed on a standard m-arc, then this arc is unique.*

Proof. Towards a contradiction assume that μ is δ-distributed on $A_1, A_2 \in \mathcal{A}_m$ and $A_1 \neq A_2$. Let $\{\ell_1^j, \ldots, \ell_m^j\} = \text{segm}(A_j)$ and $\{r_1^j, \ldots, r_m^j\} = \text{scal}(A_j)$ denote the line segments and scaling factors of A_1 and A_2 respectively. Since $A_1 \neq A_2$ and $r_m^1 = r_m^2 = 1$ there is a smallest index $1 \leq i \leq m-1$ such that $r_i^1 \neq r_i^2$. Without loss of generality assume $r_i^1 < r_i^2$. Then ℓ_{i+1}^2 lies outside of the convex hull of A_1 and in particular

$$\ell_{i+1}^2 \cap A_1 = \emptyset.$$

But this contradicts the fact that both $\mu(\ell_{i+1}^2) \geq \delta$ and $\text{supp}(\mu) \subseteq A_1$ must hold. $\qquad\square$

Lemma 3.5 shows that arc-supported measures induce unique standard m-arcs. We denote the corresponding function mapping a measure $\mu \in \mathcal{D}_{\delta,m}(\mathbb{R}^d)$ to its induced arc $A_\mu \in \mathcal{A}_m$ by $\text{arc}: \mathcal{D}_{\delta,m}(\mathbb{R}^d) \to \mathcal{A}_m : \mu \mapsto A_\mu$. For an n-parameter family of distributions, $p: \Omega \to \mathcal{D}(\mathbb{R}^d)$, we denote by

$$\Omega_{\delta,m} = \big\{ \theta \in \Omega \,:\, p(\theta) \in \mathcal{D}_{\delta,m}(\mathbb{R}^d) \big\} \subseteq \Omega$$

the set of parameters mapping to arc-supported measures. Without further assumptions on p this set might well be empty. We will see however, that this cannot happen for ReLU-invariant families as long as they contain at least one measure with a support of cardinality at least m.

We are now ready to start discussing the main steps of the proof of Theorem 2.6. It relies on the following three results.

Lemma 3.6 (Surjectivity)**.** *Under Assumption 3.1 and for any $m \in \mathbb{N}$ there exists a $\delta > 0$ such that the map $\Xi: \Omega_{\delta,m} \subseteq \mathbb{R}^n \to \mathbb{R}_+^{m-1} \times \{1\}$ given by*

$$\Xi = \text{scal} \circ \text{arc} \circ p|_{\Omega_{\delta,m}}$$

is surjective. In particular, in this case the domain $\Omega_{\delta,m}$ is non-empty.

The proof for this is given in Appendix A. In short, we show that any measure with non finite support can be transformed by ReLU layers into a measure supported on an arbitrary standard arc. This can be done in four constructive steps: (i) projecting the measure onto a one-dimensional subspace; (ii) partitioning its support along that line into segments; (iii) appropriately scaling the line segments; (iv) bending the line into arc shape by repeated rotations and projections onto the non-negative orthant. Finally, using the arc scaling factors to identify the set of arcs with $\mathbb{R}_+^{m-1} \times \{1\}$ yields the claim.

Lemma 3.7 (Local Lipschitz continuity)**.** *Under Assumption 3.1 and for any $m \in \mathbb{N}$ let $\delta > 0$ and $\Xi: \Omega_{\delta,m} \subseteq \mathbb{R}^n \to \mathbb{R}_+^{m-1} \times \{1\}$ be as in Lemma 3.6. Then Ξ is locally Lipschitz continuous.*

The proof for this is given in Appendix B. In short, we show that all three partial functions scal, arc, and p are locally Lipschitz continuous, hence also their composition Ξ is.

As a final piece before the main theorem we need a rather general result on maps between Euclidean spaces.

Lemma 3.8. *Let $B \subseteq \mathbb{R}^n$ and $f: B \to \mathbb{R}^m$ be locally Lipschitz continuous. If the image $f(B)$ is a Borel set with non-empty interior in $\mathbb{R}^m$, then $m \leq n$. In particular, if f maps surjectively onto $\mathbb{R}^m$ or $\mathbb{R}_+^m$, then $m \leq n$.*

The full proof of this is given in Appendix E. It essentially uses the fact that locally Lipschitz continuous maps cannot increase the Hausdorff dimension. Since the Hausdorff dimension of B is at most n and the Hausdorff dimension of $f(B)$ is m, the claim follows.

We now have all the ingredients to prove the main theorem.

Proof of Theorem 2.6. Towards a contradiction let Assumption 3.1 hold. Let $m > n+1$ be arbitrary. By Lemma 3.6 there exists a $\delta > 0$ so that $\Omega_{\delta,m}$ is non-empty and $\Xi: \Omega_{\delta,m} \to \mathbb{R}_+^{m-1} \times \{1\}$ is surjective. By Lemma 3.7 it is also locally Lipschitz continuous. Since $\mathbb{R}_+^{m-1} \times \{1\}$ is isometric to $\mathbb{R}_+^{m-1}$ and $m-1 > n$ this contradicts Lemma 3.8. Hence, one of the assumptions in Assumption 3.1 cannot hold, which proves Theorem 2.6. $\qquad\square$

4 Proof of the Reverse Result (Theorem 2.7)

We will now come to the proof of Theorem 2.7 and show that exploiting any of the three restrictions in Theorem 2.6 indeed allows us to find ReLU-invariant families of distributions. We split the proof into three parts and construct examples for each of the three cases. Here, we briefly provide the main proof ideas and refer to Appendices C and D for the detailed and formal proofs.

4.1 Families of Distributions in One Dimension

The first part of Theorem 2.7 corresponds to restriction (R1). We explicitly construct a family of ReLU-
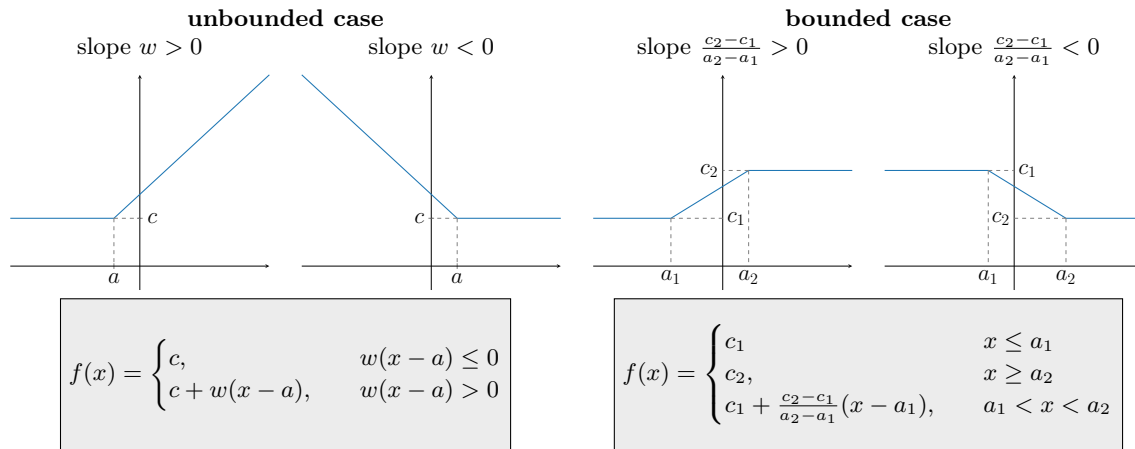
Figure 4: ReLU networks in dimension $d = 1$ can only take one of five distinct forms. We show the four non-constant cases. This includes two unbounded cases (left) and two bounded cases (right).

invariant probability distributions on $\mathbb{R}$ with a locally Lipschitz continuous parametrisation map.

The key observation is that ReLU neural networks in one dimension are a rather restricted class of functions unlike networks in higher dimensions. In fact, every one-dimensional ReLU neural network of arbitrary depth is either constant or takes one of four forms, illustrated in Figure 4, and can therefore be rewritten as a three layer network. Thus, we only require a constant number of parameters to describe the set of all one-dimensional ReLU networks. It is then straightforward to obtain a ReLU-invariant family of probability distributions by explicitly making these parameters part of the family's parametrisation.

In fact, we can use a *prototype* measure $\mu_0 \in \mathcal{D}(\mathbb{R})$ with non-finite support to obtain a parametrisation map

$$p \colon \mathbb{R}^6 \to \mathcal{D}(\mathbb{R}) \colon \theta \mapsto (f_\theta)_* \mu_0,$$

where f_θ denotes a ReLU network with weights and biases $\theta = (w_1, b_1, w_2, b_2, w_3, b_3) \in \mathbb{R}^6$. This parametrisation is shown to be locally Lipschitz continuous and ReLU-invariant in Appendix C.

4.2 Families of Distributions with Finite Support

The second part of Theorem 2.7 corresponds to (R2). We show that ReLU-invariant families of probability distributions exist in which all distributions are finitely supported. A finitely supported Radon probability measure $\mu \in \mathcal{D}(\mathbb{R}^d)$ can be expressed as a mixture of finitely many Dirac measures,

$$\mu = \sum_{i=1}^{N} c_i \delta_{\mathbf{a}_i},$$

for some $N \in \mathbb{N}$, $\mathbf{a}_i \in \mathbb{R}^d$, and $0 \leq c_i \leq 1$ with $\sum_{i=0}^{N} c_i = 1$. For any ReLU neural network layer $f \colon \mathbb{R}^d \to \mathbb{R}^d \colon \mathbf{x} \mapsto \varrho(\mathbf{W}\mathbf{x} + \mathbf{b})$ we get

$$f_*(\mu) = \sum_{i=1}^{N} c_i \delta_{\varrho(\mathbf{W}\mathbf{a}_i + \mathbf{b})},$$

which is again a mixture of not more than N Dirac measures. Hence, for any $N \in \mathbb{N}$ the set $\mathcal{D}_N = \{ \mu \in \mathcal{D}(\mathbb{R}^d) : |\operatorname{supp}(\mu)| \leq N \}$ is ReLU-invariant. It can be described with $n = N(d + 1)$ parameters by

$$p_N \colon \underbrace{\mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{N \text{ times}} \times \Delta_N \to \mathcal{D}(\mathbb{R}^d),$$

$$(\mathbf{a}_1, \ldots, \mathbf{a}_N, \mathbf{c}) \mapsto \sum_{i=1}^{N} c_i \delta_{\mathbf{a}_i}$$

with $\Delta_N = \{ \mathbf{c} \in [0, 1]^N : \sum_i c_i = 1 \}$. It is clear by the defintion of the Prokhorov metric that p_N is also Lipschitz continuous.

4.3 Families of Distributions Without Local Lipschitz Continuity

The third part of Theorem 2.7 corresponds to (R3). If we omit the local Lipschitz continuity of the parametrisation we can construct a ReLU-invariant one-parameter family of distributions in $\mathbb{R}^d$ for any dimension $d \in \mathbb{N}$. We proceed similar to the ideas in Section 4.1, however we need to take care of the fact that the set of all ReLU networks cannot easily be described by a constant number of parameters, unlike in the one-dimensional case.

Instead, each ReLU network can be represented by a finite but arbitrarily large number of parameters. This corresponds to a vector in $\mathbb{R}^\infty$ (the space of all

real-valued and eventually vanishing sequences). Additionally, we can encode these parameters by a single parameter with the help of space filling curves. More precisely, there exists a continuous map $\Gamma\colon \mathbb{R} \to \mathbb{R}^\infty$, a fact that is guaranteed by the Hahn-Mazurkiewicz theorem, see for example (Sagan, 1994, Chapter 6.8).

Putting both ideas together, we map a single parameter $\theta \in \mathbb{R}$ to $\Gamma(\theta) \in \mathbb{R}^\infty$, which represents the weights and biases of the neural network $f_{\Gamma(\theta)}$. Including this in the parametrisation p, analogously to Section 4.1, results in

$$p\colon \mathbb{R} \to \mathcal{D}(\mathbb{R}^d)\colon \theta \mapsto (f_{\Gamma(\theta)})_* \mu_0,$$

where $\mu_0 \in \mathcal{D}(\mathbb{R}^d)$ is some prototype measure. Since $d \in \mathbb{N}$ and $\mu_0 \in \mathcal{D}(\mathbb{R}^d)$ can be chosen arbitrarily, we can easily evade the other two restrictions (R1) and (R2). The detailed construction of p and proof of its ReLU-invariance can be found in Appendix D.

5   Discussion

Our work represents a comprehensive characterisation of distribution families that are invariant under transformations by layers of ReLU neural networks. The only invariant distributions are either sampling distributions (i.e. distributions of finite support) or rather degenerate and elaborately constructed distributions that are infeasible for practical applications. This justifies the use of approximation schemes such as assumed density filtering.

To conclude, we want to quickly comment on some different possible modifications to our question.

Different Activations: We have limited our analysis to the class of functions defined as layers of ReLU neural networks. In particular, the constructive steps for the proof of Lemma 3.6 (surjectivity of Ξ) in Appendix A make explicit use of the fact that the ReLU is a continuous, piecewise-linear function. Alternative continuous, piece-wise linear activations, for example the leaky ReLU (Maas et al., 2013) with a nonzero left-side slope, should allow for an analogous construction.

Different commonly used activation functions are the tanh, the logistic function, or the Heaviside function. Since the Heaviside function is not continuous and has a discrete and finite range, it transforms any distribution into a distribution with finite support. Clearly, the only invariant distributions can be sampling distributions. For smooth activation functions, such as tanh and the logistic function, we are not aware of any characterisation of invariant distributions. Extending our proof strategy to this scenario is not straightforward as it relies on specific properties of the ReLU. We leave this question open for future research.

