On the Correctness of Automatic Differentiation for Neural Networks with Machine-Representable Parameters

Wonyeol Lee [1] **Sejun Park** [2] **Alex Aiken** [1]

Abstract

Recent work has shown that forward- and reverse-mode automatic differentiation (AD) over the reals is almost always correct in a mathematically precise sense. However, actual programs work with *machine-representable numbers* (e.g., floating-point numbers), not reals. In this paper, we study the correctness of AD when the parameter space of a neural network consists solely of machine-representable numbers. In particular, we analyze two sets of parameters on which AD can be incorrect: the incorrect set on which the network is differentiable but AD does not compute its derivative, and the non-differentiable set on which the network is non-differentiable. For a neural network *with bias parameters*, we first prove that the incorrect set is always empty. We then prove a tight bound on the size of the non-differentiable set, which is linear in the number of non-differentiabilities in activation functions, and give a simple necessary and sufficient condition for a parameter to be in this set. We further prove that AD always computes a Clarke subderivative even on the non-differentiable set. We also extend these results to neural networks possibly without bias parameters.

1. Introduction

Forward- and reverse-mode automatic differentiation (AD) are popular algorithms for computing the derivative of a function represented by a program (Griewank & Walther, 2008). Diverse practical systems for AD have been developed for general-purpose programs (Baydin et al., 2016; Hascoët & Pascual, 2013; Maclaurin et al., 2015; Pearlmutter & Siskind, 2008; Revels et al., 2016; Slusanschi & Dumitrel, 2016; Walther & Griewank, 2012), and particularly for machine-learning programs (Bergstra et al., 2010; Collobert et al., 2011; Jia et al., 2014; Seide & Agarwal, 2016; Tokui et al., 2019; van Merrienboer et al., 2018), including TensorFlow (Abadi et al., 2016), PyTorch (Paszke et al., 2017), and JAX (Frostig et al., 2018). The development of such AD systems has been a driving force of the rapid advances in deep learning (and machine learning in general) in the past 10 years (Baydin et al., 2017; LeCun et al., 2015; Schmidhuber, 2015).

Recently, the correctness of AD has been actively studied for various types of programs. For programs that only use differentiable functions, AD is correct *everywhere*, i.e., it computes the derivative of a given program at all inputs (Abadi & Plotkin, 2020; Barthe et al., 2020; Brunel et al., 2020; Elliott, 2018; Huot et al., 2020; Krawiec et al., 2022; Radul et al., 2023; Smeding & Vákár, 2023; Vákár, 2021). On the other hand, for programs that use non-differentiable functions (e.g., ReLU[1]), AD can be incorrect at some inputs (Kakade & Lee, 2018).

There are two cases where AD is incorrect. The first case is when the function f represented by a given program is differentiable at some x, but AD returns a value different from the derivative of f at x. For instance, consider a program[2] that represents the identity function, defined as $\mathrm{ReLU}(x) - \mathrm{ReLU}(-x)$. If AD uses zero as a "derivative" of ReLU at $x = 0$, as is standard (e.g., in TensorFlow and PyTorch), it returns zero for this program at $x = 0$ while the true derivative is one. The second case is when f is non-differentiable at some x, but AD does not return a generalized notion of derivative (e.g., Clarke subdifferential) of f at x. For example, $\mathrm{ReLU}(x) - \frac{1}{2}\mathrm{ReLU}(-x)$ represents a function that is non-differentiable at $x = 0$ with the Clarke subdifferential $[\frac{1}{2}, 1]$, but AD outputs 0 at $x = 0$.

Although AD can be incorrect, recent works show that for a large class of programs using non-differentiable functions, AD is correct *almost everywhere*, i.e., it is incorrect at most on a Lebesgue measure-zero subset of the input domain of a program (Bolte & Pauwels, 2020a;b; Huot et al., 2023; Lee

[1]Stanford University, USA [2]Korea University, South Korea. Correspondence to: Wonyeol Lee <wonyeol.lee.cs@gmail.com>, Sejun Park <sejun.park000@gmail.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] $\mathrm{ReLU}(x) \triangleq \max\{x, 0\}$.

[2] It appeared in Kakade & Lee (2018).

et al., 2020; Mazza & Pagani, 2021).

These prior works, however, have a limitation: they consider AD over the real numbers, but in practice, inputs to a program are always *machine-representable numbers* such as 32-bit floating-point numbers. Since the set of machine-representable numbers is countable (and usually finite), it is always a Lebesgue measure-zero subset of the real numbers. Hence, AD could be incorrect on *all* machine-representable inputs according to prior works, and this is indeed possible. Consider a program[3] for a function from $\mathbb{R}$ to $\mathbb{R}$, defined as

$$\sum_{c \in \mathbb{M}} \Big[\lambda x + \Big(\frac{1}{|\mathbb{M}|} - \lambda \Big) \Big(\mathrm{ReLU}(x - c) - \mathrm{ReLU}(-x + c) \Big) \Big],$$

where $\mathbb{M} \subseteq \mathbb{R}$ is a finite set of machine-representable numbers and $\lambda \in \mathbb{R} \setminus \{1\}$ is an arbitrary constant. Then, the program represents the affine function $x \mapsto x + a$ for $a = (\lambda - \frac{1}{|\mathbb{M}|}) \times \sum_{c \in \mathbb{M}} c$, but AD incorrectly computes its derivative at any $x \in \mathbb{M}$ as λ (the arbitrarily chosen value) if zero is used as a "derivative" of ReLU at 0 as before.[4]

Given these observations, we raise the following questions: for a program that represents a neural network, at which machine-representable inputs to the program (i.e., parameters to the network) can AD be incorrect, and how many such inputs can there be? In this work, we tackle these questions and present the first theoretical results. In particular, we study the two sets of machine-representable parameters of a neural network on which AD can be incorrect: the *incorrect set*, on which the network is differentiable but AD does not compute its derivative, and the *non-differentiable set*, on which the network is non-differentiable.

Summary of results. We focus on neural networks consisting of alternating analytic pre-activation functions (e.g., fully-connected and convolution layers) and pointwise continuous activation functions (e.g., ReLU and $\mathrm{Sigmoid}$). The first set of our results (§3) is for such networks *with bias parameters* at every layer, and is summarized as follows.

- We prove that the incorrect set is *always empty*, not only over machine-representable parameters but also over real-valued ones. To our knowledge, this is the first result showing that the incorrect set can be empty for a class of neural networks using possibly non-differentiable functions; prior works only bounded the measure of this set.
- On the other hand, the non-differentiable set can be non-empty. We give a tight bound on its density over all machine-representable parameters, which has the form $n/|\mathbb{M}|$ where n is the *total number of non-differentiable points* in activation functions. This result implies that in practice, the non-differentiable set often has a low density, especially if we use high-precision parameters (e.g., use 32-bit floating-point numbers for $\mathbb{M}$, where $|\mathbb{M}| \approx 2^{32}$).
- To better describe the non-differentiable set, we provide a simple, easily verifiable *necessary and sufficient condition* for a parameter to be in the non-differentiable set. Given that deciding the non-differentiability of a neural network is NP-hard in general (Bolte et al., 2023), our result is surprising: having bias parameters is sufficient to efficiently decide the non-differentiability.
- Given that the non-differentiable set can be non-empty, a natural question arises: what does AD compute on this set? We prove that AD *always computes a Clarke subderivative* (a generalized derivative) even on the non-differentiable set. That is, AD is an efficient algorithm for computing a Clarke subderivative in this case.

The second set of our results (§4) extends the above results to neural networks possibly *without bias parameters* at some layers, and is summarized as follows.

- As we observed in the $\mathrm{ReLU}(x) - \mathrm{ReLU}(-x)$ example, the incorrect set can be non-empty in this case. Thus, we prove tight bounds on the density of both the incorrect and non-differentiable sets, which have the form $n'/|\mathbb{M}|$ where n' is linear in the total number of non-differentiable points in activation functions as well as the total number of boundary points in activation functions' zero sets.
- We provide simple, easily verifiable sufficient conditions on parameters under which AD computes the standard derivative or a Clarke subderivative.

Our theoretical results carry two main practical implications: AD for neural networks is correct on most machine-representable parameters, and it is correct more often with bias parameters. For networks with bias parameters at all layers, our results further provide an exact characterization of when AD is correct and what it computes.

We remark that many of our results, especially all the results not about the density of certain sets, hold not only for machine-representable parameters but also for real-valued ones. On the other hand, our results may not be directly applicable to neural networks with non-analytic pre-activation functions or non-pointwise activation functions; we discuss such limitations in §6.

Organization. We first introduce notation and the problem setup (§2). We then present our main results for neural networks with bias parameters (§3) and extend them to neural networks possibly without bias parameters (§4). We conclude the paper with related work and discussion (§5–7).

[3] Inspired by Bolte & Pauwels (2020b); Mazza & Pagani (2021).

[4] We can even make AD return different values at different $x \in \mathbb{M}$, by using a different λ_i for each $c_i \in \mathbb{M}$. Similarly, we can also construct a program such that at all machine-representable numbers $\mathbb{M}$, the program is non-differentiable and AD returns arbitrary values.

2. Problem Setup

2.1. Notation and Definitions

We use the following notation and definitions. Let $\mathbb{N}$ and $\mathbb{R}$ be the sets of positive integers and real numbers, respectively. For $n \in \mathbb{N}$, we use $[n] \triangleq \{1, 2, \ldots, n\}$ and $\vec{0}_n \triangleq (0, \ldots, 0) \in \mathbb{R}^n$, and often drop n from $\vec{0}_n$ when the subscript is clear from context. For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we use $x_{-i} \triangleq (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. We call $A \subseteq \mathbb{R}$ an *interval* if it is $[a, b]$, $[a, b)$, $(a, b]$, or (a, b) for some $a, b \in \mathbb{R} \cup \{\pm\infty\}$. For $A \subseteq \mathbb{R}^n$, $\mathbf{1}_A : \mathbb{R}^n \to \{0, 1\}$ denotes the indicator function of A. We say that $f : \mathbb{R}^n \to \mathbb{R}^m$ is *analytic* if it is infinitely differentiable and its Taylor series at any $x \in \mathbb{R}^n$ converges to f on some neighborhood of x. For any $f : \mathbb{R}^n \to \mathbb{R}^m$,

$$Df : \mathbb{R}^n \to \mathbb{R}^{m \times n} \cup \{\bot\}$$

denotes the standard derivative of f, where $f(x) = \bot$ denotes that f is non-differentiable at x. Lastly, for $f : \mathbb{R} \to \mathbb{R}$,

$$\mathsf{ndf}(f) \triangleq \{x \in \mathbb{R} \mid f \text{ is non-differentiable at } x\},$$
$$\mathsf{bdz}(f) \triangleq bd(\{x \in \mathbb{R} \mid f(x) = 0\})$$

denote the set of non-differentiable points of f and the boundary of the zero set of f, respectively.

2.2. Neural Networks

We define a neural network as follows. Given the number of layers $L \in \mathbb{N}$, let $N_0 \in \mathbb{N}$ be the dimension of input data, $N_l \in \mathbb{N}$ and $W_l \in \mathbb{N} \cup \{0\}$ be the number of neurons and the number of parameters at layer $l \in [L]$, and $N \triangleq N_1 + \cdots + N_L$ and $W \triangleq W_1 + \cdots + W_L$. Further, for each $l \in [L]$, let $\tau_l : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l} \to \mathbb{R}^{N_l}$ be an analytic *pre-activation function* and $\sigma_l : \mathbb{R}^{N_l} \to \mathbb{R}^{N_l}$ be a pointwise, continuous *activation function*, i.e.,

$$\sigma_l(x_1, \ldots, x_{N_l}) \triangleq \big(\sigma_{l,1}(x_1), \ldots, \sigma_{l,N_l}(x_{N_l})\big)$$

for some continuous $\sigma_{l,i} : \mathbb{R} \to \mathbb{R}$. Under this setup, we define a neural network as a function of model parameters: given input data $c \in \mathbb{R}^{N_0}$, a *neural network* $z_L(\,\cdot\,; c) : \mathbb{R}^W \to \mathbb{R}^{N_L}$ is defined as

$$z_L(w; c) \triangleq (\sigma_L \circ \tau_L^{\langle w_L \rangle} \circ \cdots \circ \sigma_1 \circ \tau_1^{\langle w_1 \rangle})(c), \quad (1)$$

where $w \triangleq (w_1, \ldots, w_L)$, $w_l \triangleq (w_{l,1}, \ldots, w_{l,W_l}) \in \mathbb{R}^{W_l}$, and $\tau_l^{\langle w_l \rangle}(x) \triangleq \tau_l(x, w_l)$. We say such z_L *has L layers, N neurons, and W parameters*.

We next define the *activation neurons* $z_l(\,\cdot\,; c) : \mathbb{R}^W \to \mathbb{R}^{N_l}$ and the *pre-activation values* $y_l(\,\cdot\,; c) : \mathbb{R}^W \to \mathbb{R}^{N_l}$ at layer $l \in [L]$, as we defined z_L above:

$$z_l(w; c) \triangleq (\sigma_l \circ \tau_l^{\langle w_l \rangle} \circ \cdots \circ \sigma_1 \circ \tau_1^{\langle w_1 \rangle})(c),$$
$$y_l(w; c) \triangleq \tau_l^{\langle w_l \rangle}(z_{l-1}(w; c)),$$

where $z_0(w; c) \triangleq c$. Since the input data c is fixed while we compute the derivative of z_L with respect to w (e.g., in order to train z_L), we often omit c and simply write $z_l(w)$ and $y_l(w)$ to denote $z_l(w; c)$ and $y_l(w; c)$, respectively.