Restricted Weights and Biases: A second variation on the class of functions would be to put restrictions on the weight matrices or bias vectors. This could be either general constraints, for example non-negativity or positive-definiteness, or more specific restrictions by allowing the weight matrices only to be chosen from a small set of allowed matrices. Both kinds of constraints arise in the context of iterative reconstruction methods for solving inverse and sparse coding problems (Hoyer, 2002) or in corresponding unrolled and learnable iterative algorithms (Gregor and LeCun, 2010; Kobler et al., 2017). Here, the restrictions on the weight matrices often stem from physical constraints of the model describing the inverse problem.

We comment on both aspects of this variation in detail in Appendix G. Interestingly, allowing only a finite or even countable set of weights and biases results in a possible parametrisation of invariant distributions that circumvent the three restrictions (R1)–(R3) in Theorem 2.6. This is because the resulting function collection $\mathcal{F}$ is then countable as well, in contrast to the continuum of functions we considered until now. To construct an invariant family in this case, we start with a prototype measure, analogously to Section 4.1 and Section 4.3. This gets mapped to a countable set of push-forward measures by functions in $\mathcal{F}$. We show that these can be locally Lipschitz-continuously parametrised making use of interpolation functions based on tree-like graphs, see Figure 9 in Appendix G. These parametrisations, however, are purely of theoretical interest, since they rely on calculating and interpolating infinitely many distributions and are thus not of practical relevance.

Approximate Invariance: Finally, we want to comment that exact invariance might not be necessary for numerical application. Instead a notion of approximate invariance can be sufficient if the resulting error can be reasonably bounded for every layer. So far, to the best of our knowledge, no distribution families are known where meaningful bounds exist. Whether this can indeed be realised is left for further investigation.

References

R. F. Astudillo and J. Neto. Propagation of uncertainty through multilayer perceptrons for robust automatic speech recognition. In *Twelfth Annual Conference of the International Speech Communication Association*, pages 461–464, 01 2011.

R. F. Astudillo, A. Abad, and I. Trancoso. Accounting for the residual uncertainty of multi-layer perceptron based features. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6859–6863, 2014.

E. Begoli, T. Bhattacharya, and D. Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 33–42, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. URL http://dl.acm.org/citation.cfm?id=2074094.2074099.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 10 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179.

J. Dugundji. *Topology*. Allyn and Bacon, Inc, 1966.

G. Edgar. *Measure, Topology, and Fractal Geometry*. Spinger-Verlag New York, 2nd edition, 2008. doi: 10.1007/978-0-387-74749-1.

R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

J. Gast and S. Roth. Lightweight probabilistic deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3369–3378, 6 2018. doi: 10.1109/CVPR.2018.00355.

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 315–323, Fort Lauderdale, FL, USA, 4 2011. PMLR. URL http://proceedings.mlr.press/v15/glorot11a.html.

K. Gregor and Y. LeCun. Learning Fast Approximations of Sparse Coding. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 399–406, Haifa, Israel, 6 2010. Omnipress. URL http://www.icml2010.org/papers/449.pdf.

J. M. Hernández-Lobato and R. P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, page 1861–1869, Lille, France, 2015. JMLR.

A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell. What do we need to build explainable ai systems for the medical domain? 2017.

P. O. Hoyer. Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002. doi: 10.1109/NNSP.2002.1030067.

P. Jylänki, A. Nummenmaa, and A. Vehtari. Expectation propagation for neural networks with sparsity-promoting priors. *The Journal of Machine Learning Research*, 15(1):1849–1901, 2014.

J. Kim and J. Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2942–2950, 2017.

E. Kobler, T. Klatzer, K. Hammernik, and T. Pock. Variational Networks: Connecting Variational Methods and Deep Learning. In V. Roth and T. Vetter, editors, *Pattern Recognition*, pages 281–293. Springer International Publishing, 2017. ISBN 978-3-319-66709-6.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

A. L. Maas, A. Y. Hannun, A. Y. Ng, et al. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

J. Macdonald, S. Wäldchen, S. Hauch, and G. Kutyniok. A rate-distortion framework for explaining neural network decisions. Preprint, arXiv:1905.11092, 2019.

M. P. McBee, O. A. Awan, A. T. Colucci, C. W. Ghobadi, N. Kadom, A. P. Kansagra, S. Tridandapani, and W. F. Auffermann. Deep learning in radiology. *Academic Radiology*, 25(11):1472–1480, 2018. doi: 10.1016/j.acra.2018.02.018.

R. Michelmore, M. Kwiatkowska, and Y. Gal. Evaluating uncertainty quantification in end-to-end autonomous driving control. Preprint, arXiv:1811.06817, 2018.

T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2001. AAI0803033.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, page 807–814, Madison, WI, USA, 2010. Omnipress.

L. Oala, J. Fehr, L. Gilli, P. Balachandran, A. W. Leite, S. Calderon-Ramirez, D. X. Li, G. Nobis, E. A. M. n. Alvarado, G. Jaramillo-Gutierrez, C. Matek, A. Shroff, F. Kherif, B. Sanguinetti, and T. Wiegand. ML4H Auditing: From Paper to Practice. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136, pages 280–317. PMLR, 12 2020. URL http://proceedings.mlr.press/v136/oala20a.html.

M. Opper. *A Bayesian Approach to On-line Learning*, page 363–378. Publications of the Newton Institute. Cambridge University Press, 1999. doi: 10.1017/CBO9780511569920.017.

P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108: 296 – 330, 2018. doi: https://doi.org/10.1016/j.neunet.2018.08.019.

P. Petersen, M. Raslan, and F. Voigtlaender. Topological properties of the set of functions generated by neural networks of fixed size. Preprint, arXiv:1806.08459, 2018.

Y. V. Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956. doi: 10.1137/1101016.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. Preprint, arXiv:1710.05941, 2017.

H. Sagan. *Space-Filling Curves.* Spinger-Verlag New York, 1994. doi: 10.1007/978-1-4612-0871-6.

D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017. doi: 10.1146/annurev-bioeng-071516-044442.

D. Soudry, I. Hubara, and R. Meir. Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

S. Sun, C. Chen, and L. Carin. Learning Structured Weight Uncertainty in Bayesian Neural Networks. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1283–1292. PMLR, 4 2017. URL https://proceedings.mlr.press/v54/sun17b.html.

C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2553–2561. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection.pdf.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

S. Wäldchen, J. Macdonald, S. Hauch, and G. Kutyniok. The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70:351—387, 1 2021. doi: 10.1613/jair.1.12359.

S. Wang and C. Manning. Fast dropout training. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pages 118–126, Atlanta, Georgia, USA, 6 2013. PMLR. URL http://proceedings.mlr.press/v28/wang13a.html.

A. Wu, S. Nowozin, E. Meeds, R. E. Turner, J. M. Hernandez-Lobato, and A. L. Gaunt. Deterministic Variational Inference for Robust Bayesian Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=B1l08oAct7.

Z. Zhang, M. W. Beck, D. A. Winkler, B. Huang, W. Sibanda, H. Goyal, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine*, 6(11), 2018.

Supplementary Material: A Complete Characterisation of ReLU-Invariant Distributions

A   Proof of Lemma 3.6: Surjectivity of Ξ

For the remainder of the section let Assumption 3.1 hold. We will prove Lemma 3.6 and show that for any $m \in \mathbb{N}$ there exists a $\delta > 0$ so that the map $\Xi \colon \Omega_{\delta,m} \to \mathbb{R}^{m-1}_+ \times \{1\}$ is surjective. We will do this by transforming (through a series of ReLU layers) any measure with infinite support into a measure that is supported on an arbitrary standard m-arc. Such an infinitely supported measure exists within the n-parameter family p by assumption. Due to the ReLU-invariance, the transformed measure is then also included in the family. As a consequence $\Omega_{\delta,m}$ must be non-empty and in fact we show that $\mathrm{arc} \circ p|_{\Omega_{\delta,m}} \colon \Omega_{\delta,m} \to \mathcal{A}_m$ is surjective. The choice of δ will depend on how "evenly spread" the support of the original, untransformed, measure is. Finally, $\mathcal{A}_m$ can be identified with $\mathbb{R}^{m-1}_+ \times \{1\}$ via the scaling factors.

As a first step we observe how transforming measures by continuous functions affects their support. The support $\mathrm{supp}(\mu)$ of a Radon measure μ is defined as the complement of the largest open μ-null set. A point $\mathbf{x}$ belongs to $\mathrm{supp}(\mu)$ if and only if every open neighbourhood of $\mathbf{x}$ has positive measure. A Borel set contained in the complement of $\mathrm{supp}(\mu)$ is a μ-null set. The converse holds for open sets, i.e. an open set intersecting $\mathrm{supp}(\mu)$ has positive measure (in fact it suffices if the intersection is relatively open in $\mathrm{supp}(\mu)$).

Lemma A.1. *For any $q, r \in \mathbb{N}$ let $\mu \in \mathcal{D}(\mathbb{R}^q)$ and let $f \colon \mathbb{R}^q \to \mathbb{R}^r$ be continuous. Then*

$$\mathrm{supp}(f_*(\mu)) = \overline{f(\mathrm{supp}(\mu))},$$

where $f_*(\mu) \in \mathcal{D}(\mathbb{R}^r)$ is the pushforward of μ under f.

Proof. We begin by showing $f(\mathrm{supp}(\mu)) \subseteq \mathrm{supp}(f_*(\mu))$. Let $\mathbf{t} \in f(\mathrm{supp}(\mu))$ and $U \subseteq \mathbb{R}^r$ be any open neighbourhood of $\mathbf{t}$. There exists $\mathbf{x} \in \mathrm{supp}(\mu)$ so that $\mathbf{t} = f(\mathbf{x})$. Since $\mathbf{t} \in U$, we get $\mathbf{x} \in f^{-1}(U)$. Thus $f^{-1}(U)$ is an open neighbourhood of $\mathbf{x}$ and since $\mathbf{x} \in \mathrm{supp}(f)$ we get $f_*(\mu)(U) = \mu(f^{-1}(U)) > 0$. Since U was an arbitrary neighbourhood of $\mathbf{t}$ this shows $\mathbf{t} \in \mathrm{supp}(f_*(\mu))$. This implies $f(\mathrm{supp}(\mu)) \subseteq \mathrm{supp}(f_*(\mu))$. But $\mathrm{supp}(f_*(\mu))$ is closed, so we can even conclude $\overline{f(\mathrm{supp}(\mu))} \subseteq \mathrm{supp}(f_*(\mu))$.

We will show the converse inclusion by showing $\overline{f(\mathrm{supp}(\mu))}^c \subseteq \mathrm{supp}(f_*(\mu))^c$. For this let $\mathbf{t} \in \overline{f(\mathrm{supp}(\mu))}^c$. Then there exists an open neighbourhood U of $\mathbf{t}$ such that $U \cap f(\mathrm{supp}(\mu)) = \emptyset$, which implies $f^{-1}(U) \cap \mathrm{supp}(\mu) = \emptyset$. But this is only possible if $f_*(\mu)(U) = \mu(f^{-1}(U)) = 0$ and therefore $\mathbf{t} \in \mathrm{supp}(f_*(\mu))^c$. $\square$

To simplify notations in the following, we begin the transformation by establishing a standardised situation that takes places exclusively in the two-dimensional subspace $\mathrm{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ of $\mathbb{R}^d$. In fact, all m-arcs are essentially one-dimensional objects embedded in a two-dimensional space. Therefore, we start by transforming the infinitely supported measure to the non-negative part of the one-dimensional space $\mathrm{span}\{\mathbf{e}_1\}$, which will then subsequently be "bent" into the correct arc living in $\mathrm{span}\{\mathbf{e}_1, \mathbf{e}_2\}$.

A.1   Projection to a One-Dimensional Subspace

Lemma A.2. *Let $\mu \in \mathcal{D}(\mathbb{R}^d)$ such that $\mathrm{supp}(\mu)$ is not finite. Then there exists $\mathbf{W} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ such that the image of $f \colon \mathbb{R}^d \to \mathbb{R}^d \colon \mathbf{x} \mapsto \varrho(\mathbf{W}\mathbf{x} + \mathbf{b})$ is contained in $\mathrm{span}\{\mathbf{e}_1\}$ and $\mathrm{supp}(f_*(\mu))$ is infinite.*

Proof. Consider the coordinate projections $\mathrm{proj}_i \colon \mathbb{R}^d \to \mathbb{R} \colon \mathbf{x} \mapsto \mathbf{e}_i^\top \mathbf{x}$. Clearly we have

$$\mathrm{supp}(\mu) \subseteq \mathrm{proj}_1(\mathrm{supp}(\mu)) \times \cdots \times \mathrm{proj}_d(\mathrm{supp}(\mu)).$$

By assumption $\text{supp}(\mu)$ is infinite and the product can only be infinite if at least one of its factors is infinite, for example $\text{proj}_j(\text{supp}(\mu))$. If $\text{proj}_j(\text{supp}(\mu)) \cap \mathbb{R}_+$ is infinite, we set $\sigma = +1$, otherwise $\text{proj}_j(\text{supp}(\mu)) \cap \mathbb{R}_-$ must be infinite and we set $\sigma = -1$. Now choosing

$$\mathbf{W} = \begin{bmatrix} \sigma \mathbf{e}_j & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}^\top \quad \text{and} \quad \mathbf{b} = \mathbf{0},$$

clearly yields $f(\mathbb{R}^d) \subseteq \text{span}\{\mathbf{e}_1\}$. Further, by Lemma A.1 we have

$$
\begin{aligned}
\text{supp}(f_*(\mu)) &= \overline{f(\text{supp}(\mu))} \\
&= \overline{\varrho\big(\sigma \text{proj}_j(\text{supp}(\mu))\big)} \times \{0\} \times \cdots \times \{0\} \\
&= \begin{cases} \overline{\text{proj}_j(\text{supp}(\mu)) \cap \mathbb{R}_+} \times \{0\} \times \cdots \times \{0\}, & \text{if } \sigma = +1 \\ \overline{-\text{proj}_j(\text{supp}(\mu)) \cap \mathbb{R}_+} \times \{0\} \times \cdots \times \{0\}, & \text{if } \sigma = -1 \end{cases},
\end{aligned}
$$

which by the choice of σ is infinite. □

Thus, for the next step we only need to consider measures with infinite support within the non-negative part of $\text{span}\{\mathbf{e}_1\}$. Two dimensions will be enough for our construction to "bend" the measures into arc shape. Any two-dimensional ReLU layer $\mathbb{R}^2 \to \mathbb{R}^2 \colon \mathbf{x} \mapsto \varrho(\mathbf{W}\mathbf{x} + \mathbf{b})$ can easily be extended to a d-dimensional ReLU layer

$$\mathbb{R}^d \to \mathbb{R}^d \colon \mathbf{x} \mapsto \varrho\left(\begin{bmatrix} \mathbf{W} & \mathbf{0}_{2\times(d-2)} \\ \mathbf{0}_{(d-2)\times 2} & \mathbf{0}_{(d-2)\times(d-2)} \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{b} \\ \mathbf{0}_{(d-2)\times 1} \end{bmatrix}\right),$$

only operating on $\text{span}\{\mathbf{e}_1, \mathbf{e}_2\} \subseteq \mathbb{R}^d$ by appropriately padding with zeros. To simplify the notation, we leave out these zero paddings and just assume without loss of generality $d = 2$ from now on.

A.2   Partitioning into Line Segments

Now we show how the infinite support on the non-negative real line can be partitioned into an arbitrary number of intervals, which in the end will correspond to the arc line segments. We can write $\text{supp}(\mu) = S \times \{0\} \subseteq \mathbb{R}^2$ for some set $S \subseteq \mathbb{R}_{\geq 0}$.

Lemma A.3. *Let $S \subseteq \mathbb{R}_{\geq 0}$ be infinite. For any $m \in \mathbb{N}$ there exist $m+1$ distinct points $0 = b_0 < \cdots < b_m < \infty$ such that $S \cap (b_{j-1}, b_j) \neq \emptyset$ for all $j = 1, \ldots, m$.*

Proof. Since S is infinite there exist m distinct points $s_1, \ldots, s_m \in S$ with $0 < s_1 < \cdots < s_m$ for any $m \in \mathbb{N}$. Setting $b_0 = 0$,

$$b_j = \frac{s_j + s_{j+1}}{2} \quad \text{for} \quad j = 1, \ldots, m-1,$$

and $b_m = s_m + 1$ we get $s_j \in (b_{j-1}, b_j)$ for $j = 1, \ldots, m$. □

Further, we can restrict the support of the measure to a compact subset by clipping the set S to the range $[b_0, b_m] = [0, b_m]$. This can be achieved using the two layer ReLU network

$$\mathbb{R}^2 \to \mathbb{R}^2 \colon \mathbf{x} \mapsto \varrho\left(-\varrho\left(-\mathbf{x} + \begin{bmatrix} b_m \\ 0 \end{bmatrix}\right) + \begin{bmatrix} b_m \\ 0 \end{bmatrix}\right)$$

with b_m as in Lemma A.3. Altogether, without loss of generality we can from now on assume that μ is supported on $[b_0, b_m] \times \{0\} \subseteq \mathbb{R}^2$ and the support intersects each subset $(b_{j-1}, b_j) \times \{0\}$ for $j = 1, \ldots, m$.

A.3   Resizing the Line Segments

After partitioning the support of a distribution into intervals that each contain a certain amount of the probability mass we now want to bend it into arc shape. Each interval in the partition will correspond to a line segment. However, the bending will distort the segment lengths, so in order to obtain an arc with specified segment lengths we first have to resize the intervals. This is also possible with ReLU layers.

Lemma A.4. *For $m \geq 2$ let $0 = a_0 < a_1 < \cdots < a_m < \infty$ and $0 = b_0 < b_1 < \cdots < b_m < \infty$. Then there exists a collection of ReLU layer transformations $\{F_j \colon \mathbb{R}^2 \to \mathbb{R}^2\}_{j=0,\dots,m}$ such that their composition $F = F_m \circ F_{m-1} \circ \cdots \circ F_1 \circ F_0$ satisfies*

$$F\left(\begin{bmatrix} b_i \\ 0 \end{bmatrix}\right) = \begin{bmatrix} a_i \\ 0 \end{bmatrix} \quad \text{for all} \quad i = 0, \dots, m,$$

and restricted to $\mathbb{R}_{\geq 0} \times \{0\}$ it is a piecewise linear map with breakpoints (possibly) at $b_0, \dots, b_m$.

Proof. The proof consists of two parts. First, we iteratively construct a collection of piecewise linear and strictly increasing functions $\{f_j \colon \mathbb{R} \to \mathbb{R}\}_{j=0,\dots,m}$ such that

$$(f_j \circ \cdots \circ f_0)(b_i) = a_i \quad \text{for all} \quad i = 0, \dots, j, \tag{3}$$

for any $j = 0, \dots, m$. In particular for $j = m$ and $f = f_m \circ f_{m-1} \circ \cdots \circ f_0$ we obtain $f(b_i) = a_i$ for all $i = 0, \dots, m$. Afterwards we show how to construct $\{F_j\}_j$ from $\{f_j\}_j$.

Since $b_0 = a_0 = 0$, we can start with

$$f_0(x) = x$$

and see that (3) is satisfied for $j = 0$. Now, assuming we have already constructed strictly increasing and piecewise linear functions $f_0, \dots, f_j$ satisfying (3), we now want to construct f_{j+1}. The idea is to choose it in such a way that it leaves the points already correctly mapped by $f_j \circ \cdots \circ f_0$ unchanged but linearly transforms the remaining points so that b_{j+1} is mapped to a_{j+1}. For brevity we denote $b_i^j = (f_j \circ \cdots \circ f_0)(b_i)$ for $i = 0, \dots, m$. By assumption $b_i^j = a_i$ for $i = 0, \dots, j$ and by strict monotonicity $b_{j+1}^j > b_j^j = a_j$. We set

$$f_{j+1}(x) = \begin{cases} x, & x \leq a_j, \\ x + \frac{a_{j+1} - b_{j+1}^j}{b_{j+1}^j - a_j}(x - a_j), & x > a_j. \end{cases}$$

It is easy to see that this function leaves all points $a_1 < a_2 \cdots < a_j$ unchanged, maps b_{j+1}^j to a_{j+1} and is continuous, monotonously increasing and piecewise linear. We can also express f_{j+1} using ϱ, i.e.

$$f_{j+1}(x) = x + \frac{a_{j+1} - b_{j+1}^j}{b_{j+1}^j - a_j} \varrho(x - a_j). \tag{4}$$

Then f_{j+1} maps b_{j+1}^j to a_{j+1} and acts as the identity map on the region $x \leq a_j$ where points are already correctly mapped, as illustrated in 5. Hence clearly (3) is also satisfied for $j + 1$. To see that f_{j+1} is strictly increasing we observe that its slope is 1 on the region $x \leq a_j$ and

$$1 + \frac{a_{j+1} - b_{j+1}^j}{b_{j+1}^j - a_j} = \frac{a_{j+1} - a_j}{b_{j+1}^j - a_j} > 0$$

on the region $x \geq a_j$. This adds at most one new breakpoint at b_j to the composition $f_{j+1} \circ \cdots \circ f_0$. Continuing this process until $j = m$ finishes the first part.

Next we derive the ReLU layers $\{F_j\}_j$ from the functions $\{f_j\}_j$. For $j = 0, \dots, m-1$ we denote

$$w_{j+1} = \frac{a_{j+1} - b_{j+1}^j}{b_{j+1}^j - a_j}$$

and rewrite (4) as

$$f_{j+1}(x) = \begin{bmatrix} 1 & w_{j+1} \end{bmatrix} \varrho\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ -a_j \end{bmatrix}\right) \quad \text{for} \quad x \geq 0.$$

Two consecutive maps can be combined as

$$f_{j+1}(f_j(x)) = \begin{bmatrix} 1 & w_{j+1} \end{bmatrix} \varrho\left(\begin{bmatrix} 1 & w_j \\ 1 & w_j \end{bmatrix} \varrho\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ -a_{j-1} \end{bmatrix}\right) + \begin{bmatrix} 0 \\ -a_j \end{bmatrix}\right)$$
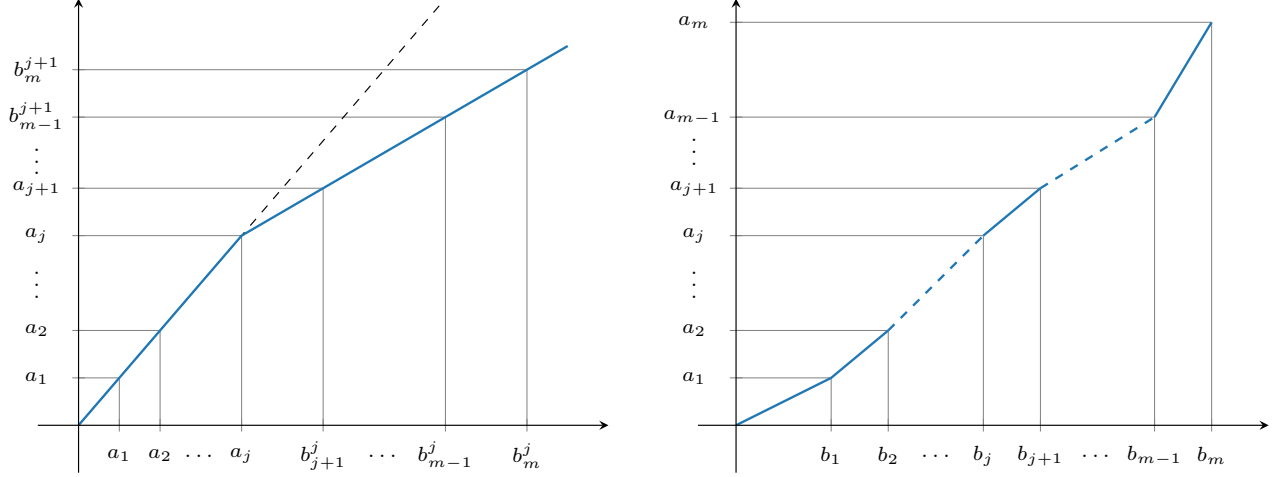
Figure 5: The function $f_{j+1}(x)$ transforms the points $0 = a_0 = b_0^j, a_1 = b_1^j, \ldots, a_j = b_j^j$ and $b_{j+1}^j, \ldots, b_m^j$ (left). The region $x \leq a_j$ is left unchanged and the region $x \geq a_j$ is linearly rescaled so that $f_{j+1}(b_{j+1}^j) = a_{j+1}$. The piecewise-linear function $f = f_m \circ \cdots \circ f_0$ maps b_j to a_j (right).

for $x \geq 0$, hence we set

$$F_j \colon \mathbb{R}^2 \to \mathbb{R}^2 \colon \mathbf{x} \mapsto \varrho\left(\begin{bmatrix} 1 & w_j \\ 1 & w_j \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ -a_j \end{bmatrix}\right)$$

for $j = 1, ..., m - 1$. It is now easy to check that with

$$F_0(\mathbf{x}) = \varrho\left(\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \mathbf{x}\right)$$

and

$$F_m(\mathbf{x}) \mapsto \varrho\left(\begin{bmatrix} 1 & w_m \\ 0 & 0 \end{bmatrix} \mathbf{x}\right)$$

we finally get

$$F\left(\begin{bmatrix} x \\ 0 \end{bmatrix}\right) = \begin{bmatrix} f(x) \\ 0 \end{bmatrix} \quad \text{for} \quad x \geq 0$$

for $F = F_m \circ F_{m-1} \circ \cdots \circ F_1 \circ F_0$. □

A.4 Bending the Arc

Let us next consider how to "bend" a one dimensional subspace into an m-arc. This iterative procedure is illustrated in Figure 6. It starts with all vertices in a straight line on the x_2-axis. First the polygonal chain is rotated by $-\phi_m$ ("to the right") and translated in negative x_1 direction ("to the left") such that the second to last vertex lies on the x_2-axis. Then a ReLU is applied, which projects all vertices up until the second to last onto the x_2-axis. The last last arc segment now has the correct angle and the process is repeated until all segments are correctly "bent". Note that each projection shrinks the unbent line segments by a factor of $\cos(\phi_m)$. To describe this transformation formally, we denote rotation matrices in $\mathbb{R}^2$ by

$$R_\alpha = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}.$$

Furthermore, for $m \geq 2$ and $0 = a_0 < a_1 < \cdots < a_m < \infty$ we define

$$\mathbf{d}_j = \begin{bmatrix} \sin(\phi_m)\cos(\phi_m)^j a_{m-j} \\ 0 \end{bmatrix}, \quad \text{for} \quad j = 1, \ldots, m,$$

as well as

$$G_0 \colon \mathbb{R}^2 \to \mathbb{R}^2 \colon \mathbf{x} \mapsto \varrho\left(R_{\frac{\pi}{2}} \mathbf{x}\right)$$
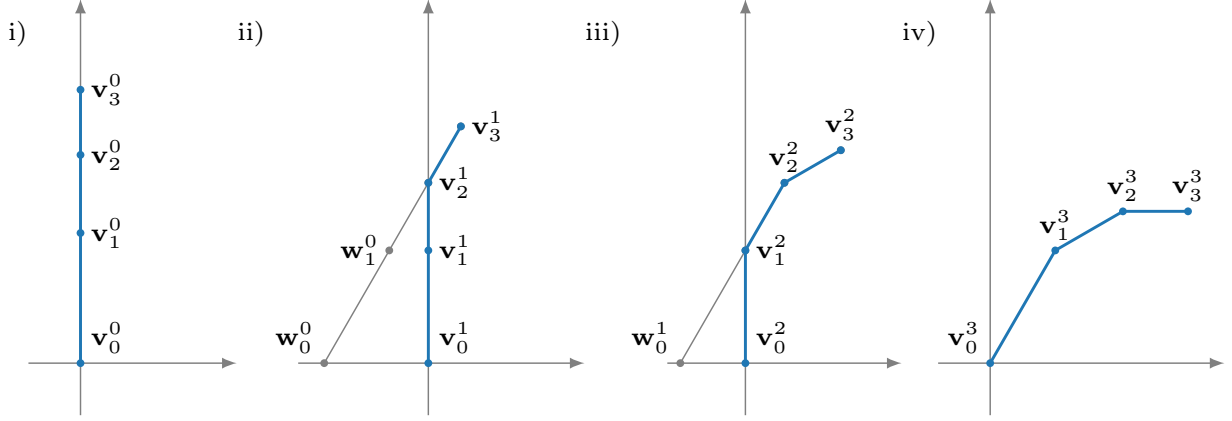
Figure 6: Illustration of the 'bending' of a 3-arc in 4 steps. Step (i) shows G_0 (90° rotation) applied to $[a_m, a_0] \times \{0\}$. Steps (ii) and (iii) show the image of $G_1 \circ G_0$ and $G_2 \circ G_1 \circ G_0$ respectively (blue) as well as the corresponding transform just before the last ReLU activation (gray). Step (iv) shows the final arc arising as the image of $G = G_3 \circ G_2 \circ G_1 \circ G_0$.

and

$$G_j \colon \mathbb{R}^2 \to \mathbb{R}^2 \colon \mathbf{x} \mapsto \varrho(R_{-\phi_m}\mathbf{x} - \mathbf{d}_j) \quad \text{for} \quad j = 1, \ldots, m,$$

which clearly are ReLU transformations.