For the set of all indices of neurons

$$\mathsf{Idx} \triangleq \{(l, i) \mid l \in [L], i \in [N_l]\}$$

and for each $(l, i) \in \mathsf{Idx}$, we use $y_{l,i}, z_{l,i} : \mathbb{R}^W \to \mathbb{R}$ and $\tau_{l,i} : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l} \to \mathbb{R}$ to denote the functions that take only the i-th output component of y_l, z_l, and τ_l, respectively. Note that we defined $\sigma_{l,i}$ above in a slightly different way: its domain is not $\mathbb{R}^{N_l}$ (i.e., the domain of σ_l) but $\mathbb{R}$.

Finally, we introduce the notion of piecewise-analytic[5] to consider possibly non-differentiable activation functions.

Definition 2.1. A function $f : \mathbb{R} \to \mathbb{R}$ is *piecewise-analytic* if there exist $n \in \mathbb{N}$, a partition $\{A_i\}_{i \in [n]}$ of $\mathbb{R}$ consisting of non-empty intervals, and analytic functions $\{f_i : \mathbb{R} \to \mathbb{R}\}_{i \in [n]}$ such that $f = f_i$ on A_i for all $i \in [n]$.

Assumption. $\sigma_{l,i}$ is piecewise-analytic for all $(l, i) \in \mathsf{Idx}$.

The class of piecewise-analytic functions includes not only all analytic functions but also many non-differentiable functions widely used in neural networks such as ReLU, LeakyReLU, and HardSigmoid. Hence, our definition of neural networks includes a rich class of practical networks: τ_l can be any analytic function (e.g., a fully-connected, convolution, or normalization layer), and σ_l can be any pointwise continuous and piecewise-analytic function (e.g., ReLU, LeakyReLU, or HardSigmoid).

In practice, we often apply AD to the composition of a neural network z_L and a loss function ℓ (e.g., Softmax followed by CrossEntropy), to compute the derivative of the loss value of z_L with respect to its parameters. We emphasize that all of our results except for lower bounds (i.e., Theorems 3.4, 4.3, and 4.5) continue to hold even if we replace z_L in their conclusions by $\ell \circ z_L$ for any analytic $\ell : \mathbb{R}^{N_L} \to \mathbb{R}^m$. For simplicity, however, we state our results only for z_L and not for $\ell \circ z_L$.

2.3. Automatic Differentiation

Given a program that represents a neural network z_L as in Eq. (1), AD essentially computes the function

$$D^{\mathsf{AD}} z_L : \mathbb{R}^W \to \mathbb{R}^{N_L \times W}$$

by applying the chain rule of differentiation to Eq. (1). That is, $D^{\mathsf{AD}} z_L$ is defined as the product of $D^{\mathsf{AD}} \tau_{l,i}$ and $D^{\mathsf{AD}} \sigma_{l,i}$ for $(l, i) \in \mathsf{Idx}$, where $D^{\mathsf{AD}} \tau_{l,i} : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l} \to$

[5] It is inspired by the notion of PAP in Lee et al. (2020).

$\mathbb{R}^{1\times(N_{l-1}+W_l)}$ and $D^{\mathsf{AD}}\sigma_{l,i} : \mathbb{R}^{N_l} \to \mathbb{R}^{1\times N_l}$ denote the "derivatives" of $\tau_{l,i}$ and $\sigma_{l,i}$ that AD uses in its computation (see Appendix A.3 for more details). Here $D^{\mathsf{AD}}z_L$, $D^{\mathsf{AD}}\tau_{l,i}$, and $D^{\mathsf{AD}}\sigma_{l,i}$ can be different from the standard derivatives Dz_L, $D\tau_{l,i}$, and $D\sigma_{l,i}$, partly because the former never return $\perp$ even at non-differentiable points while the latter always return $\perp$ at those points. We note that $D^{\mathsf{AD}}z_L$ expresses what practical AD systems (e.g., TensorFlow, PyTorch) essentially compute in *both* forward-mode and reverse-mode.

By definition, the output $D^{\mathsf{AD}}z_L$ of AD depends on the choice of $D^{\mathsf{AD}}\tau_{l,i}$ and $D^{\mathsf{AD}}\sigma_{l,i}$. To focus on the standard choices made by practical AD systems, we introduce the notion of an extended derivative.

Definition 2.2. A function $g : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ is an *extended derivative* of $f : \mathbb{R}^n \to \mathbb{R}^m$ if for all $x \in \mathbb{R}^n$ with $Df(x) \neq \perp$, it holds that $g(x) = Df(x)$.

Assumption. $D^{\mathsf{AD}}f$ is an extended derivative of f for all $f \in \{\tau_{l,i}, \sigma_{l,i} \mid (l,i) \in \mathsf{Idx}\}$.

We note that a differentiable function f has a unique extended derivative which is the standard derivative Df of f. In contrast, a non-differentiable function f has (uncountably) many extended derivatives: e.g., $\mathbf{1}_{(0,\infty)} + c \cdot \mathbf{1}_{\{0\}}$ is an extended derivative of ReLU for all $c \in \mathbb{R}$, where $\mathbf{1}_A$ denotes the indicator function of a set A.

Among many extended derivatives, some of them are used more frequently in practice, which we characterize as consistency.

Definition 2.3. For $f : \mathbb{R}^n \to \mathbb{R}^m$, an extended derivative g of f is *consistent* if for all $x \in \mathbb{R}^n$ with $Df(x) = \perp$, it holds that $g(x) = \lim_{k\to\infty} Df(x_k)$ for some $x_k \to x$.[6]

For instance, $\mathbf{1}_{(0,\infty)}$ and $\mathbf{1}_{[0,\infty)}$ are consistent extended derivatives of ReLU but $\mathbf{1}_{(0,\infty)}+c\cdot\mathbf{1}_{\{0\}}$ is not for all $c \in \mathbb{R}\setminus\{0,1\}$; among them, $D^{\mathsf{AD}}\mathrm{ReLU} = \mathbf{1}_{(0,\infty)}$ is typically used by popular AD systems (e.g., TensorFlow and PyTorch). Although $D^{\mathsf{AD}}f$ is usually consistent in practice, we do not assume it by default (and explicitly assume it only when necessary) to make our results as general as possible, and to study whether the values of extended derivatives at non-differentiable points matter to AD.

2.4. Incorrect and Non-Differentiable Sets

In practice, the parameters of a neural network cannot be arbitrary real numbers (as machines cannot represent them), but can only be machine-representable numbers $\mathbb{M} \subseteq \mathbb{R}$, where $\mathbb{M}$ is often chosen as the set of all 32-bit floating-point numbers. To this end, we consider

$$\Omega \triangleq \mathbb{M}^W \subseteq \mathbb{R}^W,$$

the set of parameters that a neural network $z_L : \mathbb{R}^W \to \mathbb{R}^{N_L}$ can take in practice. We assume that $\mathbb{M}$ is an arbitrary finite subset of $\mathbb{R}$ throughout the paper; e.g., it can be the set of n-bit floating-point (or fixed-point) numbers for any $n \in \mathbb{N}$.

To better understand the correctness of AD, we study the following two disjoint subsets of Ω on which AD can return an incorrect output.

Definition 2.4. For a neural network z_L, define the *incorrect set* and the *non-differentiable set* of z_L as

$$\mathsf{inc}_\Omega(z_L) \triangleq \{w \in \Omega \mid Dz_L(w) \neq \perp,\ D^{\mathsf{AD}}z_L(w) \neq Dz_L(w)\},$$
$$\mathsf{ndf}_\Omega(z_L) \triangleq \{w \in \Omega \mid Dz_L(w) = \perp\}.$$

These two sets correspond to the two cases when AD can be incorrect: on the incorrect set $\mathsf{inc}_\Omega(z_L)$, z_L is differentiable but AD does not compute its standard derivative; on the non-differentiable set $\mathsf{ndf}_\Omega(z_L)$, z_L is non-differentiable and AD may not compute a generalized notion of derivative (e.g., Clarke subdifferential). Here $\mathsf{ndf}_\Omega(z_L) \subseteq \Omega$ is different from $\mathsf{ndf}(f) \subseteq \mathbb{R}$, which was defined in §2.1 for $f : \mathbb{R} \to \mathbb{R}$.

3. Correctness of Automatic Differentiation for Neural Networks with Bias Parameters

Our main objective is to understand the incorrect and non-differentiable sets. In particular, we focus on neural networks with bias parameters (defined below) in this section and consider more general neural networks in §4. For the former class of neural networks, we characterize the incorrect and non-differentiable sets in §3.1 and §3.2, and establish a connection between AD and Clarke subderivatives (a generalized notion of derivative) in §3.3.

We start by defining neural networks with bias parameters.

Definition 3.1. A pre-activation function $\tau_l : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l} \to \mathbb{R}^{N_l}$ of a neural network *has bias parameters* if $W_l \geq N_l$ and there exist $f_1, \ldots, f_{N_l} : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l - N_l} \to \mathbb{R}$ such that

$$\tau_{l,i}(x,(u,v)) = f_i(x,u) + v_i$$

for all $i \in [N_l]$ and $(x,u,v) \in \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l - N_l} \times \mathbb{R}^{N_l}$. Here v_i is called the *bias parameter of* $\tau_{l,i}$. A neural network z_L *has bias parameters* if τ_l has bias parameters for all $l \in [L]$.

Many popular pre-activation functions are typically implemented with bias parameters. For example, fully-connected layers, attention layers (e.g., MultiheadAttention), and some normalization layers (e.g., LayerNorm) do so. Yet not all pre-activation functions have bias parameters in practice. For instance, convolutional layers and other normalization layers (e.g., BatchNorm) usually do not satisfy Definition 3.1: they do contain some bias terms, but each of these terms is used to compute multiple output values (instead of a single output value as in our definition).

[6] Any consistent extended derivative of f is an element of the so-called Bouligand subdifferential of f (Cui & Pang, 2021). But the converse does not hold in general.

3.1. Characterization of the Incorrect Set

We first show that the incorrect set of a neural network is always empty if the network has bias parameters, i.e., AD computes the standard derivative wherever the network is differentiable.

Theorem 3.2. *If a neural network z_L has bias parameters, then for all $w \in \mathbb{R}^W$ at which z_L is differentiable,*

$$D^{\mathsf{AD}} z_L(w) = Dz_L(w). \tag{2}$$

This implies that $|\mathsf{inc}_\Omega(z_L)| = 0$.

It should be emphasized that Eq. (2) is not only for machine-representable parameters, but also for any real-valued parameters. Compared to existing results, this result is surprising. For instance, Bolte & Pauwels (2020b); Lee et al. (2020) show that the incorrect set over $\mathbb{R}^n$ (not over $\mathbb{M}^n$) has Lebesgue measure zero for some classes of programs, but they do not give any results on whether the set can be empty. In contrast, Theorem 3.2 states that the incorrect set over $\mathbb{R}^n$ is empty for a smaller, yet still large class of programs, i.e., neural networks with bias parameters.

In Theorem 3.2, the condition that z_L has bias parameters plays a crucial role. Namely, Theorem 3.2 does not hold if this condition is dropped. For instance, consider a neural network $z_L : \mathbb{R} \to \mathbb{R}$ that is essentially the same as $f : \mathbb{R} \to \mathbb{R}$ with $f(w) = \mathrm{ReLU}(w) - \mathrm{ReLU}(-w)$ (which we discussed in §1). Then, z_L does not have bias parameters, and $\mathsf{inc}_\Omega(z_L)$ is non-empty if $D^{\mathsf{AD}}\mathrm{ReLU} = \mathbf{1}_{(0,\infty)}$ is used.

The proof of Theorem 3.2 consists of the following two arguments: for all $w \in \mathbb{R}^W$ with $Dz_L(w) \neq \bot$,

(i) if $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$, then $\partial z_L / \partial z_{l,i} = \vec{0}$ at w, and
(ii) if (i) holds, then $D^{\mathsf{AD}} z_L(w) = Dz_L(w)$.

That is, (i) if a pre-activation value $y_{l,i}$ touches a non-differentiable point of its activation function $\sigma_{l,i}$, then the derivative of z_L with respect to $z_{l,i}$ should always be zero; and (ii) Theorem 3.2 follows from (i). We point out that the proof of (i) relies heavily on the bias parameter condition. For more details, see Appendix C.

3.2. Characterization of the Non-Differentiable Set

We next show that if a neural network has bias parameters, then the density of the non-differentiable set in Ω is bounded by $n/|\mathbb{M}|$, where n is the total number of non-differentiable points in activation functions.

Theorem 3.3. *If a neural network z_L has bias parameters,*

$$\frac{|\mathsf{ndf}_\Omega(z_L)|}{|\Omega|} \leq \frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} |\mathsf{ndf}(\sigma_{l,i})|$$

where $\mathsf{ndf}(f)$ *is the set of non-differentiable points of* f.

In many practical settings, the bound in Theorem 3.3 is often small, especially under high-precision parameters. For example, $\mathbb{M}$ is frequently chosen as the set of 32-bit floating-point numbers so $|\mathbb{M}| \approx 2^{32}$, while $|\mathsf{Idx}|$ (the number of neurons) is often smaller than 2^{32} and $|\mathsf{ndf}(\sigma_{l,i})|$ is typically small (e.g., 0 for differentiable $\sigma_{l,i}$, 1 for ReLU, and 2 for HardSigmoid). This implies that in practice, the non-differentiable set often has a low density in Ω. We remark, however, that the bound in Theorem 3.3 can grow large in low-precision settings (e.g., when parameters are represented by ≤ 16-bit numbers).

Although the bound in Theorem 3.3 can be large in some cases (e.g., when $|\mathbb{M}|$ is small), we prove that the bound is in general tight up to a constant multiplicative factor.

Theorem 3.4. *For any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ with $1 \leq |\mathbb{M}| < \infty$, $n \geq 2$, and $\alpha \leq |\mathbb{M}|/(n-1)$, there is a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ with bias parameters that satisfies*

$$\frac{|\mathsf{ndf}_\Omega(z_L)|}{|\Omega|} \geq \frac{1}{2} \cdot \frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} |\mathsf{ndf}(\sigma_{l,i})|$$

and the following: z_L has $n+1$ neurons and $|\mathsf{ndf}(\sigma_{1,i})| = \alpha$ for all $i \in [N_1]$.

In Theorem 3.4, the condition $\alpha \leq |\mathbb{M}|/(n-1)$ is for achieving the constant $1/2$ in the bound. A similar bound can be derived for a larger α (i.e., $\alpha > |\mathbb{M}|/(n-1)$) but with a constant smaller than $1/2$.

Theorems 3.3 and 3.4 describe how large the non-differentiable set $\mathsf{ndf}_\Omega(z_L)$ can be, but give no clue about exactly which parameters constitute this set. To better understand this, we present an easily verifiable necessary and sufficient condition for characterizing $\mathsf{ndf}_\Omega(z_L)$.

Theorem 3.5. *If a neural network z_L has bias parameters, then the following are equivalent for all $w \in \mathbb{R}^W$.*

- *z_L is non-differentiable at w.*
- *$y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$ and $\partial^{\mathsf{AD}} z_L / \partial z_{l,i} \neq \vec{0}$ at w for some $(l,i) \in \mathsf{Idx}$.*

Here $\partial^{\mathsf{AD}} z_L / \partial z_{l,i}$ denotes the partial derivative of z_L with respect to $z_{l,i}$ that reverse-mode AD (e.g., backpropagation) computes as a byproduct of computing $D^{\mathsf{AD}} z_L$ (see Appendix E.2 for more details). Hence, Theorem 3.5 implies that we can efficiently[7] decide whether a neural network with bias parameters is non-differentiable at a (real-valued) parameter or not. This result is surprising given a recent, relevant result that deciding such non-differentiability is NP-hard in general (Bolte et al., 2023).

We now sketch the proof of Theorem 3.3, to explain how we obtain the bound in the theorem and where we use the bias

[7] in $\mathcal{O}(N_L T)$ time for a neural network $z_L : \mathbb{R}^W \to \mathbb{R}^{N_L}$ where T is the time to evaluate $z_L(w)$, because reverse-mode AD takes $\mathcal{O}(N_L T)$ time to compute $D^{\mathsf{AD}} z_L(w)$.

parameter condition. First, we prove that if $y_{l,i}(w)$ does not touch any non-differentiable point of $\sigma_{l,i}$ for all $(l,i) \in \mathsf{Idx}$, then z_L is differentiable at w. In other words,

$$\mathsf{ndf}_\Omega(z_L) \subseteq \bigcup_{(l,i)\in\mathsf{Idx}} \bigcup_{c\in\mathsf{ndf}(\sigma_{l,i})} \{w \in \Omega \mid y_{l,i}(w) = c\}. \quad (3)$$

Second, we prove that for all $(l,i) \in \mathsf{Idx}$ and $c \in \mathbb{R}$,

$$\left|\{w \in \Omega \mid y_{l,i}(w) = c\}\right| \leq |\mathbb{M}|^{W-1}. \quad (4)$$

This inequality is invalid in general, but is valid when τ_l has bias parameters. If the parameter w has a value $v = (v_1, \ldots, v_W)$ and its j-th entry v_j corresponds to the bias parameter of $\tau_{l,i}$, then $y_{l,i}(v) = f(v_{-j}) + v_j$ for some function f. Hence, for any $v_{-j} \in \mathbb{M}^{W-1}$, there is at most one $v_j \in \mathbb{M}$ achieving $y_{l,i}(v) = c$, and this implies the above inequality. Finally, we prove that Theorem 3.3 follows from the above two results. The full proofs of Theorems 3.3–3.5 are presented in Appendices B, D, and E, respectively.

3.3. Connection to Clarke Subderivatives

We have so far observed that with bias parameters, the incorrect set is always empty but the non-differentiable set may not be. A natural question is then: what does AD compute on the non-differentiable set? We answer this question by showing that AD computes a Clarke subderivative[8] everywhere (including on the non-differentiable set), if it uses *consistent* extended derivatives for activation functions.

Theorem 3.6. *If a neural network z_L has bias parameters and $D^{\mathsf{AD}}\sigma_{l,i}$ is consistent for all $(l,i) \in \mathsf{Idx}$, then for all $w \in \mathbb{R}^W$,*

$$D^{\mathsf{AD}} z_L(w) = \begin{cases} Dz_L(w) & \text{if } Dz_L(w) \neq \bot \\ \lim_{n\to\infty} Dz_L(w'_n) & \text{if } Dz_L(w) = \bot. \\ \quad \text{for some } w'_n \to w & \end{cases}$$

This implies that $D^{\mathsf{AD}} z_L$ is a Clarke subderivative of z_L.

Theorem 3.6 is not only a new result about AD, but also gives a positive answer to a long-standing open question about Clarke subgradients (Bolte et al., 2023; Clarke, 1975; Kakade & Lee, 2018): are there a sufficiently large class $\mathcal{F}$ of *scalar* functions and a deterministic algorithm $\mathcal{A}$ that computes a *Clarke subgradient* (i.e., subderivative) of $f \in \mathcal{F}$ at $x \in \mathbb{R}^n$ efficiently (i.e., in time $\mathcal{O}(T)$ that is independent of n, where T is time to evaluate $f(x)$)? In other words, is there a so-called "Cheap Subgradient Principle"? For instance, Kakade & Lee (2018) propose an efficient algorithm $\mathcal{A}'$ (for some $\mathcal{F}'$) but $\mathcal{A}'$ is not deterministic, whereas Barton et al. (2018); Khan & Barton (2015) propose deterministic algorithms $\mathcal{A}''$ (for some $\mathcal{F}''$) but $\mathcal{A}''$ are not efficient. In contrast, Theorem 3.6 implies that for neural networks with bias parameters, a Clarke subgradient at any (real-valued) parameter can be computed deterministically and efficiently, even by the vanilla reverse-mode AD. In this sense, we provide a new understanding on the computational aspects of Clarke subgradients.

We note that Theorem 3.6 no longer holds without any of its conditions: having bias parameters and using consistent extended derivatives. One can confirm this using the following examples: $z_L(w) = \mathrm{ReLU}(w) - \mathrm{ReLU}(-w)$ with $D^{\mathsf{AD}}\mathrm{ReLU} = \mathbf{1}_{(0,\infty)}$ (in which z_L does not have bias parameters as observed in §3.1), and $\widehat{z}_L(w) = \mathrm{ReLU}(w)$ with $D^{\mathsf{AD}}\mathrm{ReLU} = \mathbf{1}_{(0,\infty)} + c \cdot \mathbf{1}_{\{0\}}$ for any $c \in \mathbb{R} \setminus [0,1]$ (in which $D^{\mathsf{AD}}\mathrm{ReLU}$ is not consistent). For the proof of Theorem 3.6, see Appendix F.

4. Correctness of Automatic Differentiation for Neural Networks without Bias Parameters

In this section, we investigate the correctness of AD for neural networks that may or may not have bias parameters. For such general networks, however, considering only the properties of activation functions such as $\mathsf{ndf}(\sigma_{l,i})$ (as we did in §3) is insufficient to derive non-trivial bounds on the size of the incorrect and non-differentiable sets, as long as general pre-activation functions are used.

To illustrate this, consider neural networks $z_L, \widehat{z}_L : \mathbb{R} \to \mathbb{R}$ that are essentially the same as $f, \widehat{f} : \mathbb{R} \to \mathbb{R}$ with $f(w) = \mathrm{ReLU}(h(w)) - \mathrm{ReLU}(-h(w))$ and $\widehat{f}(w) = \mathrm{ReLU}(h(w))$, where $h : \mathbb{R} \to \mathbb{R}$ is some analytic pre-activation function satisfying $h(x) = 0$ and $Dh(x) = 1$ for all $x \in \mathbb{M}$. Suppose that $D^{\mathsf{AD}}\mathrm{ReLU} = \mathbf{1}_{(0,\infty)}$. Then, we have $\mathsf{inc}_\Omega(z_L) = \mathsf{ndf}_\Omega(\widehat{z}_L) = \Omega$ even though z_L and $\widehat{z}_L$ have only ≤ 2 non-differentiable points in their activation functions. The main culprit of having such large $\mathsf{inc}_\Omega(z_L)$ and $\mathsf{ndf}_\Omega(\widehat{z}_L)$, even with a tiny number of non-differentiable points in activation functions, is that z_L and $\widehat{z}_L$ use the unrealistic pre-activation function h which does not have bias parameters.

To exclude such extreme cases and focus on realistic neural networks, we will often consider *well-structured biaffine* pre-activation functions when they do not have bias parameters.

Definition 4.1. A pre-activation function $\tau_l : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l} \to \mathbb{R}^{N_l}$ is *well-structured biaffine* if there are $M_i \in \mathbb{R}^{N_{l-1}\times W_l}$ and $c_i \in \mathbb{R}$ for all $i \in [N_l]$ such that

$$\tau_{l,i}(x,u) = x^\top M_i u + c_i$$

and each column of M_i has at most one non-zero entry.

Any fully-connected or convolution layers are well-structured biaffine when they do not have bias parameters. Thus, a large class of neural networks is still under our

[8] The *Clarke subdifferential* of $f : \mathbb{R}^n \to \mathbb{R}^m$ at $x \in \mathbb{R}^n$ refers to the convex hull of $\{\lim_{n\to\infty} Df(x_n) \mid x_n \to x,\ Df(x_n) \neq \bot\} \subseteq \mathbb{R}^{m\times n}$, and an element of the Clarke subdifferential is called a *Clarke subderivative* (Clarke, 1990; Kakade & Lee, 2018).

consideration even after we impose the above restriction. Yet some pre-activation functions (e.g., normalization and attention layers) are not well-structured biaffine whether or not they have bias parameters.

We now present our results for neural networks possibly without bias parameters, extending Theorems 3.2–3.6.

4.1. Bounds for Non-Differentiable and Incorrect Sets

We first bound the density of the non-differentiable and incorrect sets in Ω, extending Theorem 3.3.

Theorem 4.2. *If a pre-activation function τ_l has bias parameters or is well-structured biaffine for all $l \in [L]$, then*

$$\frac{|\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L)|}{|\Omega|} \leq \frac{1}{|\mathbb{M}|} \sum_{(l,i)\in\mathsf{Idx}} \Big|\mathsf{ndf}(\sigma_{l,i}) \cup \big(\mathsf{bdz}(\sigma_{l,i}) \cap S_{l+1}\big)\Big|,$$

where $\mathsf{bdz}(f)$ is the boundary of f's zero set (see §2.1), and

$$S_l \triangleq \begin{cases} \emptyset & \text{if } l > L \text{ or } \tau_l \text{ has bias parameters} \\ \mathbb{R} & \text{otherwise.} \end{cases}$$

We note that if z_L has bias parameters, Theorem 4.2 reduces to Theorem 3.3 since $\mathsf{inc}_\Omega(z_L) = \emptyset$ (by Theorem 3.2) and $S_l = \emptyset$ for all l (by its definition) in such a case.

As in Theorem 3.3, the bound in Theorem 4.2 is often small for neural networks that use practical activation functions, since $|\mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i})|$ is typically small for those activation functions (e.g., 1 for ReLU and 2 for HardSigmoid).

We now show that the additional term $\mathsf{bdz}(\sigma_{l,i})$ in Theorem 4.2 is indeed necessary by providing a matching lower bound up to a constant factor.

Theorem 4.3. *For any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ with $1 \leq |\mathbb{M}| < \infty$, $n \geq 4$, and $\alpha \leq |\mathbb{M}|/(n-1)$, there is a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that satisfies*

$$\frac{|\mathsf{ndf}_\Omega(z_L)|}{|\Omega|} \geq \frac{1}{9} \cdot \frac{1}{|\mathbb{M}|} \sum_{(l,i)\in\mathsf{Idx}} \Big|\mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i})\Big|$$

and the following: (i) τ_l is well-structured biaffine without bias parameters for all $l < L$, and has bias parameters for $l = L$; (ii) z_L has $n+1$ neurons; and (iii) $|\mathsf{ndf}(\sigma_{1,i})| = \alpha$ and $|\mathsf{bdz}(\sigma_{1,i})| = 0$ for all i. We obtain the same result for (i), (ii'), and (iii'): (ii') z_L has $2n+1$ neurons; and (iii') $|\mathsf{ndf}(\sigma_{1,i})| = 0$ and $|\mathsf{bdz}(\sigma_{1,i})| = \alpha$ for all i.