Lemma A.5. *Let $m \geq 2$ and $0 = a_0 < a_1 < \cdots < a_m < \infty$ and G_j as above. Then*

$$G \colon [a_0, a_m] \times \{0\} \subseteq \mathbb{R}^2 \to \mathbb{R}^2 \colon \mathbf{x} \mapsto (G_m \circ G_{m-1} \circ \cdots \circ G_1 \circ G_0)(\mathbf{x})$$

parametrises an m-arc with vertices $\mathbf{v}_i = G\left(\begin{bmatrix} a_i \\ 0 \end{bmatrix}\right)$, line segments $\ell_i = G([a_{i-1}, a_i] \times \{0\})$, and scaling factors $r_i = \cos^{m-i}(\phi_m)(a_i - a_{i-1})$ for $i = 1, \ldots, m$.

Proof. We need to show that G is (in its first component) a piecewise-linear function with breakpoints a_i, and that $\mathbf{v}_0 = \mathbf{0}$ as well as

$$\mathbf{v}_i = \mathbf{v}_{i-1} + \cos^{m-i}(\phi_m)(a_i - a_{i-1})\begin{bmatrix} \sin(i\phi_m) \\ \cos(i\phi_m) \end{bmatrix}, \quad \text{for} \quad i = 1, \ldots, m. \tag{5}$$

For this we denote

$$\mathbf{v}_i^j = (G_j \circ \cdots \circ G_0)\left(\begin{bmatrix} a_i \\ 0 \end{bmatrix}\right) \quad \text{for} \quad i, j = 0, \ldots, m.$$

By the choice of ϕ_m we know $\sin(\phi_m) > 0$ and $\cos(\phi_m) > 0$. Hence, the first component of the bias vectors $\mathbf{d}_j$ satisfy $\sin(\phi_m)\cos(\phi_m)^j a_{m-j} \geq 0$ for $j = 1, \ldots, m$. Therefore, $\mathbf{v}_0^j = \mathbf{0}$ for all j and in particular $\mathbf{v}_0 = \mathbf{v}_0^m = \mathbf{0}$. Further, we will show inductively over j that for any i, j we have

$$\mathbf{v}_i^j - \mathbf{v}_{i-1}^j = \begin{cases} \cos(\phi_m)^j(a_i - a_{i-1})\begin{bmatrix} 0 \\ 1 \end{bmatrix}, & i \leq m - j, \\ \cos(\phi_m)^{m-i}(a_i - a_{i-1})\begin{bmatrix} \sin((i+j-m)\phi_m) \\ \cos((i+j-m)\phi_m) \end{bmatrix}, & i > m - j. \end{cases} \tag{6}$$

Then (5) follows from (6) with $j = m$.

To start the inductive proof we observe that for $j = 0$ and $i = 1, \ldots, m$ we have

$$
\begin{aligned}
\mathbf{v}_i^0 - \mathbf{v}_{i-1}^0 &= G_0\left(\begin{bmatrix} a_i \\ 0 \end{bmatrix}\right) - G_0\left(\begin{bmatrix} a_{i-1} \\ 0 \end{bmatrix}\right) \\
&= R_{\frac{\pi}{2}}\begin{bmatrix} a_i \\ 0 \end{bmatrix} - R_{\frac{\pi}{2}}\begin{bmatrix} a_{i-1} \\ 0 \end{bmatrix} \\
&= (a_i - a_{i-1})\begin{bmatrix} 0 \\ 1 \end{bmatrix},
\end{aligned}
$$

satisfying (6).

Next, let us assume that (6) holds for some $j < m$. For any i we denote $\mathbf{w}_i^j = R_{-\phi_m}\mathbf{v}_i^j - \mathbf{d}_j$ and therefore $\mathbf{v}_i^{j+1} = G_{j+1}(\mathbf{v}_i^j) = \varrho(\mathbf{w}_i^j)$ and

$$
\mathbf{w}_i^j - \mathbf{w}_{i-1}^j = \begin{cases} \cos(\phi_m)^j (a_i - a_{i-1})\begin{bmatrix} \sin(\phi_m) \\ \cos(\phi_m) \end{bmatrix}, & i \leq m - j, \\[3mm] \cos(\phi_m)^{m-i}(a_i - a_{i-1})\begin{bmatrix} \sin((i+j+1-m)\phi_m) \\ \cos((i+j+1-m)\phi_m) \end{bmatrix}, & i > m - j. \end{cases}
$$

We also have $\sin((i+j+1-m)\phi_m) > 0$ and $\cos((i+j+1-m)\phi_m) \geq 0$ for $i > m - j$ by the choice of ϕ_m and therefore $\mathbf{w}_m^j \geq \mathbf{w}_{m-1}^j \geq \cdots \geq \mathbf{w}_1^j \geq \mathbf{w}_0^j = -\mathbf{d}_j$ component-wise, where even all inequalities are strict in the first component. We immediatley see that all $\mathbf{w}_i^j$ lie in the upper half-plane. Further,

$$
\begin{aligned}
\mathbf{w}_{m-j}^j &= \mathbf{w}_0^j + \sum_{i=1}^{m-j}(\mathbf{w}_i^j - \mathbf{w}_{i-1}^j) \\
&= -\mathbf{d}_j + \sum_{i=1}^{m-j}\cos(\phi_m)^j (a_i - a_{i-1})\begin{bmatrix} \sin(\phi_m) \\ \cos(\phi_m) \end{bmatrix} \\
&= -\begin{bmatrix} \sin(\phi_m)\cos(\phi_m)^j a_{m-j} \\ 0 \end{bmatrix} + \cos(\phi_m)^j a_{m-j}\begin{bmatrix} \sin(\phi_m) \\ \cos(\phi_m) \end{bmatrix} \\
&= \cos(\phi_m)^{j+1}a_{m-j}\begin{bmatrix} 0 \\ 1 \end{bmatrix},
\end{aligned}
$$

from which we can conclude that

$$(\mathbf{w}_i^j)_1 \geq 0, \quad \text{for} \quad i = 1, \ldots, m \tag{7}$$

$$(\mathbf{w}_i^j)_2 < 0, \quad \text{for} \quad i = 1, \ldots, m - j - 1, \tag{8}$$

$$(\mathbf{w}_i^j)_2 \geq 0, \quad \text{for} \quad i = m - j, \ldots, m. \tag{9}$$

Now $\mathbf{v}_i^{j+1} = \varrho(\mathbf{w}_i^j)$ together with (7) to (9) implies

$$
\begin{aligned}
(\mathbf{v}_i^{j+1})_1 &= (\mathbf{w}_i^j)_1, \quad \text{for} \quad i = 1, \ldots, m, \\
(\mathbf{v}_i^{j+1})_2 &= 0, \quad \text{for} \quad i = 1, \ldots, m - j - 1, \\
(\mathbf{v}_i^{j+1})_2 &= (\mathbf{w}_i^j)_2, \quad \text{for} \quad i = m - j, \ldots, m,
\end{aligned}
$$

and therefore

$$
\mathbf{v}_i^{j+1} - \mathbf{v}_{i-1}^{j+1} = \begin{cases} \cos(\phi_m)^{j+1}(a_i - a_{i-1})\begin{bmatrix} 0 \\ 1 \end{bmatrix}, & i \leq m - (j+1), \\[3mm] \cos(\phi_m)^{m-i}(a_i - a_{i-1})\begin{bmatrix} \sin((i+j+1-m)\phi_m) \\ \cos((i+j+1-m)\phi_m) \end{bmatrix}, & i > m - (j+1), \end{cases}
$$

which shows that (6) holds for $j + 1$ and concludes the inductive step.

Each G_{j+1} adds only one new breakpoint at to the overall function, corresponding to point $\mathbf{w}_{m-j}^j$ where the second component of $(G_j \circ \cdots \circ G_0)(\mathbf{x})$ switches sign. This corresponds to the breakpoint a_{m-j} of $\tilde{G}$. $\qquad\square$

We are now ready to give the main proof of this section.

Proof of Lemma 3.6. Let $m \in \mathbb{N}$ be arbitrary. By assumption there exists a measure $\mu \in \{p(\theta)\}_{\theta \in \Omega}$ with infinite support. As discussed in Appendix A.1 and in particular by using Lemma A.2 we can without loss of generality assume $d = 2$ and $\mathrm{supp}(\mu)$ is compact and contained in the non-negative part of $\mathrm{span}\{\mathbf{e}_1\}$. In other words we can rewrite $\mathrm{supp}(\mu) = S \times \{0\} \subseteq [0, b] \times \{0\}$ for an infinite set $S \subseteq \mathbb{R}_{\geq 0}$ and some $b > 0$. By Lemma A.3 we can choose points $0 = b_0 < b_1 < \cdots < b_m = b$ so that S intersects each (b_{j-1}, b_j). By the choice of S and since $(b_{j-1}, b_j) \times \{0\}$ is relatively open in $S \times \{0\}$ we have $\mu((b_{j-1}, b_j) \times \{0\}) > 0$ for all j. Set $0 < \delta \leq \min_j \mu((b_{j-1}, b_j) \times \{0\})$.

Now let $A \in \mathcal{A}_m$ be an arbitrary standard m-arc and denote its set of vertices by $(\mathbf{v}_0, \ldots, \mathbf{v}_m) = \mathrm{vert}(A)$, its line segments by $(\ell_1, \ldots, \ell_m) = \mathrm{segm}(A)$, and its scaling factors by $(r_1, \ldots, r_m) = \mathrm{scal}(A)$. We set $a_0 = 0$ and iteratively $a_j = a_{j-1} + r_j \cos^{j-m}(\phi_m)$ for $j = 1, \ldots, m$. By Lemma A.4 there is a ReLU neural network F satisfying

$$F\left(\begin{bmatrix} b_j \\ 0 \end{bmatrix}\right) = \begin{bmatrix} a_j \\ 0 \end{bmatrix} \quad \text{for} \quad j = 1, \ldots, m,$$

and then by Lemma A.5 another ReLU neural network G satisfying

$$G\left(\begin{bmatrix} a_i \\ 0 \end{bmatrix}\right) = \mathbf{v}_j \quad \text{for} \quad j = 1, \ldots, m.$$

Altogether $G \circ F$ is a ReLU neural network transforming $\mathbf{b}_j = \begin{bmatrix} b_i & 0 \end{bmatrix}^\top$ to the vertex $\mathbf{v}_j$ and $[b_{j-1}, b_j] \times \{0\}$ to the line segment ℓ_j. By Lemma A.1 we have

$$\mathrm{supp}((G \circ F)_* \mu) = \overline{(G \circ F)(\mathrm{supp}(\mu))} \subseteq \overline{(G \circ F)([b_0, b_m] \times \{0\})} \subseteq A.$$

Also, for each j we get

$$((G \circ F)_* \mu)(\ell_j) = \mu\big((G \circ F)^{-1}(\ell_j)\big) \geq \mu((b_{j-1}, b_j) \times \{0\}) \geq \delta$$

and therefore $(G \circ F)_* \mu \in \mathcal{D}_{\delta,m}(\mathbb{R}^2)$ and $\mathrm{arc}((G \circ F)_* \mu) = A$. Using the ReLU-invariance, we know that $(G \circ F)_* \mu \in \{p(\theta)\}_{\theta \in \Omega_{\delta,m}} \subseteq \{p(\theta)\}_{\theta \in \Omega}$. Altogether, since $A \in \mathcal{A}_m$ was arbitrary we conclude that $\mathrm{arc} \circ p|_{\Omega_{\delta,m}}$ is surjective onto $\mathcal{A}_m$. Clearly, $\mathrm{scal} \colon \mathcal{A}_m \to \mathbb{R}_+^{m-1} \times \{1\}$ is a bijection, hence also $\Xi \colon \Omega_{\delta,m} \to \mathbb{R}_+^{m-1} \times \{1\}$ is surjective. $\qquad\square$

B   Proof of Lemma 3.7: Local Lipschitz Continuity of Ξ

We will now show that the map $\Xi \colon \Omega_{\delta,m} \to \mathbb{R}_+^m$ is locally Lipschitz continuous by showing this for each of its three composite parts $p|_{\Omega_{\delta,m}} \colon \Omega_{\delta,m} \to \mathcal{D}_{\delta,m}(\mathbb{R}^d)$, $\mathrm{arc} \colon \mathcal{D}_{\delta,m}(\mathbb{R}^d) \to \mathcal{A}_m$, and $\mathrm{scal} \colon \mathcal{A}_m \to \mathbb{R}_+^m$.

Firstly, $p|_{\Omega_{\delta,m}} \colon \Omega_{\delta,m} \to \mathcal{D}_{\delta,m}(\mathbb{R}^d)$ is locally Lipschitz continuous by Assumption 3.1, so there is nothing to show.

Secondly, the local Lipschitz continuity of $\mathrm{arc} \colon \mathcal{D}_{\delta,m}(\mathbb{R}^d) \to \mathcal{A}_m$ can be derived from geometric observations.

Let us recall the definition of the Prokhorov metric (Prokhorov, 1956):

$$d_P(\mu, \nu) = \inf\{\epsilon > 0 \,:\, \mu(B) \leq \nu(B^\epsilon) + \epsilon \text{ and } \nu(B) \leq \mu(B^\epsilon) + \epsilon \text{ for any } B \in \mathcal{B}(\mathbb{R}^d)\},$$

where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel σ-Algebra on $\mathbb{R}^d$ and $B^\epsilon = \{\, x \in \mathbb{R}^d \,:\, \exists y \in B \text{ with } \|x - y\| < \epsilon \,\}$ is the open ϵ-neighbourhood of a Borel set $B \in \mathcal{B}(\mathbb{R}^d)$.[3]

Lemma B.1. *The map* $\mathrm{arc} \colon (\mathcal{D}_{\delta,m}(\mathbb{R}^d), d_P) \to (\mathcal{A}_m, d_{\mathcal{A}}) \colon \mu \mapsto A_\mu$, *is locally Lipschitz continuous with Lipschitz constant* $\frac{2\sqrt{2}}{\sin(\phi_m)}$.

Proof. Let $\mu_0 \in \mathcal{D}_{\delta,m}(\mathbb{R}^d)$ be arbitrary. We will show Lipschitz continuity on the open ball of radius $\frac{\delta}{2}$ around μ_0. For this let $\mu_1, \mu_2 \in B_{\frac{\delta}{2}}(\mu_0) \cap \mathcal{D}_{\delta,m}(\mathbb{R}^d)$ and therefore $d_P(\mu_1, \mu_2) \leq \delta$. We know that μ_1 and μ_2 are δ-supported on two unique standard m-arcs $A_1 = \mathrm{arc}(\mu_1) = A_{\mu_1} \in \mathcal{A}_m$ and $A_2 = \mathrm{arc}(\mu_2) = A_{\mu_2} \in \mathcal{A}_m$ and by definition

---

[3]The original definition by Prokhorov only considers closed sets B, however this results in the same metric.

these lie in the same two-dimensional subspace $\text{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ of $\mathbb{R}^d$. So without loss of generality we carry out the remaining proof in $\mathbb{R}^2$. Let $(\ell_1^1, \ldots, \ell_m^1) = \text{segm}(A_1)$ and $(\ell_1^2, \ldots, \ell_m^2) = \text{segm}(A_2)$ denote the line segments of A_1 and A_2 respectively. We denote the affine hulls of individual line segments as $h_i^j = \text{aff}\left(\ell_i^j\right)$ and further denote the Euclidean distance of two corresponding affine hulls by $d_i = \text{dist}(h_i^1, h_i^2)$. An example for this can be seen in Figure 7.

We want to upper bound the Euclidean distances $\|\mathbf{v}_i^1 - \mathbf{v}_i^2\|_2$ of all corresponding vertices of the two arcs. For any i one of four cases can occur, as visualised in Figure 8. Firstly, if $\mathbf{v}_i^1 = \mathbf{v}_i^2$ (cf. Figure 8, top left) we trivially get $\|\mathbf{v}_i^1 - \mathbf{v}_i^2\|_2 = 0$. Secondly, if $\mathbf{v}_i^1 \neq \mathbf{v}_i^2$ but $h_i^1 = h_i^2$ (cf. Figure 8, top right) it is not hard to see that

$$\|\mathbf{v}_i^1 - \mathbf{v}_i^2\|_2 \leq \frac{d_{i+1}}{\sin(\phi_m)}.$$

Thirdly, if $\mathbf{v}_i^1 \neq \mathbf{v}_i^2$ but $h_{i+1}^1 = h_{i+1}^2$ (cf. Figure 8, bottom left) we similarly get

$$\|\mathbf{v}_i^1 - \mathbf{v}_i^2\|_2 \leq \frac{d_i}{\sin(\phi_m)}.$$

In the fourth case, where $\mathbf{v}_i^1 \neq \mathbf{v}_i^2$, $h_i^1 \neq h_i^2$, and $h_{i+1}^1 \neq h_{i+1}^2$ the four intersections of the affine spaces form a parallelogram (cf. Figure 8, bottom right). We obtain

$$\|\mathbf{v}_i^1 - \mathbf{v}_i^2\|_2 \leq \frac{\sqrt{2}(d_i + d_{i+1})}{\sin(\phi_m)},$$

using the parallelogram identity. Thus, it remains to bound all the distances d_i.

For this, assume that for some i we have $d_i > 0$. We say that ℓ_i^1 is the *outer* line segment and ℓ_i^2 is the *inner* line segment if $\text{dist}\left(\mathbf{0}, h_i^1\right) > \text{dist}\left(\mathbf{0}, h_i^2\right)$ and vice versa if $\text{dist}\left(\mathbf{0}, h_i^1\right) < \text{dist}\left(\mathbf{0}, h_i^2\right)$. Without loss of generality assume that ℓ_i^1 is the outer line segment. Then for any point $\mathbf{p} \in A_2$ we have $\text{dist}(\mathbf{p}, \ell_i^1) \geq d_i$. Thus $\mu_2((\ell_i^1)^\epsilon) = 0$ for any $\epsilon < d_i$. But $\mu_1(\ell_i^1) \geq \delta$ and therefore by definition of the Prokhorov metric

$$d_i \leq d_P(\mu_1, \mu_2).$$

Altogether, we obtain

$$d_{\mathcal{A}}(A_1, A_2) = \max_i \|\mathbf{v}_i^1 - \mathbf{v}_i^2\|_2 \leq \frac{2\sqrt{2}}{\sin(\phi_m)} d_P(\mu_1, \mu_2). \qquad \square$$

Thirdly, the local Lipschitz continuity of $\text{scal} \colon \mathcal{A}_m \to \mathbb{R}_+^m$ is a straight-forward calculation if $\mathbb{R}_+^m$ is equipped with the metric induced by the ℓ_∞-norm. By norm equivalence the same also holds for $\mathbb{R}_+^m$ viewed as a subspace of Euclidean $\mathbb{R}^m$.

Lemma B.2. *The map* $\text{scal} \colon (\mathcal{A}_m, d_{\mathcal{A}}) \to (\mathbb{R}_+^m, \|\cdot\|_\infty) \colon A \mapsto (r_1, \ldots, r_m)$, *is Lipschitz continuous with Lipschitz constant 2.*

Proof. Let $A_1, A_2 \in \mathcal{A}_m$ be arcs with vertices $(\mathbf{v}_1^1, \ldots, \mathbf{v}_m^1) = \text{vert}(A_1)$ and $(\mathbf{v}_1^2, \ldots, \mathbf{v}_m^2) = \text{vert}(A_2)$ and scaling factors $(r_1^1, \ldots, r_m^1) = \text{scal}(A_1)$ as well as $(r_1^2, \ldots, r_m^2) = \text{scal}(A_2)$ respectively. Since for all $i = 1, \ldots, m$ we have

$$r_i^1 \leq r_i^2 + \text{dist}\left(\mathbf{v}_{i-1}^1, \mathbf{v}_{i-1}^2\right) + \text{dist}\left(\mathbf{v}_i^1, \mathbf{v}_i^2\right)$$

and vice versa

$$r_i^2 \leq r_i^1 + \text{dist}\left(\mathbf{v}_{i-1}^1, \mathbf{v}_{i-1}^2\right) + \text{dist}\left(\mathbf{v}_i^1, \mathbf{v}_i^2\right),$$

we obtain

$$\|\text{scal}(A_1) - \text{scal}(A_2)\|_\infty = \max_{i=1,\ldots,m} \left|r_i^1 - r_i^2\right| \leq 2 d_{\mathcal{A}}(A_1, A_2). \qquad \square$$

Altogether we can conclude this section by combining all pieces and proving the local Lipschitz continuity of $\Xi \colon \Omega_{\delta,m} \to \mathbb{R}_+^m$.

Proof of Lemma 3.7. The claim follows directly from Assumption 3.1 and Lemmas B.1 and B.2 since the composition of locally Lipschitz continuous functions is again locally Lipschitz continuous. $\qquad \square$
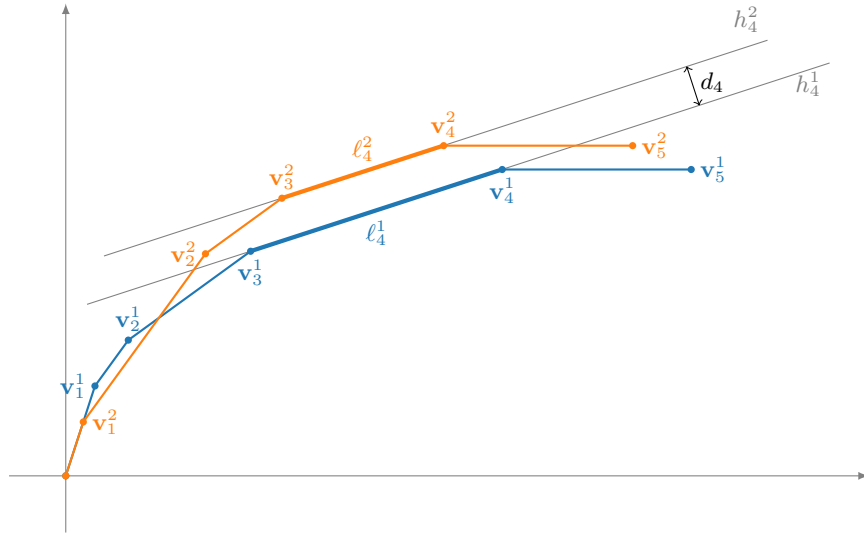
Figure 7: Example of two standard m-arcs A_1 (blue) and A_2 (orange) for $m = 5$. Supporting affine subspaces h_4^1 and h_4^2 and their distance d_4 are shown exemplarily for the fourth line segments. At this segment A_2 is the outer arc and A_1 is the inner arc.
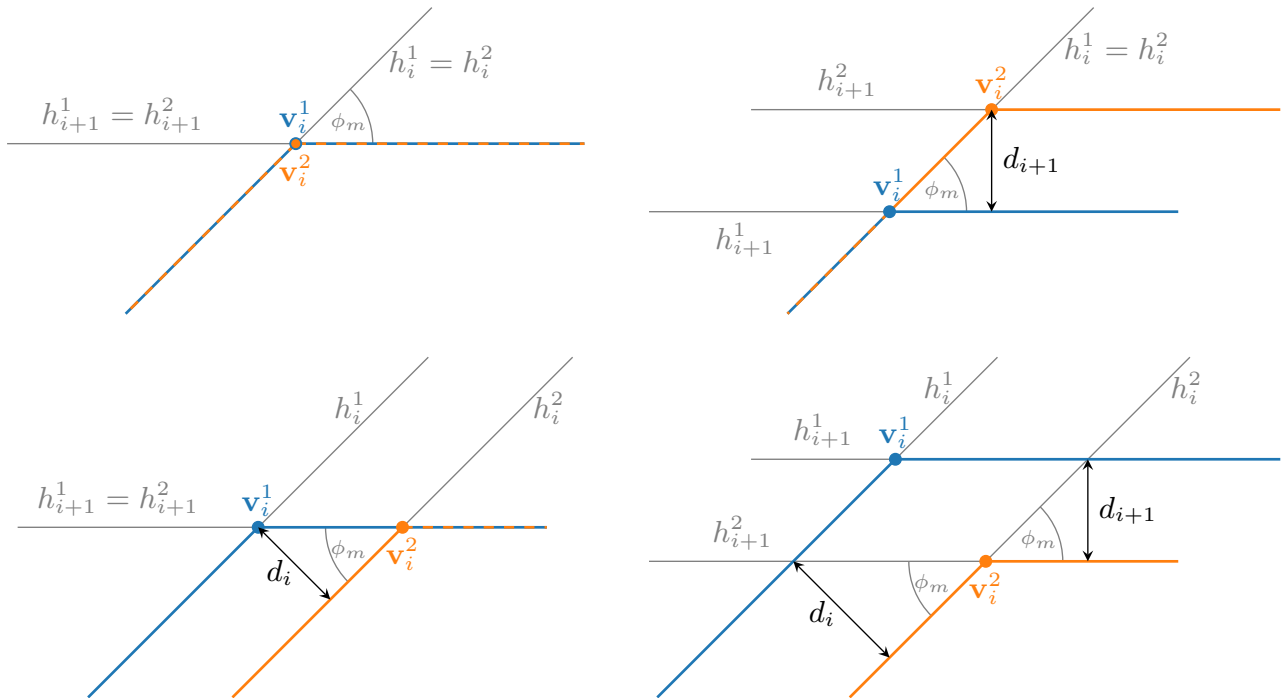

Figure 8: Four possible arrangements of corresponding vertices $\mathbf{v}_i^1$ and $\mathbf{v}_i^2$ of two arcs. In case one both vertices are equal (top left). In cases two and three the vertices are different but either the affine subspaces h_i^1 and h_i^2 or h_{i+1}^1 and h_{i+1}^2 coincide (top right and bottom left). In the fourth case all affine subspaces differ and their intersections form a parallelogram with $\mathbf{v}_i^1$ and $\mathbf{v}_i^2$ at opposite corners (bottom right).