We next give an intuition for why the zero set of $\sigma_{l,i}$ (from which the additional term $\mathsf{bdz}(\sigma_{l,i})$ is defined) appears in Theorem 4.2, by examining its proof. The proof consists of two main parts that extend Eqs. (3) and (4) from the proof sketch of Theorem 3.3: we first show

$$\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L) \subseteq \bigcup_{(l,i)\in\mathsf{Idx},\, c\in\mathsf{ndf}(\sigma_{l,i})} \{w \in \Omega \mid y_{l,i}(w) = c\}$$

and then find a reasonable bound on $|\Lambda_{l,i,c}|$ for $\Lambda_{l,i,c} \triangleq \{w \in \Omega \mid y_{l,i}(w) = c\}$, the set of parameters on which the pre-activation value $y_{l,i}$ touches the non-differentiable point c of $\sigma_{l,i}$. Among the two parts, the zero set of $\sigma_{l,i}$ arises from the second part (i.e., bounding $|\Lambda_{l,i,c}|$), especially when τ_l does not have bias parameters and is well-structured biaffine. For simplicity, assume that τ_l is a fully-connected layer with constant biases, i.e., $y_{l,i}(w) = \sum_{j\in[N_{l-1}]} z_{l-1,j}(w) \cdot w_{j+a} + b$ for some constants a, b. Based on this, we decompose $\Lambda_{l,i,c}$ into $\Lambda' \cup \Lambda''$:

$$\Lambda' \triangleq \{w \in \Omega \mid y_{l,i}(w) = c,\ z_{l-1,j}(w) \neq 0 \text{ for some } j\},$$
$$\Lambda'' \triangleq \{w \in \Omega \mid y_{l,i}(w) = c,\ z_{l-1,j}(w) = 0 \text{ for all } j\}.$$

Then, we can show $|\Lambda'| \leq |\mathbb{M}|^{W-1}$ as in Eq. (4), since w_{j+a} acts like a bias parameter of $y_{l,i}$ for any j with $z_{l-1,j}(w) \neq 0$. To bound $|\Lambda''|$, however, we cannot apply a similar approach due to the lack of j with $z_{l-1,j}(w) \neq 0$. Instead, we directly count the number of parameters $w \in \Omega$ achieving $z_{l-1,j}(w) = 0$ for all j (i.e., $\sigma_{l-1,j}(y_{l-1,j}(w)) = 0$ for all j), and this requires the zero set of $\sigma_{l-1,j}$. For the full proofs of Theorems 4.2 and 4.3, see Appendices B and D.

4.2. Bounds for the Incorrect Set

For the non-differentiable set, Theorems 4.2 and 4.3 provide tight bounds on its size. For the incorrect set, it turns out that we can further improve the upper bound in Theorem 4.2 and get a similar lower bound to Theorem 4.3.

Theorem 4.4. *If a pre-activation function τ_l has bias parameters or is well-structured biaffine for all $l \in [L]$, then*

$$\frac{|\mathsf{inc}_\Omega(z_L)|}{|\Omega|} \leq \frac{1}{|\mathbb{M}|} \sum_{(l,i)\in\mathsf{Idx}} \Big|\big(\mathsf{ndf}(\sigma_{l,i}) \cap S_l\big) \cup \big(\mathsf{bdz}(\sigma_{l,i}) \cap S_{l+1}\big)\Big|,$$

where S_l is defined as in Theorem 4.2.

Theorem 4.5. *For any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ with $1 \leq |\mathbb{M}| < \infty$, $n \geq 4$, and $\alpha \leq |\mathbb{M}|/(n-1)$, there is a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that satisfies*

$$\frac{|\mathsf{inc}_\Omega(z_L)|}{|\Omega|} \geq \frac{1}{13} \cdot \frac{1}{|\mathbb{M}|} \sum_{(l,i)\in\mathsf{Idx}} \Big|\mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i})\Big|$$

and the following: (i) τ_l is well-structured biaffine without bias parameters for all $l < L$, and has bias parameters for $l = L$; (ii) z_L has $2n+1$ neurons; and (iii) $|\mathsf{ndf}(\sigma_{1,i})| = \alpha$ and $|\mathsf{bdz}(\sigma_{1,i})| = 0$ for all i. We obtain the same result for (i), (ii'), and (iii'): (ii') z_L has $3n+1$ neurons; and (iii') $|\mathsf{ndf}(\sigma_{1,i})| = 0$ and $|\mathsf{bdz}(\sigma_{1,i})| = \alpha$ for all i.

We note that if z_L has bias parameters, Theorem 4.4 reduces to $|\mathsf{inc}_\Omega(z_L)| = 0$ as in Theorem 3.2 since $S_l = \emptyset$ for all l in the case. On the other hand, if z_L does not have bias parameters, then the incorrect set can be non-empty as discussed in §3.1, and more importantly, its size can be bounded by Theorem 4.4. To see why the bounds on $|\mathsf{inc}_\Omega(z_L)|$ depend on both $\mathsf{ndf}(\sigma_{l,i})$ and $\mathsf{bdz}(\sigma_{l,i})$, refer to the proofs of Theorems 4.4 and 4.5 in Appendices C and D.

4.3. Sufficient Conditions for Computing Standard Derivatives and Clarke Subderivatives

We extend Theorems 3.5 and 3.6 to general neural networks without the well-structured biaffinity restriction, by characterizing two sufficient conditions on parameters under which AD computes the standard derivative or a Clarke subderivative.

Theorem 4.6. *Let $w \in \mathbb{R}^W$. If $y_{l,i}(w) \notin \mathsf{ndf}(\sigma_{l,i})$ for all $(l,i) \in \mathsf{Idx}$ such that τ_l does not have bias parameters or $\partial^{\mathsf{AD}} z_L / \partial z_{l,i} \neq \vec{0}$ at w, then*

$$D^{\mathsf{AD}} z_L(w) = Dz_L(w) \neq \bot.$$

Theorem 4.7. *Let $w \in \mathbb{R}^W$. Assume that $D^{\mathsf{AD}}\sigma_{l,i}$ is consistent for all $(l,i) \in \mathsf{Idx}$. If $y_{l,i}(w) \notin \mathsf{ncdf}(\sigma_{l,i})$ for all $(l,i) \in \mathsf{Idx}$ such that τ_l does not have bias parameters, then*

$$D^{\mathsf{AD}} z_L(w) = \begin{cases} Dz_L(w) & \text{if } Dz_L(w) \neq \bot \\ \lim_{n\to\infty} Dz_L(w'_n) & \text{if } Dz_L(w) = \bot \\ \quad \text{for some } w'_n \to w & \end{cases}$$

and so $D^{\mathsf{AD}} z_L(w)$ is a Clarke subderivative of z_L at w. Here $\mathsf{ncdf}(f)$ denotes the set of real numbers at which $f : \mathbb{R} \to \mathbb{R}$ is not continuously differentiable.

The two sufficient conditions on w given in Theorems 4.6 and 4.7 are simple enough to be checked efficiently in practice; thus, we can use them to validate whether the output of AD is the standard derivative or a Clarke subderivative. If w does not satisfy either of the sufficient conditions, AD may not compute the standard derivative or a Clarke subderivative; the first example discussed in §3.3 illustrates both cases. We remark that the sufficient condition in Theorem 4.7 involves $\mathsf{ncdf}(\sigma_{l,i})$ (not $\mathsf{ndf}(\sigma_{l,i})$), since we use continuous differentiability (not differentiability) in the proof to properly handle the limit of derivatives $Dz_L(w'_n)$. For the proofs of Theorems 4.6 and 4.7, see Appendices E and F.

5. Related Work

The correctness of AD has been extensively studied, especially in the past few years. When a program uses only differentiable functions, AD is shown to compute its standard derivative at all real-valued inputs (Abadi & Plotkin, 2020; Barthe et al., 2020; Brunel et al., 2020; Elliott, 2018; Huot et al., 2020; Krawiec et al., 2022; Radul et al., 2023; Smeding & Vákár, 2023; Vákár, 2021). In contrast, when a program uses non-differentiable functions, the program itself can be non-differentiable, and AD can return a value different from its standard derivative, at some real-valued inputs. Nevertheless, for a large class of programs, such inputs are shown to be in a Lebesgue measure-zero subset of the real-valued input domain (Bolte & Pauwels, 2020a;b; Huot et al., 2023; Lee et al., 2020; Mazza & Pagani, 2021). All these works consider the case when inputs to AD are real-valued, while our work focuses on the case when the inputs are machine-representable.

The Clarke subdifferential and its connection to AD have been studied for decades. Some classes of functions (e.g., subdifferentially regular or strictly differentiable) are shown to admit exact chain rules for the Clarke subdifferential (e.g., Theorems 2.3.9, 2.3.10, and 2.6.6 of Clarke (1990) and Theorem 10.6 of Rockafellar & Wets (1998)), and this implies that AD always computes a Clarke subderivative for a certain class of programs. However, this class of programs is restrictive, excluding even simple neural networks (e.g., $(1 - \mathrm{ReLU}(x))^2$) (Davis et al., 2020). In contrast, our Theorem 3.6 shows that AD always computes a Clarke subderivative of neural networks with bias parameters. For piecewise differentiable functions, the Clarke subdifferential can be expressed in terms of the standard derivatives of underlying differentiable functions (e.g., Proposition 4.3.1 of Scholtes (2012)), but this result is not directly related to AD.

A variety of algorithms (other than AD) have been proposed to compute a Clarke subgradient of a scalar program, correctly and efficiently. For a large class of programs $f : \mathbb{R}^n \to \mathbb{R}$ and an input $x \in \mathbb{R}^n$, the algorithm by Kakade & Lee (2018) computes a Clarke subgradient of f at x in time $\mathcal{O}(T)$ almost surely, while the algorithms by Barton et al. (2018); Khan & Barton (2015) compute the quantity in time $\mathcal{O}(nT)$ deterministically, where T denotes time to evaluate $f(x)$. Our Theorem 3.6 provides a relevant result as described above, but we point out that our work is about analyzing the correctness of vanilla (forward/reverse-mode) AD, not about proposing a new algorithm.

Recently, Bertoin et al. (2021) empirically studied how the choice of $D^{\mathsf{AD}}\mathrm{ReLU}(0)$ changes the output of AD and the training of neural networks. In contrast, our work theoretically studies the correctness of AD. Further connections between this and our work are discussed in §6.

6. Discussion

Connections to Bertoin et al. (2021). Bertoin et al. empirically studied the *bifurcation zone* of a neural network with ReLU, given an input dataset: the set of the network parameters on which the output of AD using $D^{\mathsf{AD}}\mathrm{ReLU}(0) = 0$ is

different from that using $D^{\mathtt{AD}}\mathrm{ReLU}(0) = 1$ for some input data. The bifurcation zone is closely related to the non-differentiable and incorrect sets as follows: the bifurcation zone (over machine-representable parameters) is always a subset of the union of the non-differentiable set and two incorrect sets (one for $D^{\mathtt{AD}}\mathrm{ReLU}(0) = 0$ and the other for $D^{\mathtt{AD}}\mathrm{ReLU}(0) = 1$) over all input data in the given dataset.

For various neural networks (MLP, VGG, ResNet) and datasets (MNIST, CIFAR10, SVHN, ImageNet), Bertoin et al. estimated the density of the bifurcation zone over 32-bit floating-point parameters (i.e., the number of 32-bit parameters in the bifurcation zone over the total number of 32-bit parameters) using Monte Carlo sampling. They reported two results among many others: when AD uses 64-bit precision in its computation, the estimated density is exactly 0 in all cases they considered; and when AD uses 32- or 16-bit precision, the estimated density is often large and even goes up to 1. The first result is consistent with our results: if we use 32-bit parameters, the non-differentiable and incorrect sets would often have small densities in practice. Meanwhile, the second result does not contradict our results, since our results assume that AD computes its output without any rounding errors. Given these observations, it would be an interesting direction to rigorously study the correctness of AD under floating-point operations.

Extensions. As mentioned in §2.2, all our theorems except for those on lower bounds (i.e., Theorems 3.4, 4.3, and 4.5) continue to hold even if we replace z_L in their conclusions by $\ell \circ z_L$ for any analytic $\ell : \mathbb{R}^{N_L} \to \mathbb{R}^m$. Among them, Theorems 3.3 and 4.2 are easily extended to a more general case with multiple input data: they remain valid even if we replace z_L in their conclusions by $\ell(z_L(\cdot; x_1), \ldots, z_L(\cdot; x_k))$ for any $x_1, \ldots, x_k \in \mathbb{R}^{N_0}$ and analytic $\ell : \mathbb{R}^{N_L} \to \mathbb{R}^m$, where we need to multiply k to the upper bounds in the theorems. The remaining theorems (i.e., Theorems 3.2, 3.5, 3.6, 4.4, 4.6, and 4.7), on the other hand, are not easily extended to the case with multiple input data, at least based on our current proofs. Studying such extensions could be another interesting future direction.

Limitations. Our results have some limitations. For example, all of our results are for a class of neural networks consisting of alternating analytic pre-activation functions and pointwise continuous activation functions. Hence, if a network contains non-pointwise activation functions (e.g., MaxPool) or a residual connection bypassing a non-analytic activation function (e.g., ReLU), then our results may not be directly applicable. Our results for general neural networks (e.g., Theorems 4.2 and 4.4) additionally assume pre-activation functions to have bias parameters or to be well-structured biaffine, which does not allow, e.g., BatchNorm layers and attention layers without bias parameters. Nevertheless, we believe that our results still cover a large class of neural networks, especially compared to prior works studying theoretical aspects of neural networks (Jacot et al., 2018; Kidger & Lyons, 2020; Laurent & von Brecht, 2018; Lu et al., 2017; Park et al., 2021). We believe extending our work to more general neural networks is an interesting direction for future work.

7. Conclusion

In this paper, we theoretically study for the first time the correctness of AD for neural networks with machine-representable parameters. In particular, we provide various theoretical results on the incorrect and non-differentiable sets of a neural network, as well as closely related questions such as when AD is correct and what it computes. Our results have two major practical implications: AD is correct at most machine-representable parameters when applied to neural networks, and it is correct more often if more layers of the network have bias parameters. Furthermore, our theoretical analyses suggest new applications of AD for identifying differentiability and computing Clarke subderivatives, not only for machine-representable parameters but also for any real-valued ones.