C   Families of Distributions in One Dimension

In this section we will give a prove of the first part of Theorem 2.7 corresponding to (R1). For this we explicitly construct a family of ReLU-invariant probability distributions on $\mathbb{R}$ with a locally Lipschitz continuous parametrisation map.

We proceed in three steps. We first analyse the set of one-dimensional ReLU networks and show that it can be completely described by only six parameters. We then use these to parametrise all ReLU neural networks as functions in $C(K, \mathbb{R})$ on some compact domain $K \subseteq \mathbb{R}$ via a so called *realisation* map $R \colon \mathbb{R}^6 \to C(K, \mathbb{R})$. Finally, we use a pushforward map $Q \colon C(K, \mathbb{R}) \to \mathcal{D}(\mathbb{R})$ mapping neural network functions $f \in C(K, \mathbb{R})$ to the pushforward of a fixed *prototype* measure $\mu_0 \in \mathcal{D}(\mathbb{R})$ under f to generate the family of probability distribution. The prototype measure is assumed to be supported within K. The one-parameter family $p \colon \mathbb{R}^6 \to \mathcal{D}(\mathbb{R})$ is then simply given as $p = Q \circ R$.

We will follow a similar idea to obtain ReLU-invariant families in higher dimension when proving the third part of Theorem 2.7 corresponding to (R3) in Appendix D. However in this case the number of parameters to describe the neural networks grows with their depth leading to the need for an arbitrary large number of parameters to describe the distributions in the family. Although it is possible to generalise our construction to higher dimensions this comes at the cost of losing the local Lipschitz continuity of the parametrisation map.

One-dimensional ReLU networks: Let us first show that any one dimensional ReLU network can be rewritten as a three layer network. For this let $L \in \mathbb{N}$ and $f \colon \mathbb{R} \to \mathbb{R}$ be an L-layer ReLU network given as

$$f = f_L \circ f_{L-1} \circ \cdots \circ f_2 \circ f_1$$

with layers

$$f_i \colon \mathbb{R} \to \mathbb{R} \colon x \mapsto \varrho(w_i x + b_i) \quad \text{for} \quad i = 1, \dots, L.$$

We observe that f is constant as soon as any weight w_i is zero and a constant function can easily be rewritten as a three layer network. So from now on we assume $w_i \neq 0$ for all layers. We denote by

$$A_i = \{\, x \in \mathbb{R} \,:\, w_i x + b_i \geq 0 \,\} \quad \text{for} \quad i = 1, \dots, L$$

the domain on which the i-th layer is affine linear with slope w_i. It is constant on the complement A_i^c. Further we denote

$$B_i = (f_{i-1} \circ \cdots \circ f_1)^{-1}(A_i) \quad \text{for} \quad i = 2, \dots, L$$

and $B_1 = A_1$. Each A_i is a convex and closed subset of $\mathbb{R}$ and $f_{i-1} \circ \cdots \circ f_1$ is continuous and monotone as a composition of continuous and monotone functions. Hence, also all B_i are closed and convex. In $\mathbb{R}$ these are exactly the (possibly unbounded) closed intervals.

Lemma C.1. *Let $U \subseteq \mathbb{R} \setminus \bigcap_{i=1}^{L} B_i$ be connected. Then $f|_U$ is constant.*

Proof. We have $\mathbb{R} \setminus \bigcap_{i=1}^{L} B_i = \bigcup_{i=1}^{L} B_i^c$ and therefore $U = \bigcup_{i=1}^{L} U \cap B_i^c$. Since B_i is closed B_i^c is open and thus $U \cap B_i^c$ is relatively open in U. But by the choice of B_i we know $(f_i \circ \cdots \circ f_1)(U \cap B_i^c) \equiv 0$, hence $f|_{U \cap B_i^c}$ is constant. But since f is continuous and U is connected this already implies that $f|_U$ is constant. $\qquad \square$

The one-dimensional ReLU network f is piecewise constant except for the (possibly empty) region $\bigcap_{i=1}^{L} B_i$ where all layers are non constant and act affine linearly. On this region f is affine linear with slope $\prod_{i=1}^{L} w_i$. As discussed above each B_i is convex and closed, hence the same holds for their intersection. Two different cases can occur, which determine the overall shape of the function f. This is also visualised in Figure 4 of the main paper.

Firstly, if $\bigcap_{i=1}^{L} B_i$ is bounded it must be empty, a single point, or a bounded and closed interval. Then f is bounded. In fact it is constant or a piecewise linear step function with one step and two constant regions. These functions can be represented by three layer networks.

Secondly, if $\bigcap_{i=1}^{L} B_i$ is an unbounded closed interval, then f looks like a shifted and scaled ReLU function and can again be represented by a three layer ReLU network.

Lemma C.2. *The ReLU network f has one of the following forms.*

- *f is bounded and has two constant and one affine linear region:*

$$f(x) = \begin{cases} c_1, & x \leq a_1 \\ c_2, & x \geq a_2 \\ c_1 + \frac{c_2 - c_1}{a_2 - a_1}(x - a_1), & a_1 < x < a_2 \end{cases}$$

for some $c_1, c_2 \geq 0$ and $a_1 < a_2$. This includes the degenerate case where f is constant by choosing $c_1 = c_2$.

- *f is unbounded and has one constant and one affine linear region:*

$$f(x) = \begin{cases} c, & w(x - a) \leq 0 \\ c + w(x - a), & w(x - a) > 0 \end{cases}$$

for some $c \geq 0$, $a \in \mathbb{R}$, and $w \neq 0$.

Proof. This directly follows from Lemma C.1, the continuity and monotonicity of the ReLU network f and the fact that f is lower bounded by zero. $\square$

Lemma C.3. *The two types of piecewise linear functions from Lemma C.2 can both be written as one dimensional ReLU networks with three layers.*

Proof. We give explicit formulas for choosing the weights w_1, w_2, w_3 and biases b_1, b_2, b_3 of the three layer network.

In the first case of a bounded function we choose

$$w_1 = 1, \qquad\qquad b_1 = -a_1,$$
$$w_2 = -\frac{|c_2 - c_1|}{a_2 - a_1}, \qquad\qquad b_2 = |c_2 - c_1|,$$
$$w_3 = -\operatorname{sign}(c_2 - c_1), \qquad\qquad b_3 = c_2,$$

and in the second case of an unbounded function we choose

$$w_1 = \operatorname{sign}(w), \qquad b_1 = -\operatorname{sign}(w)a,$$
$$w_2 = |w|, \qquad b_2 = c,$$
$$w_3 = 1, \qquad b_3 = 0.$$

The claim now follows by straightforward calculations. $\square$

Neural Network Parametrisation: Neural networks can be viewed algebraically (as collections of weights and biases) or analytically (as functions in some function space). This distinction was made by Petersen and Voigtlaender (2018) and used by Petersen et al. (2018) to study the dependence of neural network functions on their weights and biases. The mapping from weights and biases to the corresponding neural network function is referred to as the *realisation* map. Petersen et al. (2018) show the continuity of the realisation map for neural networks of fixed size with continuous activation function (in particular this includes the ReLU activation), when they are viewed as functions in $C(K, \mathbb{R}^d)$ on a compact domain K. If the activation function is Lipschitz continuous (which again is the case for the ReLU activation) also the realisation map is Lipschitz continuous.

As we have seen all ReLU neural networks in one dimension can be rewritten into a three layer network and can thus be parametrised by three weight and three bias parameters. Let $K \subseteq \mathbb{R}$ be any compact non-empty domain. For any collection of weights and biases $\theta = (w_1, b_1, w_2, b_2, w_3, b_3) \in \mathbb{R}^6$ we denote the function that is the ReLU neural network realisation of these weights and biases as $f_\theta \colon K \to \mathbb{R}$. Then the realisation map is

$$R \colon \mathbb{R}^6 \to C(K, \mathbb{R}^d) \colon \theta \mapsto f_\theta.$$

Petersen et al. (2018, Proposition 5.1) show the following result.

Lemma C.4. *The realisation map $R \colon \mathbb{R}^6 \to C(K, \mathbb{R}^d)$ is Lipschitz continuous.*

Pushforward Mapping: We want to connect neural network functions with the effect they have on probability measures and obtain an invariant family of distributions. For this let $\mu_0 \in \mathcal{D}(\mathbb{R})$ be any probability measure on $\mathbb{R}$ with $\mathrm{supp}(\mu_0) \subseteq K$. We call this the *prototype* measure and will derive all other measures in the family as pushforwards of μ_0 under ReLU neural networks. We define the pushforward function

$$Q \colon C(K, \mathbb{R}) \to \mathcal{D}(\mathbb{R}) \colon f \mapsto f_* \mu_0.$$

The restriction of Q to the subset of ReLU neural network functions will be used to generate the invariant family of distributions.

Lemma C.5. *The pushforward function $Q \colon C(K, \mathbb{R}) \to \mathcal{D}(\mathbb{R})$ is Lipschitz continuous.*

Proof. Let $f, g \in C(K, \mathbb{R})$ and $A \in \mathcal{B}(\mathbb{R})$. Assuming $f^{-1}(A) \neq \emptyset$, for any $x \in f^{-1}(A)$ we have $f(x) \in A$ and $|f(x) - g(x)| \leq \|f - g\|_\infty$. Consequently, for any $\epsilon > \|f - g\|_\infty$ we get $g(x) \in A^\epsilon$ and therefore $x \in g^{-1}(A^\epsilon)$. Thus $f^{-1}(A) \subseteq g^{-1}(A^\epsilon)$. Conversely, if $f^{-1}(A) = \emptyset$ then the same inclusion trivially holds. Analogously $g^{-1}(A) \subseteq f^{-1}(A^\epsilon)$ can be shown for any $\epsilon > \|f - g\|_\infty$. This directly implies

$$\begin{aligned}
d_P(Q(f), Q(g)) &= d_P(f_* \mu_0, g_* \mu_0) \\
&= \inf \big\{ \, \epsilon > 0 \, : \, f_* \mu_0(A) \leq g_* \mu_0(A^\epsilon) + \epsilon \text{ and} \\
&\qquad\qquad g_* \mu_0(A) \leq f_* \mu_0(A^\epsilon) + \epsilon \text{ for any } A \in \mathcal{B}(\mathbb{R}) \, \big\} \\
&= \inf \big\{ \, \epsilon > 0 \, : \, \mu_0(f^{-1}(A)) \leq \mu_0(g^{-1}(A^\epsilon)) + \epsilon \text{ and} \\
&\qquad\qquad \mu_0(g^{-1}(A)) \leq \mu_0(f^{-1}(A^\epsilon)) + \epsilon \text{ for any } A \in \mathcal{B}(\mathbb{R}) \, \big\} \\
&\leq \|f - g\|_\infty. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square
\end{aligned}$$

ReLU-Invariance: We can now define the family of distributions as

$$p \colon \mathbb{R}^6 \to \mathcal{D}(\mathbb{R}) \colon \theta \mapsto (Q \circ R)(\theta).$$

The local Lipschitz continuity of p is clear from the previous steps. It remains to establish the ReLU-invariance.

Lemma C.6. *The family of distributions p is ReLU-invariant.*

Proof. Let $\theta \in \mathbb{R}^6$ be arbitrary and $p(\theta) \in \mathcal{D}(\mathbb{R})$ the corresponding probability measure. Further, let $f \colon \mathbb{R} \to \mathbb{R} \colon x \mapsto \varrho(wx + b)$ for some $w, b \in \mathbb{R}$ be any ReLU layer. We need to show that there exists a $\omega \in \mathbb{R}^6$ such that $p(\omega) = f_* p(\theta)$.

By construction $p(\theta) = (Q \circ R)(\theta) = (R(\theta))_* \mu_0 = (f_\theta)_* \mu_0$, where $f_\theta = R(\theta)$ is a ReLU neural network. Clearly also $f \circ f_\theta$ is a ReLU neural network, and since any ReLU network can be rewritten as a three-layer network, there exists $\omega \in \mathbb{R}^6$ such that $R(\omega) = f \circ f_\theta$. Finally,

$$p(\omega) = (f \circ f_\theta)_* \mu_0 = f_* (f_\theta)_* \mu_0 = f_* p(\theta)$$

yields the claim. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

D Families of Distributions Without Local Lipschitz Continuity

In this section we will prove the third part of Theorem 2.7 corresponding to (R3).

We define a family of distributions each of which can be described by a finite but arbitrarily large number of parameters. We use space-filling curves to fuse these arbitrarily many parameters into a single parameter.

The high-level proof idea is quite straightforward and analogous to Appendix C: If we have an arbitrarily large number of parameters available to describe each distribution in the family we can simply use these parameters to specify the weights of the neural network transformations with respect to which we want to obtain invariance. The challenge in the general d-dimensional setting is to show that this leads to a continuous parametrisation.

We start by clarifying what we mean by a finite but arbitrarily large number of parameters. We use the conventional notation $\mathbb{R}^\omega$ for the product of countably many copies of $\mathbb{R}$ equipped with the product topology. In other words, $\mathbb{R}^\omega$ is the space of all real-valued sequences. Further we denote by

$$\mathbb{R}^\infty = \big\{ \, (x_1, x_2, \ldots) \in \mathbb{R}^\omega \, : \, x_i \neq 0 \text{ for only finitely many } i \, \big\}$$

the subset of sequences that are eventually zero. The space $\mathbb{R}^\infty$ will serve as the parameter space. Each element in $\mathbb{R}^\infty$ effectively uses only a finite number of parameters (since all the remaining ones are zero), however this number can be arbitrarily large.

We proceed in three steps. We first introduce a one-parameter description $\Gamma\colon \mathbb{R} \to \mathbb{R}^\infty$ of the finitely-but-arbitrarily-many parameter set $\mathbb{R}^\infty$. We then use a subset $\Omega_\infty \subseteq \mathbb{R}^\infty$ to parametrise ReLU neural networks as functions in $C(K, \mathbb{R}^d)$ using a realisation map $R\colon \Omega_\infty \to C(K, \mathbb{R}^d)$. Finally, we use a pushforward map $Q\colon C(K, \mathbb{R}^d) \to \mathcal{D}(\mathbb{R}^d)$ mapping neural network functions $f \in C(K, \mathbb{R}^d)$ to the pushforward of a fixed *prototype* measure $\mu_0 \in \mathcal{D}(\mathbb{R}^d)$ under f to generate the family of probability distribution. As before the prototype measure is assumed to be supported within the domain K. The one-parameter family $p\colon \Omega \to \mathcal{D}(\mathbb{R}^d)$ is then given as $p = Q \circ R \circ \Gamma$ for some suitable domain Ω. We will now discuss each of the steps in more detail.

Space-Filling Curves: We want to continuously describe $\mathbb{R}^\infty$ with a single parameter. This can be achieved by gluing together continuous space-filling curves from the unit interval to cubes of arbitrary dimension. The existence of such maps is guaranteed by the Hahn-Mazurkiewicz theorem, see for example (Sagan, 1994, Chapter 6.8).

Lemma D.1. *There exists a continuous and surjective function $\Gamma\colon \mathbb{R} \to \mathbb{R}^\infty$.*

The construction of Γ is deferred to Appendix F.

Neural Network Parametrisation: We need to extend the realisation map from Appendix C to arbitrary dimensions and arbitrary numbers of layers. Further we will show that the results of Petersen et al. (2018) concerning continuous network parametrisations can be extended to networks of arbitrary depth.

A ReLU neural network in d dimensions is characterised by the number of its layers as well as $d^2 + d$ parameters for the weights and biases for each of the layers. Using the first component in $\mathbb{R}^\infty$ to encode the number of layers we can parametrise all such neural networks with the set

$$\Omega_\infty = \left\{ (L, x_1, x_2, \dots) \in \mathbb{R}^\infty \ : \ L \in \mathbb{N} \text{ and } x_i = 0 \text{ for all } i > L\big(d^2 + d\big) \right\} \subseteq \mathbb{R}^\infty.$$

It will be convenient to introduce notations for the subsets of parameters with a fixed number of layers. For $L \in \mathbb{N}$ we write $\Omega_L = \{L\} \times \mathbb{R}^{L(d^2+d)} \times \{0\}^\infty \subseteq \Omega_\infty$ and observe that

$$\Omega_\infty = \bigcup_{L \in \mathbb{N}} \Omega_L.$$

In fact this is the partition of Ω_∞ into its connected components.

For any $\theta = (L, x_1, x_2, \dots) \in \Omega_\infty$ we can regroup its leading non-zero components into alternating blocks of size d^2 and d and write

$$\theta = (L, \mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, \dots, \mathbf{W}_L, \mathbf{b}_L, 0, \dots)$$

with $\mathbf{W}_i \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_i \in \mathbb{R}^d$. As before let $K \subseteq \mathbb{R}^d$ be any non-empty compact domain and denote the function that is the ReLU neural network realisation of the weights $\{\mathbf{W}_1, \dots, \mathbf{W}_d\}$ and biases $\{\mathbf{b}_1, \dots, \mathbf{b}_d\}$ as $f_\theta\colon K \to \mathbb{R}^d$. Then we can define the extended realisation map

$$R\colon \Omega_\infty \to C(K, \mathbb{R}^d)\colon \theta \mapsto f_\theta.$$

Lemma D.2. *The extended realisation map $R\colon \Omega_\infty \to C(K, \mathbb{R}^d)$ is continuous.*

Proof. To show the continuity of $R\colon \Omega_\infty \to C(K, \mathbb{R}^d)$ it suffices to show the continuity on all connected components Ω_L of the domain. But for each fixed $L \in \mathbb{N}$ we know from the results of Petersen et al. (2018, Proposition 5.1) that the realisation map is continuous from Ω_L to $C(K, \mathbb{R}^d)$. This already proves the overall continuity of R. □

Pushforward Mapping: We need to extend our pushforward map from Appendix C to arbitrary dimension. As before let $\mu_0 \in \mathcal{D}(\mathbb{R}^d)$ be any probability measure on $\mathbb{R}^d$ with $\mathrm{supp}(\mu_0) \subseteq K$. We define the pushforward function analogous to before as

$$Q\colon C(K, \mathbb{R}^d) \to \mathcal{D}(\mathbb{R}^d)\colon f \mapsto f_*\mu_0.$$

The restriction of Q to the subset of ReLU neural network functions will be used to generate the invariant family of distributions.

Lemma D.3. *The pushforward function $Q: C(K, \mathbb{R}^d) \to \mathcal{D}(\mathbb{R}^d)$ is continuous.*

The proof works exactly as in the one-dimensional case in Appendix C.

ReLU-Invariance: We can now define the one-parameter family of distributions as

$$p: \Omega \to \mathcal{D}(\mathbb{R}^d): \theta \mapsto (Q \circ R \circ \Gamma)(\theta)$$

with the domain $\Omega = \Gamma^{-1}(\Omega_\infty)$ chosen so that Γ maps it exactly to the feasible neural network parameters Ω_∞ and thus $R \circ \Gamma$ maps it to all ReLU neural networks in d dimensions. The continuity of p is clear from the three previous steps. It remains to establish the ReLU-invariance.

Lemma D.4. *The one-parameter family of distributions p is ReLU-invariant.*

Proof. Let $\theta \in \Omega$ be arbitrary and $p(\theta) \in \mathcal{D}(\mathbb{R}^d)$ the corresponding probability measure. Further, let $f: \mathbb{R}^d \to \mathbb{R}^d: \mathbf{x} \mapsto \varrho(\mathbf{W}\mathbf{x} + \mathbf{b})$ for some $\mathbf{W} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ be any ReLU layer. We need to show that there exists a $\omega \in \Omega$ such that $p(\omega) = f_* p(\theta)$.

By construction we have $p(\theta) = (Q \circ R \circ \Gamma)(\theta) = (R(\Gamma(\theta)))_* \mu_0 = \widetilde{f}_* \mu_0$, where $\widetilde{f} = R(\Gamma(\theta))$ is a ReLU neural network. Clearly also $f \circ \widetilde{f}$ is a ReLU neural network, so there exists $\omega \in \Omega$ such that $R(\Gamma(\omega)) = f \circ \widetilde{f}$. Finally,

$$p(\omega) = (f \circ \widetilde{f})_* \mu_0 = f_* \widetilde{f}_* \mu_0 = f_* p(\theta)$$

yields the claim. $\square$

E Metric Spaces and Hausdorff Dimension

In this section we will proof Lemma 3.8. For this we will first review some concepts and results regarding general metric spaces and their Hausdorff dimension, which we denote by $\dim_H$. Applying these to Euclidean $\mathbb{R}^n$ or subspaces thereof will yield the desired result.

A topological space is called a Lindelöf space if every open cover of it has a countable subcover. This is weaker than compactness, where the existence of finite subcovers is required. For metric spaces the notions Lindelöf, separable, and second-countable are all equivalent. It is easy to see that any σ-compact space is Lindelöf. Subspaces of separable metric spaces are again separable. We start with a collection of useful properties of the Hausdorff dimension, see for example (Edgar, 2008, Chapter 6).

Lemma E.1. *Let (X, d_X) and (Y, d_Y) be metric spaces, $A, B \subseteq X$ Borel sets, $(A_i)_{i \in \mathbb{N}}$ a countable collection of Borel sets $A_i \subseteq X$, and $f: X \to Y$ (globally) Lipschitz continuous. Then*

(i) $A \subseteq B \Rightarrow \dim_H(A) \leq \dim_H(B)$,

(ii) $\dim_H(A \cup B) = \max\{\dim_H(A), \dim_H(B)\}$,

(iii) $\dim_H\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sup_{i \in \mathbb{N}}\{\dim_H(A_i)\}$,

(iv) $\dim_H(f(X)) \leq \dim_H(X)$.

If (X, d_X) is a separable space, then the last part can be generalised to locally Lipschitz continuous maps.

Lemma E.2. *Let (X, d_X) and (Y, d_Y) be metric spaces, $f: X \to Y$ locally Lipschitz continuous, and X separable. Then*

$$\dim_H(f(X)) \leq \dim_H(X).$$

Proof. For every $x \in X$ there exists an open neighbourhood U_x of x such that f restricted to U_x is Lipschitz continuous. Also $(U_x)_{x \in X}$ is an open cover of X. As X is separable and thus Lindelöf there exists a countable subcover $(U_{x_i})_{i \in \mathbb{N}}$. From Lemma E.1 we conclude

$$\dim_H(f(U_{x_i})) \leq \dim_H(U_{x_i}) \leq \dim_H(X)$$

for all $i \in \mathbb{N}$. Since $f(X) = f(\bigcup_{i \in \mathbb{N}} U_{x_i}) = \bigcup_{i \in \mathbb{N}} f(U_{x_i})$, we can again use Lemma E.1 to conclude

$$\dim_H(f(X)) = \dim_H(\bigcup_{i \in \mathbb{N}} f(U_{x_i})) \leq \sup_{i \in \mathbb{N}} \dim_H(f(U_{x_i})) \leq \dim_H(X). \qquad \square$$

We now come to the special case of the Euclidean space $\mathbb{R}^n$. Every Borel subset of $\mathbb{R}^n$ with non-empty interior has Hausdorff dimension n, again see for example (Edgar, 2008, Chapter 6). A direct consequence of this is the following result.

Lemma E.3. *We have* $\dim_H(\mathbb{R}^n) = n$ *and* $\dim_H(\mathbb{R}^n_+) = n$.

We are now ready to prove Lemma 3.8.

Proof of Lemma 3.8. Since $B \subseteq \mathbb{R}^n$ is a separable metric space, f is locally Lipschitz continuous, and $f(B) \subseteq \mathbb{R}^m$ has non-empty interior, we can use Lemmas E.2 and E.3 to obtain

$$m = \dim_H(f(B)) \leq \dim_H(B) \leq \dim_H(\mathbb{R}^n) = n. \qquad \square$$

F   Space-Filling Curves in Arbitrary Dimensions

In this section we describe the construction of the continuous and surjective function $\Gamma\colon \mathbb{R} \to \mathbb{R}^\infty$ and thus prove Lemma D.1.

We start with constructing a map from the unit interval to the n-dimensional cube for any $n \in \mathbb{N}$. We can then glue these maps together to obtain Γ.

Let $g\colon [0,1] \to [0,1]^2$ be any continuous and surjective space-filling curve. Examples for such curves are the Sierpiński curve, the Hilbert curve, or the Peano curve, see (Sagan, 1994) for an overview of various space-filling curves. We extend this to higher-dimensional cubes by iteratively defining curves $g_n\colon [0,1] \to [0,1]^n$ for $n \geq 2$ as

$$g_2 = g$$
$$g_n = (\mathrm{id}_{n-2} \times g) \circ g_{n-1}, \qquad n \geq 3,$$

which are again continuous and surjective. We obtain continuous and surjective curves $h_n\colon [0,1] \to [-n,n]^n$ from the unit interval to the scaled symmetric n-dimensional cubes by scaling and translating. Since we ultimately want glue all these h_n together we also want to assure $h_n(0) = h_n(1) = \mathbf{0}$ for all $n \geq 2$. Thus we define

$$h_n(t) = \begin{cases} 4nt(2g_n(0) - \mathbf{1}), & t \in \left[0, \frac{1}{4}\right], \\ n\left(2g_n\left(2t - \frac{1}{2}\right) - \mathbf{1}\right), & t \in \left[\frac{1}{4}, \frac{3}{4}\right], \\ 4n(1-t)(2g_n(1) - \mathbf{1}), & t \in \left[\frac{3}{4}, 1\right]. \end{cases}$$

Finally, we glue the pieces together to get a continuous map from one real parameter to any arbitrary finite number of parameters. We define the map $\Gamma\colon \mathbb{R} \to \mathbb{R}^\omega$ that maps the interval $[n, n+1]$ surjectively to the $[-n,n]^n$ cube on the first n coordinates of $\mathbb{R}^\omega$, that is

$$\Gamma(t) = \begin{cases} h_{\lfloor t \rfloor}(t - \lfloor t \rfloor) \times \{0\}^\infty, & t \geq 2, \\ \{0\}^\infty & , \quad t < 2. \end{cases}$$

The only critical points regarding the continuity of Γ are the integers where we transition from one interval to the next and thus $\lfloor t \rfloor$ changes. But by assuring $h_n(0) = h_n(1) = \mathbf{0}$ for all $n \geq 2$ we achieve a continuous gluing at the interval transitions. Thus Γ is continuous restricted to each of the intervals $(-\infty, 2], [2,3], [3,4], \ldots$ respectively. These form a locally finite cover of $\mathbb{R}$ by closed sets, hence we can use the pasting Lemma, see for example (Dugundji, 1966, Chapter III.9), to conclude the continuity of Γ on all of $\mathbb{R}$. The map is not surjective onto $\mathbb{R}^\omega$. However for our purpose we only require surjectivity onto $\mathbb{R}^\infty$ and this is clearly satisfied, since

$$\mathbb{R}^\infty = \bigcup_{n \geq 2} [-n,n]^n \times \{0\}^\infty = \bigcup_{n \geq 2} \Gamma([n, n+1]).$$

G   Restricting the Set of Weight Matrices

In this section we briefly discuss some variations of characterisations of $\mathcal{F}$-invariant families of distributions, where $\mathcal{F}$ is not the entire set of all ReLU layers. If the considered collection $\mathcal{F}$ of functions is much more restricted, it might be possible to obtain invariant families of distributions.