Acknowledgments

We thank anonymous reviewers for providing helpful comments. WL and AA were supported by the Advanced Simulation and Computing (ASC) program of the US Department of Energy's National Nuclear Security Administration (NNSA) via the PSAAP-III Center at Stanford, Grant No. DE-NA0002373 and by the Department of Energy's Office of Advanced Scientific Computing Research (ASCR) under contract DE-AC03-76SF00515. SP was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University).

References

Abadi, M. and Plotkin, G. D. A simple differentiable programming language. *Proceedings of the ACM on Programming Languages*, 4(POPL):38:1–38:28, 2020.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P. A., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: A system for large-scale machine learning. In *Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 265–283, 2016.

Barthe, G., Crubillé, R., Lago, U. D., and Gavazzo, F. On the

versatility of open logical relations - continuity, automatic differentiation, and a containment theorem. In *European Symposium on Programming (ESOP)*, pp. 56–83, 2020.

Barton, P. I., Khan, K. A., Stechlinski, P., and Watson, H. A. J. Computationally relevant generalized derivatives: Theory, evaluation and applications. *Optimization Methods and Software*, 33(4-6):1030–1072, 2018.

Baydin, A. G., Pearlmutter, B. A., and Siskind, J. M. Diffsharp: An AD library for .NET languages. In *International Conference on Algorithmic Differentiation (AD)*, 2016. Also arXiv:1611.03423.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18:153:1–153:43, 2017.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. Theano: A CPU and GPU math compiler in Python. In *Python in Science Conference (SciPy)*, pp. 18–24, 2010.

Bertoin, D., Bolte, J., Gerchinovitz, S., and Pauwels, E. Numerical influence of ReLU'(0) on backpropagation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 468–479, 2021.

Bolte, J. and Pauwels, E. Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning. *Mathematical Programming*, 188:19–51, 2020a.

Bolte, J. and Pauwels, E. A mathematical model for automatic differentiation in machine learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 10809–10819, 2020b.

Bolte, J., Boustany, R., Pauwels, E., and Pesquet-Popescu, B. On the complexity of nonsmooth automatic differentiation. In *International Conference on Learning Representations (ICLR)*, 2023.

Brunel, A., Mazza, D., and Pagani, M. Backpropagation in the simply typed lambda-calculus with linear negation. *Proceedings of the ACM on Programming Languages*, 4 (POPL):64:1–64:27, 2020.

Burden, R. L., Faires, J. D., and Burden, A. M. *Numerical analysis*. Cengage learning, 10th edition, 2015.

Clarke, F. H. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Clarke, F. H. *Optimization and nonsmooth analysis*. Classics in Applied Mathematics: Volume 5. SIAM, 1990.

Collobert, R., Kavukcuoglu, K., and Farabet, C. Torch7: A Matlab-like environment for machine learning. In *NIPS BigLearn Workshop*, 2011.

Cui, Y. and Pang, J.-S. *Modern nonconvex nondifferentiable optimization*. MOS-SIAM Series on Optimization. SIAM, 2021.

Davis, D., Drusvyatskiy, D., Kakade, S. M., and Lee, J. D. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1): 119–154, 2020.

Elliott, C. The simple essence of automatic differentiation. *Proceedings of the ACM on Programming Languages*, 2 (ICFP):70:1–70:29, 2018.

Frostig, R., Johnson, M., and Leary, C. Compiling machine learning programs via high-level tracing. In *SysML Conference*, 2018.

Griewank, A. and Walther, A. *Evaluating derivatives: Principles and techniques of algorithmic differentiation*. SIAM, 2nd edition, 2008.

Hascoët, L. and Pascual, V. The Tapenade automatic differentiation tool: Principles, model, and specification. *ACM Transactions on Mathematical Software*, 39(3):20:1–20:43, 2013.

Huot, M., Staton, S., and Vákár, M. Correctness of automatic differentiation via diffeologies and categorical gluing. In *International Conference on Foundations of Software Science and Computation Structures (FoSSaCS)*, pp. 319–338, 2020.

Huot, M., Lew, A. K., Mansinghka, V. K., and Staton, S. ωPAP spaces: Reasoning denotationally about higher-order, recursive probabilistic and differentiable programs. *arXiv:2302.10636*, 2023.

Jacot, A., Hongler, C., and Gabriel, F. Neural tangent kernel: Convergence and generalization in neural networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 8580–8589, 2018.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R. B., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In *International Conference on Multimedia (MM)*, pp. 675–678, 2014.

Kakade, S. M. and Lee, J. D. Provably correct automatic sub-differentiation for qualified programs. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7125–7135, 2018.

Khan, K. A. and Barton, P. I. A vector forward mode of automatic differentiation for generalized derivative evaluation. *Optimization Methods and Software*, 30(6): 1185–1212, 2015.

Kidger, P. and Lyons, T. Universal approximation with deep narrow networks. In *Conference on Learning Theory (COLT)*, pp. 2306–2327, 2020.

Krawiec, F., Jones, S. P., Krishnaswami, N., Ellis, T., Eisenberg, R. A., and Fitzgibbon, A. W. Provably correct, asymptotically efficient, higher-order reverse-mode automatic differentiation. *Proceedings of the ACM on Programming Languages*, 6(POPL):48:1–48:30, 2022.

Laurent, T. and von Brecht, J. The multilinear structure of ReLU networks. In *International Conference on Machine Learning (ICML)*, pp. 2914–2922, 2018.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Lee, W., Yu, H., Rival, X., and Yang, H. On correctness of automatic differentiation for non-differentiable functions. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6719–6730, 2020.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6232–6240, 2017.

Maclaurin, D., Duvenaud, D., and Adams, R. P. Autograd: Effortless gradients in Numpy. In *ICML AutoML Workshop*, 2015.

Mazza, D. and Pagani, M. Automatic differentiation in PCF. *Proceedings of the ACM on Programming Languages*, 5 (POPL):28:1–28:27, 2021.

Park, S., Yun, C., Lee, J., and Shin, J. Minimum width for universal approximation. In *International Conference on Learning Representations (ICLR)*, 2021.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

Pearlmutter, B. A. and Siskind, J. M. Reverse-mode AD in a functional framework: Lambda the ultimate backpropagator. *ACM Transactions on Programming Languages and Systems*, 30(2):7:1–7:36, 2008.

Radul, A., Paszke, A., Frostig, R., Johnson, M. J., and Maclaurin, D. You only linearize once: Tangents transpose to gradients. *Proceedings of the ACM on Programming Languages*, 7(POPL):43:1–43:29, 2023.

Revels, J., Lubin, M., and Papamarkou, T. Forward-mode automatic differentiation in Julia. *arXiv:1607.07892*, 2016.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*. A Series of Comprehensive Studies in Mathematics: Volume 317. Springer Science & Business Media, 1998.

Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

Scholtes, S. *Introduction to piecewise differentiable equations*. SpringerBriefs in Optimization. Springer Science & Business Media, 2012.

Seide, F. and Agarwal, A. CNTK: Microsoft's open-source deep-learning toolkit. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2135, 2016.

Slusanschi, E. and Dumitrel, V. ADiJaC – automatic differentiation of Java classfiles. *ACM Transactions on Mathematical Software*, 43(2):9:1–9:33, 2016.

Smeding, T. and Vákár, M. Efficient dual-numbers reverse AD via well-known program transformations. *Proceedings of the ACM on Programming Languages*, 7(POPL): 54:1–54:28, 2023.

Tokui, S., Okuta, R., Akiba, T., Niitani, Y., Ogawa, T., Saito, S., Suzuki, S., Uenishi, K., Vogel, B., and Vincent, H. Y. Chainer: A deep learning framework for accelerating the research cycle. In *International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 2002–2011, 2019.

Vákár, M. Reverse AD at higher types: Pure, principled and denotationally correct. In *European Symposium on Programming (ESOP)*, pp. 607–634, 2021.

van Merrienboer, B., Moldovan, D., and Wiltschko, A. B. Tangent: Automatic differentiation using source-code transformation for dynamically typed array programming. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6259–6268, 2018.

Walther, A. and Griewank, A. Getting started with ADOL-C. In *Combinatorial Scientific Computing*, chapter 7, pp. 181–202. Chapman & Hall/CRC Computational Science, 2012.

Contents of Appendix

A Formal Setup **13**
A.1 Piecewise-Analytic Functions . . . 13
A.2 Neural Networks . . . 14
A.3 Automatic Differentiation . . . 16

B Upper Bounds on $|\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L)|$ **18**
B.1 Lemmas (Basic) . . . 18
B.2 Lemmas (Technical: Part 1) . . . 18
B.3 Theorem 3.3 (Main Lemmas) . . . 19
B.4 Theorem 3.3 (Main Proof) . . . 20
B.5 Lemmas (Technical: Part 2) . . . 20
B.6 Theorem 4.2 (Main Lemmas) . . . 23
B.7 Theorem 4.2 (Main Proof) . . . 24

C Upper Bounds on $|\mathsf{inc}_\Omega(z_L)|$ **25**
C.1 Lemmas (Basic) . . . 25
C.2 Lemmas (Technical: Part 1) . . . 25
C.3 Lemmas (Technical: Part 2) . . . 27
C.4 Theorem 3.2 (Main Lemmas) . . . 30
C.5 Theorem 3.2 (Main Proof) . . . 31
C.6 Lemmas (Technical: Part 3) . . . 32
C.7 Theorem 4.4 (Main Lemma) . . . 33
C.8 Theorem 4.4 (Main Proof) . . . 34

D Lower Bounds on $|\mathsf{ndf}_\Omega(z_L)|$ and $|\mathsf{inc}_\Omega(z_L)|$ **35**
D.1 Theorem 3.4 (Main Proof) . . . 35
D.2 Theorem 4.3 (Main Proof) . . . 36
D.3 Theorem 4.5 (Main Proof) . . . 37

E Computation of Standard Derivatives **39**
E.1 Lemmas (Basic) . . . 39
E.2 Lemmas (Technical: Part 1) . . . 39
E.3 Lemmas (Technical: Part 2) . . . 40
E.4 Theorems 3.5 and 4.6 (Main Lemmas) . . . 41
E.5 Theorems 3.5 and 4.6 (Main Proofs) . . . 42

F Computation of Clarke Subderivatives **43**
F.1 Lemmas (Basic) . . . 43
F.2 Lemmas (Technical) . . . 44
F.3 Theorems 3.6 and 4.7 (Main Lemmas) . . . 45
F.4 Theorems 3.6 and 4.7 (Main Proofs) . . . 47

A. Formal Setup

In the appendix, we use the following notation. For $A \subseteq \mathbb{R}^n$, $int(A)$ and $bd(A)$ denote the interior and the boundary of A.

A.1. Piecewise-Analytic Functions

Definition A.1. For $A \subseteq \mathbb{R}^n$, define $pbd(A)$ as

$$pbd(A) \triangleq A \setminus int(A).$$

We call $pbd(A)$ the *proper boundary of* A. Note that $pbd(A) = bd(A) \cap A$ holds for any A.

Definition A.2. A function $f : \mathbb{R} \to \mathbb{R}$ is *piecewise-differentiable* (or *piecewise-C^1*) if there exist $n \in \mathbb{N}$, a partition $\{A_i\}_{i\in[n]}$ of $\mathbb{R}$ consisting of non-empty intervals, and differentiable (or C^1) functions $\{f_i : \mathbb{R} \to \mathbb{R}\}_{i\in[n]}$ such that $f = f_i$ on A_i for all $i \in [n]$. We call such $\{(A_i, f_i)\}_{i\in[n]}$ a *piecewise-differentiable (or piecewise-C^1) representation of* f. Moreover, for an extended derivative $g : \mathbb{R} \to \mathbb{R}$ of f, we say that the representation $\{(A_i, f_i)\}_{i\in[n]}$ *defines* g if $g = Df_i$ on A_i for all $i \in [n]$. We define a *piecewise-analytic representation of* f in a similar way.

Lemma A.3. *Let $\{A_i\}_{i\in S}$ be any partition of $\mathbb{R}^n$. Then,*

$$\bigcup_{i\in S} bd(A_i) = \bigcup_{i\in S} pbd(A_i). \tag{5}$$

Proof. The direction $\supseteq$ is clear, since $pbd(X) \subseteq bd(X)$ for any $X \subseteq \mathbb{R}^n$. To prove the other direction $\subseteq$, it suffices to show that for any $i \in S$ and $x \in bd(A_i)$, we have $x \in pbd(A_j)$ for some $j \in S$. Here we assume $x \notin A_i$; if not, choosing $j = i$ completes the proof. Let $j \in S$ be the index with $x \in A_j$, where such j always exists since $\{A_i\}_{i\in S}$ is a partition of $\mathbb{R}$. Then, it suffices to show $x \in bd(A_j)$, because this and $x \in A_j$ implies $x \in pbd(A_j)$. To prove $x \in bd(A_j)$, consider any open neighborhood $U \subseteq \mathbb{R}^n$ of x. Then, there is $x' \in U \cap A_i$ (by $x \in bd(A_i)$ and $x \notin A_i$). This implies that $x' \notin U \cap A_j$ (by $A_i \cap A_j = \emptyset$ from $i \neq j$) and $x \in U \cap A_j$ (by $x \in A_j$). Hence, we have $x \in bd(A_j)$ as desired. □

Theorem A.4. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous, piecewise-analytic function, and $g : \mathbb{R} \to \mathbb{R}$ be an extended derivative of f. Then, the following hold.*

(i) There is a piecewise-differentiable representation $\{(A_i, f_i)\}_{i\in[n]}$ of f that defines g and satisfies the following:

$$\bigcup_{i\in[n]} bd(A_i) = \bigcup_{i\in[n]} pbd(A_i) = \mathsf{ndf}(f).$$

(ii) If g is consistent, there is a piecewise-C^1 representation $\{(A_i, f_i)\}_{i\in[n]}$ of f that defines g and satisfies the following:

$$\bigcup_{i\in[n]} bd(A_i) = \bigcup_{i\in[n]} pbd(A_i) = \mathsf{ncdf}(f), \qquad int(A_i) \neq \emptyset \text{ for all } i \in [n],$$

where $\mathsf{ncdf}(f) \subseteq \mathbb{R}$ denotes the set of real numbers at which $f : \mathbb{R} \to \mathbb{R}$ is not continuously differentiable.

Proof. We prove the two claims as follows. Note that by Lemma A.3, we do not need to prove the equality between the union of $bd(A_i)$ and that of $pbd(A_i)$ in the claims.

Claim (i). Let $\{(\widetilde{A}_i, \widetilde{f}_i)\}_{i\in[\widetilde{n}]}$ be a piecewise-analytic representation of f that defines g and satisfies

$$(\widetilde{A}_1, \ldots, \widetilde{A}_{\widetilde{n}}) = \Big((x_0, x_1), \ldots, (x_k, x_{k+1}), \{x_1\}, \ldots, \{x_k\}\Big)$$

for some $-\infty = x_0 < x_1 < \cdots < x_k < x_{k+1} = \infty$. Such a representation always exists, because f is piecewise-analytic and g is an extended derivative of f. Note that $\mathsf{ndf}(f) \subseteq \{x_1, \ldots, x_k\}$ because f is differentiable on (x_{i-1}, x_i) for all $i \in [k+1]$ (since $\widetilde{f}_i$ is analytic and it coincides with f on $\widetilde{A}_i = (x_{i-1}, x_i)$). We then construct $\{(A_i, f_i)\}_{i\in[n]}$ from $\{(\widetilde{A}_i, f_i)\}_{i\in[\widetilde{n}]}$, by merging all adjacent intervals $\widetilde{A}_i$ (and associated functions $\widetilde{f}_i$) into a single interval (and a single function) such that the class of the singleton interval in $\{A_i\}$ are the same as $\{\{x\} \mid x \in \mathsf{ndf}(f)\}$. Then,

$$\bigcup_{i\in[n]} pbd(A_i) = \bigcup_{x\in\mathsf{ndf}(f)} \{x\} = \mathsf{ndf}(f)$$

by construction; f_i is differentiable for all $i \in [n]$; and $\{(A_i, f_i)\}_{i\in[n]}$ defines g since g is an extended derivative of f. Hence, $\{(A_i, f_i)\}_{i\in[n]}$ is a piecewise-differentiable representation of f that defines g and satisfies the equation in the statement.

Claim (ii). By a similar argument, there is a piecewise-C^1 representation $\{(\widetilde{A}_i, \widetilde{f}_i)\}_{i\in[\widetilde{n}]}$ of f that defines g and satisfies

$$\bigcup_{i\in[\widetilde{n}]} pbd(\widetilde{A}_i) = \mathsf{ncdf}(f).$$

Note that here we need $\mathsf{ncdf}(f)$ (instead of $\mathsf{ndf}(f)$) in the above equation, to obtain a piecewise-C^1 (instead of piecewise-differentiable) representation of f. We then construct $\{(A_i, f_i)\}_{i\in[n]}$ from $\{(\widetilde{A}_i, \widetilde{f}_i)\}_{i\in[\widetilde{n}]}$, by merging each singleton interval $\widetilde{A}_i$ (and the associated function $\widetilde{f}_i$) with one of the two adjacent intervals (and its associated function) such that $\{(A_i, f_i)\}_{i\in[n]}$ defines g. Such a construction always exists, because f is continuous, g is consistent, and $\widetilde{f}_i$ is C^1 for all $i \in [\widetilde{n}]$. Then,

$$\bigcup_{i\in[n]} pbd(A_i) = \mathsf{ncdf}(f), \qquad int(A_i) \neq \emptyset \qquad \text{for all } i \in [n]$$

by construction; and f_i is C^1 for all $i \in [\widetilde{n}]$ since f is continuous. Hence, $\{(A_i, f_i)\}_{i\in[n]}$ is a piecewise-C^1 representation of f that defines g and satisfies the equation given in the statement. □

A.2. Neural Networks

Definition A.5. For each $(l, i) \in \mathsf{Idx}$, let

$$\{(\mathcal{I}^k_{l,i}, \sigma^k_{l,i})\}_{k\in[K_{l,i}]}$$

be a piecewise-differentiable representation of $\sigma_{l,i} : \mathbb{R} \to \mathbb{R}$ that defines $D^{\mathtt{AD}}\sigma_{l,i}$ (an extended derivative of $\sigma_{l,i}$ defined in §2.3), where $K_{l,i} \in \mathbb{N}$, $\mathcal{I}^k_{l,i} \subseteq \mathbb{R}$, and $\sigma^k_{l,i} : \mathbb{R} \to \mathbb{R}$. We assume that the representation satisfies the following:

$$\bigcup_{k\in[K_{l,i}]} bd(\mathcal{I}^k_{l,i}) = \bigcup_{k\in[K_{l,i}]} pbd(\mathcal{I}^k_{l,i}) = \mathsf{ndf}(\sigma_{l,i}).$$

Note that such a representation always exists by Theorem A.4.

Definition A.6. Define Γ, the set of indices denoting which piece of each activation function is used, as

$$\Gamma \triangleq \{\gamma : \mathsf{Idx} \to \mathbb{N} \mid \gamma(l, i) \in [K_{l,i}] \text{ for all } (l, i) \in \mathsf{Idx}\}.$$

Definition A.7. Let $\gamma \in \Gamma$ and $l \in [L]$. Define $\mathcal{R}^\gamma \subseteq \mathbb{R}^W$, $y^\gamma_l, z^\gamma_l : \mathbb{R}^W \to \mathbb{R}^{N_l}$, $\sigma^\gamma_l : \mathbb{R}^{N_l} \to \mathbb{R}^{N_l}$ as:

$$\begin{aligned}
\mathcal{R}^\gamma &\triangleq \{w \in \mathbb{R}^W \mid y_{l,i}(w) \in \mathcal{I}^{\gamma(l,i)}_{l,i} \text{ for all } (l, i) \in \mathsf{Idx}\},\\
y^\gamma_l(w) &\triangleq \tau_l\big(z^\gamma_{l-1}(w), \pi_l(w)\big), \qquad z^\gamma_l(w) \triangleq \sigma^\gamma_l\big(y^\gamma_l(w)\big),\\
\sigma^\gamma_l(x) &\triangleq \big(\sigma^{\gamma(l,1)}_{l,1}(x_1), \ldots, \sigma^{\gamma(l,N_l)}_{l,N_l}(x_{N_l})\big),
\end{aligned}$$

where $\pi_l : \mathbb{R}^W \to \mathbb{R}^{W_l}$ denotes the projection function that extracts $w_l \in \mathbb{R}^{W_l}$ from $(w_1, \ldots, w_L) \in \mathbb{R}^W$, and $z^\gamma_0 : \mathbb{R}^W \to \mathbb{R}^{N_0}$ is defined as $z^\gamma_0 \triangleq z_0$.

Lemma A.8. *$\{\mathcal{R}^\gamma\}_{\gamma\in\Gamma}$ is a partition of $\mathbb{R}^W$.*

Proof. This follows immediately from that $\{\mathcal{I}^k_{l,i}\}_{k\in[K_{l,i}]}$ is a partition of $\mathbb{R}$ for all $(l, i) \in \mathsf{Idx}$ (since $\{(\mathcal{I}^k_{l,i}, \sigma^k_{l,i})\}_{k\in[K_{l,i}]}$ is a representation of $\sigma_{l,i}$). □

Lemma A.9. *For all $l \in [L]$ and $\gamma \in \Gamma$, y_l and z_l are continuous, and y^γ_l and z^γ_l are differentiable.*

Proof. The continuity of y_l and z_l follows directly from that $\tau_{l'}$, $\pi_{l'}$, and $\sigma_{l',i'}$ are continuous for all $(l', i') \in \mathsf{Idx}$. Similarly, the differentiability of y^γ_l and z^γ_l follows directly from that $\tau_{l'}$, $\pi_{l'}$, and $\sigma^{k'}_{l',i'}$ are differentiable for all $(l', i') \in \mathsf{Idx}$ and $k' \in [K_{l',i'}]$. □

Lemma A.10. *Let $\gamma \in \Gamma$. Then,*

$$\mathcal{R}^\gamma = \{w \in \mathbb{R}^W \mid y^\gamma_{l,i}(w) \in \mathcal{I}^{\gamma(l,i)}_{l,i} \textit{ for all } (l, i) \in \mathsf{Idx}\}.$$

Note that the RHS uses $y^\gamma_{l,i}$ instead of $y_{l,i}$.

Proof. Let $\gamma \in \Gamma$. Define $\mathcal{R}^{\gamma}_{\le l}, \mathcal{S}^{\gamma}_{\le l} \subseteq \mathbb{R}^W$ for $l \in [L]$ as

$$\mathcal{R}^{\gamma}_{\le l} \triangleq \{w \in \mathbb{R}^W \mid y_{l',i}(w) \in \mathcal{I}^{\gamma(l',i)}_{l',i} \text{ for all } (l',i) \in \mathsf{Idx} \text{ with } l' \le l\},$$
$$\mathcal{S}^{\gamma}_{\le l} \triangleq \{w \in \mathbb{R}^W \mid y^{\gamma}_{l',i}(w) \in \mathcal{I}^{\gamma(l',i)}_{l',i} \text{ for all } (l',i) \in \mathsf{Idx} \text{ with } l' \le l\}.$$

It suffices to show the following claim which generalizes this lemma: all $l \in [L]$,

$$y_l(w) = y^{\gamma}_l(w) \text{ for all } w \in \mathcal{R}^{\gamma}_{\le l-1}, \qquad \mathcal{R}^{\gamma}_{\le l} = \mathcal{S}^{\gamma}_{\le l}.$$

We prove this claim by induction on l.

Case $l = 1$**.** Since $z_0 = z^{\gamma}_0$, we have the first claimed equation:

$$y_1(w) = \tau_1(z_0(w), w_1) = \tau_1(z^{\gamma}_0(w), w_1) = y^{\gamma}_1(w)$$

for all $w \in \mathbb{R}^W$. From this, we have the second claimed equation:

$$\mathcal{R}^{\gamma}_{\le 1} = \bigcap_{i \in [N_1]} \{w \in \mathbb{R}^W \mid y_{1,i}(w) \in \mathcal{I}^{\gamma(1,i)}_{1,i}\} = \bigcap_{i \in [N_1]} \{w \in \mathbb{R}^W \mid y^{\gamma}_{1,i}(w) \in \mathcal{I}^{\gamma(1,i)}_{1,i}\} = \mathcal{S}^{\gamma}_{\le 1}.$$

Case $l > 1$**.** We obtain the first claimed equation as follows: for all $w \in \mathcal{R}^{\gamma}_{\le l-1}$,

$$\begin{aligned} y^{\gamma}_l(w) &= \tau_l\big(\sigma^{\gamma}_{l-1}(y^{\gamma}_{l-1}(w)), \pi_l(w)\big) \\ &= \tau_l\big(\sigma^{\gamma}_{l-1}(y_{l-1}(w)), \pi_l(w)\big) \\ &= \tau_l\big(\sigma_{l-1}(y_{l-1}(w)), \pi_l(w)\big) = y_l(w). \end{aligned}$$

Here the second line uses $y^{\gamma}_{l-1}(w) = y_{l-1}(w)$, which holds by induction hypothesis on $l-1$ with $w \in \mathcal{R}^{\gamma}_{\le l-1} \subseteq \mathcal{R}^{\gamma}_{\le l-2}$. And the third line uses $\sigma^{\gamma(l-1,i)}_{l-1,i}(y_{l-1,i}(w)) = \sigma_{l-1,i}(y_{l-1,i}(w))$ for all $i \in [N_{l-1}]$, which holds because $y_{l-1,i}(w) \in \mathcal{I}^{\gamma(l-1,i)}_{l-1,i}$ (by $w \in \mathcal{R}^{\gamma}_{\le l-1}$) and $\{(\mathcal{I}^k_{l-1,i}, \sigma^k_{l-1,i})\}_{k \in [K_{l-1,i}]}$ is a representation of $\sigma_{l-1,i}$. Using this result, we obtain the second claimed equation as follows:

$$\begin{aligned} \mathcal{R}^{\gamma}_{\le l} &= \mathcal{R}^{\gamma}_{\le l-1} \cap \bigcap_{i \in [N_l]} \{w \in \mathcal{R}^{\gamma}_{\le l-1} \mid y_{l,i}(w) \in \mathcal{I}^{\gamma(l,i)}_{l,i}\} \\ &= \mathcal{R}^{\gamma}_{\le l-1} \cap \bigcap_{i \in [N_l]} \{w \in \mathcal{R}^{\gamma}_{\le l-1} \mid y^{\gamma}_{l,i}(w) \in \mathcal{I}^{\gamma(l,i)}_{l,i}\} \\ &= \mathcal{S}^{\gamma}_{\le l-1} \cap \bigcap_{i \in [N_l]} \{w \in \mathcal{S}^{\gamma}_{\le l-1} \mid y^{\gamma}_{l,i}(w) \in \mathcal{I}^{\gamma(l,i)}_{l,i}\} = \mathcal{S}^{\gamma}_{\le l}, \end{aligned}$$

where the second line uses $y^{\gamma}_{l,i}(w) = y_{l,i}(w)$ for all $w \in \mathcal{R}^{\gamma}_{\le l-1}$, which we already proved, and the third line uses $\mathcal{R}^{\gamma}_{\le l-1} = \mathcal{S}^{\gamma}_{\le l-1}$, which holds by induction hypothesis on $l-1$. □