G.1 General Restrictions of the Weight Matrices

The setting we considered so far has no restrictions on the weight matrices. We make use of that in our constructive proof by repeatedly relying on rotation and projection matrices. Putting restrictions on the matrices, such as non-negativity, symmetry, positive-definiteness, or allowing only diagonal matrices, would prohibit our proof strategy. Using only diagonal matrices renders the functions in $\mathcal{F}$ separable in their input components and thus effectively reduces to the one-dimensional case. For this we have already discussed the invariant distributions in Section 4.1. For all other matrix restrictions it remains to be investigated whether our approach can be adapted.

G.2 Restrictions on the Number of Weight Matrices and Bias Vectors

Another practical concern can be the number of different possible weight matrices. So far we considered a continuum of matrices and biases. One possible restriction is to consider a finite collections $\mathcal{F}$ instead.

We begin with the extreme case, in which the collection contains only one measurable function $\mathcal{F} = \{f \colon \mathbb{R}^d \to \mathbb{R}^d\}$, for example a ReLU layer $f(\mathbf{x}) = \varrho(\mathbf{Wx} + \mathbf{b})$ with a fixed weight matrix $\mathbf{W}$ and bias vector $\mathbf{b}$. As in Sections 4.1 and 4.3 we can construct an invariant family of distributions starting with a prototype distribution $\mu_0 \in \mathcal{D}(\mathbb{R}^d)$. Next, we iteratively define

$$\mu_n = f_* \mu_{n-1}, \quad n \in \mathbb{N},$$

and using $\eta(t) = t - \lfloor t \rfloor$ also the intermediate interpolations

$$\mu_t = (1 - \eta(t))\mu_{\lfloor t \rfloor} + \eta(t)\mu_{\lfloor t \rfloor + 1}, \quad t \geq 0. \tag{10}$$

We quickly show the Lipschitz continuity of $t \mapsto \mu_t$. Let $0 \leq t_1 < t_2 < \infty$. Without loss of generality, we can assume $t_2 \leq t_1 + 1$, since the Prokhorov metric of two probability distributions is always bounded by 1. First, we consider the case $\lfloor t_2 \rfloor = \lfloor t_1 \rfloor = m$. Then

$$\begin{aligned}
d_P(\mu_{t_2}, \mu_{t_1}) &\leq \|\mu_{t_2} - \mu_{t_1}\|_{\mathrm{TV}} \\
&\leq \|(\eta(t_2) - \eta(t_1))(\mu_{m+1} - \mu_m)\|_{\mathrm{TV}} \\
&\leq |\eta(t_2) - \eta(t_1)| \|\mu_{m+1} - \mu_m\|_{\mathrm{TV}} \\
&\leq |t_2 - t_1|(\|\mu_{m+1}\|_{\mathrm{TV}} + \|\mu_m\|_{\mathrm{TV}}) \\
&= 2|t_2 - t_1|,
\end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm.

Second, we consider the case $\lfloor t_2 \rfloor = \lfloor t_1 \rfloor + 1 = m$. Then

$$\mu_{t_2} - \mu_{t_1} = (1 - \eta(t_2) - \eta(t_1))\mu_m + \eta(t_2)\mu_{m+1} - (1 - \eta(t_1))\mu_{m-1}$$

and since

$$|1 - \eta(t_2) - \eta(t_1)| \leq |1 - \eta(t_1)| + |\eta(t_2)| = t_2 - t_1$$

we obtain

$$\begin{aligned}
d_P(\mu_{t_2}, \mu_{t_1}) &\leq \|\mu_{t_2} - \mu_{t_1}\|_{\mathrm{TV}} \\
&\leq |1 - \eta(t_2) - \eta(t_1)| + |\eta(t_2)| + |1 - \eta(t_1)| \\
&\leq 2|t_2 - t_1|.
\end{aligned}$$

Thus, the one-parameter mapping $t \mapsto \mu_t$ is Lipschitz continuous. It also satisfies $f_* \mu_t = \mu_{t+1}$ for any $t \geq 0$, which means that it is f-invariant (hence $\mathcal{F}$-invariant).

This idea can also be extended to finite (or even countable) collections $\mathcal{F}$. For brevity we only illustrate the concept for a collection $\mathcal{F} = \{f_0, f_1\}$ of two functions. Again, we start by choosing an arbitrary prototype measure, which we will simply denote μ in this case. For any binary string $\mathbf{z} \in \{0, 1\}^*$, where ε refers to the empty string, we define

$$\begin{aligned}
\mu_\varepsilon &= \mu, \\
\mu_{\mathbf{z}} &= (f_{z_n} \circ \cdots \circ f_{z_1})_* \mu.
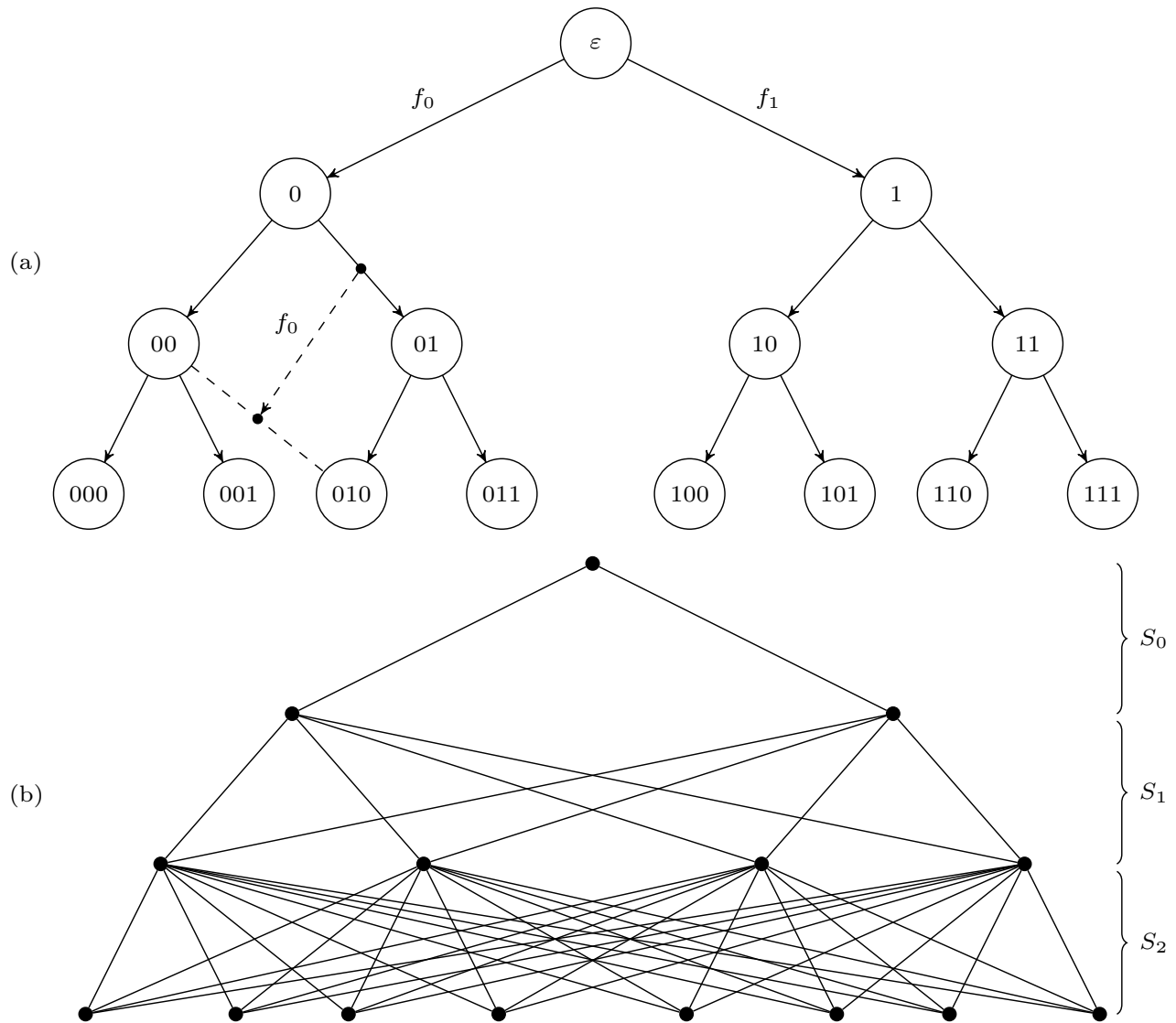\end{aligned}$$

Figure 9: (a) Consecutively applying either f_0 or f_1 to a prototype measure μ results in an infinite tree structure (only the first three levels are shown). Transforming an intermediate (interpolated) point by f_0 or f_1 (dashed) results in a point outside the tree structure. (b) Adding additional edges to represent also the transformations of all interpolated points results in a multipartite graph in which all vertices of consecutive levels are connected. The subgraphs S_i induced by two consecutive levels of vertices are complete bipartite graphs.

The procedure of obtaining measures this way can be associated to a perfect binary tree of infinite depth with root ε. A binary string $\mathbf{z}_1$ is the child of another string $\mathbf{z}_2$ if it extends it by exactly one digit 0 or 1, see Figure 9. Note that different strings can result in the same measure so that the set of measures $\mu_\mathbf{z}$ is not in one-to-one correspondence to the vertices of the tree. However, each measure is represented by at least on vertex.

In order to find a parametrisation that includes all the measures $\mu_\mathbf{z}$ and that is locally Lipschitz continuous and $\mathcal{F}$-invariant, we will turn the tree into an undirected graph. The parametrisation will extend the interpolation idea from (10) and results from a walk through all nodes of the graph.

Let $l(\mathbf{z})$ denote the length of a binary string $\mathbf{z}$. We define the graph $G = (V, E)$ with vertices $V = \{0,1\}^*$ and edges $E = \{(\mathbf{z}_1, \mathbf{z}_2) \in V \times V : |l(\mathbf{z}_1) - l(\mathbf{z}_2)| = 1\}$, i.e. every vertex is a binary string and all strings of consecutive lengths are connected by an edge. The resulting graph is a multipartite graph where the independent sets are strings of the same length. Let us define subgraphs S_i of G for $i \in \mathbb{N}_0$ that are induced by vertex sets

$$V_i = \{0,1\}^i \cup \{0,1\}^{i+1}.$$

The resulting subgraphs are complete bipartite graphs, see Figure 9. We define a walk $\mathcal{W} \in \{V\}^*$ on G as a sequence of vertices where two consecutive vertices are connected by an edge. We are now looking for an infinite walk on G that passes through all edges at least once.

Since we can pass an edge multiple times such a walk is easy to construct: Each subgraph S_i has only finitely many edges, which means there exists a walk that passes through them (except for S_0 we can even find an Eulerian cycle for each subgraph). Without loss of generality we can assume that the walk on S_i starts and ends in $\mathbf{0}_i$ (the string of i zeros) and denote it $\mathcal{W}_i$. Let $\sqcap$ symbolise the concatenation of two walks. We set

$$\mathcal{W} = \prod_{k=0}^{\infty} \mathcal{W}_i,$$

which is possible since in G the two vertices $\mathbf{0}_i$ and $\mathbf{0}_{i+1}$ are connected by an edge for every $i \in \mathbb{N}$. Denoting the i-th vertext visited by $\mathcal{W}$ as $\mathcal{W}(i)$ we can now define

$$\mu_t = (1 - \eta(t))\mu_{\mathcal{W}(\lfloor t \rfloor)} + \eta(t)\mu_{\mathcal{W}(\lfloor t \rfloor + 1)}, \tag{11}$$

with $\eta(t)$ as before.

To see that this fulfils our criteria consider the following. All distributions in the parametrised family are of the form

$$(1-s)\mu_{\mathbf{z}_1} + s\mu_{\mathbf{z}_2}, \quad \text{with} \quad \mathbf{z}_1 \in \{0,1\}^k, \mathbf{z}_2 \in \{0,1\}^{k+1}, k \in \mathbb{N}_0, s \in [0,1].$$

Transforming it by the function f_i for $i \in \{0,1\}$ results in the pushforward measure

$$(1-s)\mu_{\mathbf{z}_1'} + s\mu_{\mathbf{z}_2'}, \quad \text{with} \quad \mathbf{z}_1' = (\mathbf{z}_1, i), \mathbf{z}_2' = (\mathbf{z}_2, i).$$

But since $\mathbf{z}_1'$ and $\mathbf{z}_2'$ still have length difference 1, they share an edge in S^{k+1}. The walk $\mathcal{W}$ passes through all edges, so we can define $m \in \mathbb{N}$ as the smallest number where either $\mathcal{W}_m = \mathbf{z}_1'$ and $\mathcal{W}_{m+1} = \mathbf{z}_2'$ or $\mathcal{W}_m = \mathbf{z}_2'$ and $\mathcal{W}_{m+1} = \mathbf{z}_1'$. Let us without loss of generality assume that the former holds. Then

$$(1-s)\mu_{\mathbf{z}_1'} + s\mu_{\mathbf{z}_2'} = \mu_{m+s},$$

which shows $t \mapsto \mu_t$ is $\mathcal{F}$-invariant. The Lipschitz continuity follows analogously to the case with only a single function above.

This idea can even be extended to countably many different measurable functions. Instead of a single binary tree and corresponding graph G, we could define a sequence of trees T_i and corresponding graphs G_i where T_i is i-ary, corresponding to the first i functions. Every tree T_i is a subtree of T_j whenever $j > i$ and correspondingly the associated graph G_i is a subgraph of G_j. Thus any walk on G_i is also valid on G_j. If now $\mathcal{W}_i$ denotes a walk on G_i starting at the root ε, covering every edge up to the i-th level in G_i, and returning back to the root ε, then we can set

$$\mathcal{W} = \prod_{k=1}^{\infty} \mathcal{W}_i.$$

It follows that every edge of every graph in the sequence will eventually be reached. Defining the parametrisation as in (11), ensures that both $\mathcal{F}$-invariance as well as Lipschitz continuity holds.