Lemma A.11. *Let $\gamma \in \Gamma$. Then, for all $l \in [L]$ and $w \in \mathcal{R}^{\gamma}$,*

$$y^{\gamma}_l(w) = y_l(w), \qquad z^{\gamma}_l(w) = z_l(w).$$

Proof. Let $\gamma \in \Gamma$. The claim shown in the proof of Lemma A.10 implies the first part of the conclusion (since $\mathcal{R}^{\gamma}_{\le l-1} \supseteq \mathcal{R}^{\gamma}$): for all $l \in [L]$ and $w \in \mathcal{R}^{\gamma}$, $y^{\gamma}_l(w) = y_l(w)$. From this, we obtain the second part of the conclusion: for all $l \in [L]$ and $w \in \mathcal{R}^{\gamma}$,

$$z^{\gamma}_l(w) = \sigma^{\gamma}_l(y^{\gamma}_l(w)) = \sigma^{\gamma}_l(y_l(w)) = \sigma_l(y_l(w)) = z_l(w),$$

where the second equality follows from the first part of the conclusion, and the third equality from $\sigma^{\gamma(l,i)}_{l,i}(y_{l,i}(w)) = \sigma_{l,i}(y_{l,i}(w))$ which holds because $y_{l,i}(w) \in \mathcal{I}^{\gamma(l,i)}_{l,i}$ (by $w \in \mathcal{R}^{\gamma}$) and $\{(\mathcal{I}^k_{l,i}, \sigma^k_{l,i})\}_{k \in [K_{l,i}]}$ is a representation of $\sigma_{l,i}$. □

A.3. Automatic Differentiation

As discussed in §1, AD operates not on mathematical functions, but on programs that represent those functions. To this end, we define a program $\mathtt{P}$ that represents a function from $\mathbb{R}^W$ to $\mathbb{R}$ as follows:

$$\mathtt{P} ::= r \mid \mathtt{w}_{l,j} \mid f\,(\mathtt{P}_1, \ldots, \mathtt{P}_n)$$

where $r \in \mathbb{R}$, $l \in [L]$, $j \in [W_l]$, $f \in \{\tau_{l,i}, \sigma_{l,i} \mid (l,i) \in \mathsf{Idx}\}$, and $n \in \mathbb{N}$. This definition says that a program $\mathtt{P}$ can be either a real-valued constant r, a real-valued parameter $\mathtt{w}_{l,j}$, or the application of a function $f : \mathbb{R}^n \to \mathbb{R}$ to subprograms $\mathtt{P}_1, \ldots, \mathtt{P}_n$. In this paper, we focus on particular programs $\mathtt{P}_{y_{l,i}}$ and $\mathtt{P}_{z_{l,i}}$ that represent the functions $y_{l,i}(\,\cdot\,; c), z_{l,i}(\,\cdot\,; c) : \mathbb{R}^W \to \mathbb{R}$ and are defined in a canonical way as follows:

$$\begin{aligned}
\mathtt{P}_{y_{l,i}} &\triangleq \tau_{l,i}\,(\mathtt{P}_{z_{l-1,1}}, \ldots, \mathtt{P}_{z_{l-1,N_{l-1}}}, \mathtt{w}_{l,1}, \ldots, \mathtt{w}_{l,W_l})\,, \\
\mathtt{P}_{z_{l,i}} &\triangleq \sigma_{l,i}\,(\mathtt{P}_{y_{l,i}})\,,
\end{aligned}$$

where $\mathtt{P}_{z_{0,i'}} \triangleq c_{i'}$ for $i' \in [N_0]$ represents the constant function $z_{0,i'}(\,\cdot\,; c) : \mathbb{R}^W \to \mathbb{R}$.

Given a program $\mathtt{P}$, we define $[\![\mathtt{P}]\!] : \mathbb{R}^W \to \mathbb{R}$ as the function represented by $\mathtt{P}$, and $[\![\mathtt{P}]\!]^{\mathtt{AD}} : \mathbb{R}^W \to \mathbb{R}^{1\times W}$ as the function that AD essentially computes when applied to $\mathtt{P}$. These functions are defined inductively as follows (Abadi & Plotkin, 2020; Baydin et al., 2017; Lee et al., 2020):

$$\begin{aligned}
[\![r]\!](w) &\triangleq r, \\
[\![\mathtt{w}_{l,j}]\!](w) &\triangleq w_{l,j}, \\
[\![f\,(\mathtt{P}_1, \ldots, \mathtt{P}_n)]\!](w) &\triangleq f\big([\![\mathtt{P}_1]\!](w), \ldots, [\![\mathtt{P}_n]\!](w)\big), \\
[\![r]\!]^{\mathtt{AD}}(w) &\triangleq \mathbb{0}, \\
[\![\mathtt{w}_{l,j}]\!]^{\mathtt{AD}}(w) &\triangleq \mathbb{1}_{l,j}, \\
[\![f\,(\mathtt{P}_1, \ldots, \mathtt{P}_n)]\!]^{\mathtt{AD}}(w) &\triangleq D^{\mathtt{AD}} f\big([\![\mathtt{P}_1]\!](w), \ldots, [\![\mathtt{P}_n]\!](w)\big) \cdot \big[[\![\mathtt{P}_1]\!]^{\mathtt{AD}}(w) \,/\, \cdots \,/\, [\![\mathtt{P}_n]\!]^{\mathtt{AD}}(w)\big].
\end{aligned}$$

Here $w_{l,j} \in \mathbb{R}$ is defined as $(w_{1,1}, w_{1,2} \ldots, w_{L,W_L}) \triangleq w$, $\mathbb{0}, \mathbb{1}_{l,j} \in \mathbb{R}^{1\times W}$ denote the zero matrix and the matrix whose entries are all zeros except for a single one at the $(W_1 + \cdots + W_{l-1} + j)$-th entry, $D^{\mathtt{AD}} f : \mathbb{R}^n \to \mathbb{R}^{1\times n}$ denotes a "derivative" of f used by AD, and $[M_1 \,/\, \cdots \,/\, M_n]$ denotes the matrix that stacks up matrices $M_1, \ldots, M_n$ vertically. Note that $[\![f\,(\mathtt{P}_1, \ldots, \mathtt{P}_n)]\!]^{\mathtt{AD}}$ captures the essence of AD: it computes derivatives based on the chain rule for differentiation.

Using the above definitions, we define $D^{\mathtt{AD}} z_L : \mathbb{R}^W \to \mathbb{R}^{N_L \times W}$ as what AD essentially computes when applied to a program that canonically represents a neural network $z_L : \mathbb{R}^W \to \mathbb{R}^{N_L}$:

$$D^{\mathtt{AD}} z_L(w) \triangleq \big[[\![\mathtt{P}_{z_{L,1}}]\!]^{\mathtt{AD}}(w) \,/\, \cdots \,/\, [\![\mathtt{P}_{z_{L,N_L}}]\!]^{\mathtt{AD}}(w)\big].$$

Note that $D^{\mathtt{AD}} z_L$ depends on the "derivative" of (pre-)activation functions (i.e., $D^{\mathtt{AD}}\sigma_{l,i}$ and $D^{\mathtt{AD}}\tau_{l,i}$) used by AD.

Lemma A.12. *For any $\gamma \in \Gamma$ and $w \in \mathcal{R}^\gamma$,*

$$D^{\mathtt{AD}} z_L(w) = D z_L^\gamma(w).$$

Proof. Let $\gamma \in \Gamma$. We prove the following claim: for all $l \in [L] \cup \{0\}$, $i \in [N_l]$, and $w \in \mathcal{R}^\gamma$,

$$D z_{l,i}^\gamma(w) = [\![\mathtt{P}_{z_{l,i}}]\!]^{\mathtt{AD}}(w).$$

Note that this claim implies the conclusion since

$$D z_L^\gamma(w) = [D z_{L,1}^\gamma(w) \,/\, \cdots \,/\, D z_{L,N_L}^\gamma(w)] = [[\![\mathtt{P}_{z_{L,1}}]\!]^{\mathtt{AD}}(w) \,/\, \cdots \,/\, [\![\mathtt{P}_{z_{L,N_L}}]\!]^{\mathtt{AD}}(w)] = D^{\mathtt{AD}} z_L(w).$$

We prove the claim by induction on l.

Case $l = 0$. Let $i \in [N_l]$ and $w \in \mathcal{R}^\gamma$. Since $\mathtt{P}_{z_{0,i}}$ is a constant program, $D z_{0,i}^\gamma(w) = \mathbb{0} = [\![\mathtt{P}_{z_{0,i}}]\!]^{\mathtt{AD}}(w)$ as desired.

Case $l > 0$. Let $i \in [N_l]$ and $w \in \mathcal{R}^\gamma$. Observe that

$$[\![\mathtt{P}_{y_{l,i}}]\!]^{\mathtt{AD}}(w) = [\![\tau_{l,i}\,(\mathtt{P}_{z_{l-1,1}}, \ldots, \mathtt{P}_{z_{l-1,N_{l-1}}}, \mathtt{w}_{l,1}, \ldots, \mathtt{w}_{l,W_l})]\!]^{\mathtt{AD}}(w)$$

$$
\begin{aligned}
&= D\tau_{l,i}\big([\![\mathtt{P}_{z_{l-1,1}}]\!](w), \ldots, [\![\mathtt{P}_{z_{l-1,N_{l-1}}}]\!](w), [\![\mathtt{w}_{l,1}]\!](w), \ldots, [\![\mathtt{w}_{l,N_l}]\!](w)\big) \\
&\qquad \cdot \big[[\![\mathtt{P}_{z_{l-1,1}}]\!]^{\mathtt{AD}}(w) \,/\, \cdots \,/\, [\![\mathtt{P}_{z_{l-1,N_{l-1}}}]\!]^{\mathtt{AD}}(w) \,/\, [\![\mathtt{w}_{l,1}]\!]^{\mathtt{AD}}(w) \,/\, \cdots \,/\, [\![\mathtt{w}_{l,W_l}]\!]^{\mathtt{AD}}(w)\big] \\
&= D\tau_{l,i}\big(z_{l-1}(w), \pi_l(w)\big) \cdot \big[Dz^{\gamma}_{l-1,1}(w) \,/\, \cdots \,/\, Dz^{\gamma}_{l-1,N_{l-1}}(w) \,/\, \mathbb{1}_{l,1} \,/\, \cdots \,/\, \mathbb{1}_{l,N_l}\big] \\
&= D\tau_{l,i}\big(z_{l-1}(w), \pi_l(w)\big) \cdot \big[Dz^{\gamma}_{l-1}(w) \,/\, D\pi_l(w)\big] \\
&= D\tau_{l,i}\big((z_{l-1}, \pi_l)(w)\big) \cdot D(z^{\gamma}_{l-1}, \pi_l)(w), \qquad (6)
\end{aligned}
$$

where $(f, g) : \mathbb{R}^n \to \mathbb{R}^{m_1+m_2}$ is defined as $(f, g)(x) \triangleq (f(x), g(x))$ for $f : \mathbb{R}^n \to \mathbb{R}^{m_1}$ and $g : \mathbb{R}^n \to \mathbb{R}^{m_2}$. Here the third line uses $[\![\mathtt{P}_{z_{l-1,i'}}]\!](w) = z_{l-1,i'}(w)$ and $[\![\mathtt{P}_{z_{l-1,i'}}]\!]^{\mathtt{AD}}(w) = Dz^{\gamma}_{l-1,i'}(w)$ for all $i' \in [N_{l-1}]$, where the latter holds by induction hypothesis on $l-1$.

Using the observation above, we obtain the claim:

$$
\begin{aligned}
[\![\mathtt{P}_{z_{l,i}}]\!]^{\mathtt{AD}}(w) &= [\![\sigma_{l,i}\,(\mathtt{P}_{y_{l,i}})\,]\!]^{\mathtt{AD}}(w) \\
&= D^{\mathtt{AD}}\sigma_{l,i}\big([\![\mathtt{P}_{y_{l,i}}]\!](w)\big) \cdot [\![\mathtt{P}_{y_{l,i}}]\!]^{\mathtt{AD}}(w) \\
&= D^{\mathtt{AD}}\sigma_{l,i}\big(y_{l,i}(w)\big) \cdot D\tau_{l,i}\big((z_{l-1}, \pi_l)(w)\big) \cdot D(z^{\gamma}_{l-1}, \pi_l)(w) \\
&= D\sigma^{\gamma}_{l,i}\big(y_{l,i}(w)\big) \cdot D\tau_{l,i}\big((z_{l-1}, \pi_l)(w)\big) \cdot D(z^{\gamma}_{l-1}, \pi_l)(w) \\
&= D\sigma^{\gamma}_{l,i}\big((\tau_{l,i} \circ (z^{\gamma}_{l-1}, \pi_l))(w)\big) \cdot D\tau_{l,i}\big((z^{\gamma}_{l-1}, \pi_l)(w)\big) \cdot D(z^{\gamma}_{l-1}, \pi_l)(w) \\
&= D(\sigma^{\gamma}_{l,i} \circ \tau_{l,i} \circ (z^{\gamma}_{l-1}, \pi_l))(w) \\
&= Dz^{\gamma}_{l,i}(w).
\end{aligned}
$$

Here the third line uses $[\![\mathtt{P}_{y_{l,i}}]\!](w) = y_{l,i}(w)$ and Eq. (6), and the fourth line uses $D^{\mathtt{AD}}\sigma_{l,i}(y_{l,i}(w)) = D\sigma^{\gamma(l,i)}_{l,i}(y_{l,i}(w))$, which holds because $y_{l,i}(w) \in \mathcal{I}^{\gamma(l,i)}_{l,i}$ (by $w \in \mathcal{R}^{\gamma}$) and $\{(\mathcal{I}^{k}_{l,i}, \sigma^{k}_{l,i})\}_{k \in [K_{l,i}]}$ defines $D^{\mathtt{AD}}\sigma_{l,i}$. The fifth line uses $y_{l,i}(w) = y^{\gamma}_{l,i}(w)$ and $z_{l-1}(w) = z^{\gamma}_{l-1}(w)$ (by Lemma A.11 with $w \in \mathcal{R}^{\gamma}$), and the sixth line uses the chain rule, which is applicable to $(\sigma^{\gamma}_{l,i} \circ \tau_{l,i} \circ (z^{\gamma}_{l-1}, \pi_l))$ because $\sigma^{\gamma}_{l,i}$, $\tau_{l,i}$, z^{γ}_{l-1}, and π_l are differentiable (as z^{γ}_{l-1} is differentiable by Lemma A.9). $\square$

B. Upper Bounds on $|\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L)|$

B.1. Lemmas (Basic)

Lemma B.1. *For any $A, B \subseteq \mathbb{R}^n$,*

$$pbd(A \cup B) \subseteq pbd(A) \cup pbd(B), \qquad pbd(A \cap B) \subseteq pbd(A) \cup pbd(B).$$

Proof. Let $A, B \subseteq \mathbb{R}^n$. Then, $int(A \cup B) \supseteq int(A) \cup int(B)$ and $int(A \cap B) = int(A) \cap int(B)$. Using these, we obtain:

$$\begin{aligned}
pbd(A \cup B) &= (A \cup B) \setminus int(A \cup B) \\
&= (A \setminus int(A \cup B)) \cup (B \setminus int(A \cup B)) \\
&\subseteq (A \setminus int(A)) \cup (B \setminus int(B)) \\
&= pbd(A) \cup pbd(B), \\
pbd(A \cap B) &= (A \cap B) \setminus int(A \cap B) \\
&= (A \cap B) \setminus (int(A) \cap int(B)) \\
&= ((A \cap B) \setminus int(A)) \cup ((A \cap B) \setminus int(B)) \\
&\subseteq (A \setminus int(A)) \cup (B \setminus int(B)) \\
&= pbd(A) \cup pbd(B).
\end{aligned}$$

□

Lemma B.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function defined as $f(x) = g(x_{-n}) + c \cdot x_n$ for any $g : \mathbb{R}^n \to \mathbb{R}$ and $c \in \mathbb{R} \setminus \{0\}$, where x_{-n} denotes $(x_1, \ldots, x_{n-1})$. Then,*

$$\left|\{x \in \mathbb{M}^n \mid f(x) = 0\}\right| \leq |\mathbb{M}|^{n-1}.$$

Proof. Using the definition of f and $c \neq 0$, we obtain the conclusion:

$$\begin{aligned}
\left|\{x \in \mathbb{M}^n \mid f(x) = 0\}\right| &= \left|\{(x_{-n}, x_n) \in \mathbb{M}^{n-1} \times \mathbb{M} \mid f(x_{-n}, x_n) = 0\}\right| \\
&= \textstyle\sum_{x_{-n} \in \mathbb{M}^{n-1}} \left|\{x_n \in \mathbb{M} \mid x_n = -g(x_{-n})/c\}\right| \\
&\leq \textstyle\sum_{x_{-n} \in \mathbb{M}^{n-1}} 1 = |\mathbb{M}|^{n-1}.
\end{aligned}$$

□

B.2. Lemmas (Technical: Part 1)

Definition B.3. For a neural network $z_L : \mathbb{R}^W \to \mathbb{R}^{N_L}$, define the incorrect set and the non-differentiable set of z_L *over* $\mathbb{R}^W$ (not over Ω) as:

$$\begin{aligned}
\mathsf{inc}_\mathbb{R}(z_L) &\triangleq \{w \in \mathbb{R}^W \mid Dz_L(w) \neq \bot,\ D^{\mathtt{AD}} z_L(w) \neq Dz_L(w)\}, \\
\mathsf{ndf}_\mathbb{R}(z_L) &\triangleq \{w \in \mathbb{R}^W \mid Dz_L(w) = \bot\}.
\end{aligned}$$

Lemma B.4. *We have*

$$\mathsf{ndf}_\mathbb{R}(z_L) \cup \mathsf{inc}_\mathbb{R}(z_L) \subseteq \bigcup_{\gamma \in \Gamma} pbd(\mathcal{R}^\gamma).$$

Proof. First, observe that for all $\gamma \in \Gamma$,

$$D^{\mathtt{AD}} z_L(w) = Dz_L^\gamma(w) = Dz_L(w) \qquad \text{for all } w \in int(\mathcal{R}^\gamma),$$

where the first equality is by Lemma A.12, and the second equality is obtained by applying the following fact to $(z_L^\gamma, z_L, int(\mathcal{R}^\gamma))$: for any $f, g : \mathbb{R}^n \to \mathbb{R}^m$ and open $U \subseteq \mathbb{R}^n$, if f is differentiable on U and $f = g$ on U, then g is differentiable on U and $Df = Dg$ on U. Note that the previous fact is applicable since $int(\mathcal{R}^\gamma)$ is open, z_L^γ is differentiable (by Lemma A.9), and $z_L^\gamma = z_L$ on $int(\mathcal{R}^\gamma)$ by Lemma A.11.

From the above equation, we have

$$\bigcup_{\gamma \in \Gamma} int(\mathcal{R}^\gamma) \subseteq \mathbb{R}^W \setminus \big(\mathsf{ndf}_\mathbb{R}(z_L) \cup \mathsf{inc}_\mathbb{R}(z_L)\big).$$

From this, we obtain the conclusion:

$$\begin{aligned}\mathsf{ndf}_{\mathbb{R}}(z_L) \cup \mathsf{inc}_{\mathbb{R}}(z_L) &\subseteq \mathbb{R}^W \setminus \bigcup_{\gamma\in\Gamma} int(\mathcal{R}^\gamma) \\ &= \Big(\bigcup_{\gamma\in\Gamma} \mathcal{R}^\gamma\Big) \setminus \Big(\bigcup_{\gamma\in\Gamma} int(\mathcal{R}^\gamma)\Big) = \bigcup_{\gamma\in\Gamma} \big(\mathcal{R}^\gamma \setminus int(\mathcal{R}^\gamma)\big) = \bigcup_{\gamma\in\Gamma} pbd(\mathcal{R}^\gamma),\end{aligned}$$

where the first equality is by Lemma A.8, and the last equality is by the definition of $pbd(-)$. □

Lemma B.5. *We have*

$$\bigcup_{\gamma\in\Gamma} pbd(\mathcal{R}^\gamma) \subseteq \bigcup_{(l,i)\in\mathsf{Idx}} \bigcup_{c\in\mathsf{ndf}(\sigma_{l,i})} pbd\big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) = c\}\big).$$

Proof. First, we have

$$\begin{aligned}\bigcup_{\gamma\in\Gamma} pbd(\mathcal{R}^\gamma) &= \bigcup_{\gamma\in\Gamma} pbd\Big(\bigcap_{(l,i)\in\mathsf{Idx}} \{w \in \mathbb{R}^W \mid y_{l,i}(w) \in \mathcal{I}_{l,i}^{\gamma(l,i)}\}\Big) \\ &\subseteq \bigcup_{\gamma\in\Gamma} \bigcup_{(l,i)\in\mathsf{Idx}} pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in \mathcal{I}_{l,i}^{\gamma(l,i)}\}\Big) \\ &= \bigcup_{(l,i)\in\mathsf{Idx}} \bigcup_{\gamma\in\Gamma} pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in \mathcal{I}_{l,i}^{\gamma(l,i)}\}\Big) \\ &= \bigcup_{(l,i)\in\mathsf{Idx}} \bigcup_{k\in[K_{l,i}]} pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in \mathcal{I}_{l,i}^{k}\}\Big),\end{aligned} \tag{7}$$

where the first line uses the definition of $\mathcal{R}^\gamma$, the second line uses Lemma B.1, and the last line uses that $\{\gamma(l,i) \mid \gamma \in \Gamma\} = [K_{l,i}]$ for all (l,i). Note that in the last two lines, we change the way we count the proper boundary of all subregions: from per subregion to per activation neuron.

Next, for any $(l,i) \in \mathsf{Idx}$ and $k \in [K_{l,i}]$, we have

$$\begin{aligned}&pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in \mathcal{I}_{l,i}^{k}\}\Big) \\ &= pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in pbd(\mathcal{I}_{l,i}^{k})\} \cup \{w \in \mathbb{R}^W \mid y_{l,i}(w) \in int(\mathcal{I}_{l,i}^{k})\}\Big) \\ &\subseteq pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in pbd(\mathcal{I}_{l,i}^{k})\}\Big) \cup pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in int(\mathcal{I}_{l,i}^{k})\}\Big) \\ &= pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in pbd(\mathcal{I}_{l,i}^{k})\}\Big),\end{aligned} \tag{8}$$

where the third line is by Lemma B.1 and the last line is by the following: $pbd(A) = \emptyset$ for any open $A \subseteq \mathbb{R}^n$; and $\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in int(\mathcal{I}_{l,i}^{k})\}$ is open, because $y_{l,i}$ is continuous (by Lemma A.9) and the inverse image of an open set by a continuous function is open.

Finally, combining the above results, we obtain the conclusion:

$$\begin{aligned}\bigcup_{\gamma\in\Gamma} pbd(\mathcal{R}^\gamma) &\subseteq \bigcup_{(l,i)\in\mathsf{Idx}} \bigcup_{k\in[K_{l,i}]} pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) \in pbd(\mathcal{I}_{l,i}^{k})\}\Big) \\ &\subseteq \bigcup_{(l,i)\in\mathsf{Idx}} \bigcup_{c\in\mathsf{ndf}(\sigma_{l,i})} pbd\Big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) = c\}\Big),\end{aligned}$$

where the first line uses Eqs. (7) and (8), and the second line uses $\bigcup_{k\in[K_{l,i}]} pbd(\mathcal{I}_{l,i}^{k}) = \mathsf{ndf}(\sigma_{l,i})$ (by Definition A.5) □

B.3. Theorem 3.3 (Main Lemmas)

Lemma B.6. *We have*

$$\mathsf{ndf}_{\Omega}(z_L) \cup \mathsf{inc}_{\Omega}(z_L) \subseteq \bigcup_{(l,i)\in\mathsf{Idx}} \bigcup_{c\in\mathsf{ndf}(\sigma_{l,i})} \{w \in \Omega \mid y_{l,i}(w) = c\}.$$

Proof. We obtain the conclusion by chaining Lemma B.4, Lemma B.5, and the following: $pbd(A) \subseteq A$ for any $A \subseteq \mathbb{R}^W$, and $\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L) = \big(\mathsf{ndf}_\mathbb{R}(z_L) \cup \mathsf{inc}_\mathbb{R}(z_L)\big) \cap \Omega$. □

Lemma B.7. *Let $(l, i) \in \mathsf{Idx}$ and $c \in \mathbb{R}$. Suppose that τ_l has bias parameters. Then, for $S = \{w \in \Omega \mid y_{l,i}(w) = c\}$,*

$$|S| \leq |\mathbb{M}|^{W-1}.$$

Proof. Suppose that τ_l has bias parameters and S is given as above. Then, by the definition of having bias parameters, $W_l \geq N_l$ and there is $\tau'_{l,i} : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l - N_l} \to \mathbb{R}$ for all $i \in [N_l]$ such that

$$\tau_{l,i}(x, (u, v)) = \tau'_{l,i}(x, u) + v_i \qquad \text{for all } (u, v) \in \mathbb{R}^{W_l - N_l} \times \mathbb{R}^{N_l}.$$

From this, we have

$$y_{l,i}(w) = \tau_{l,i}(z_{l-1}(w), w_l) = \tau'_{l,i}\big(z_{l-1}(w_{1,1}, \ldots, w_{l-1,W_{l-1}}, 0, \ldots, 0), (w_{l,1}, \ldots, w_{l,W_l - N_l})\big) + w_{l,W_l - N_l + i},$$

where we also use that z_{l-1} depends only on $w_1, \ldots, w_{l-1}$. Note that the function $f : \mathbb{R}^W \to \mathbb{R}$ defined by $f(w) \triangleq y_{l,i}(w) - c$ satisfies the preconditions of Lemma B.2 (after reordering the input variables of f) due to the term $w_{l,W_l - N_l + i}$. Using this, we obtain the desired result:

$$|S| = |\{w \in \Omega \mid f(w) = 0\}| \leq |\mathbb{M}|^{W-1},$$

where the inequality is by Lemma B.2 applied to f. □

B.4. Theorem 3.3 (Main Proof)

Theorem B.8. *If z_L has bias parameters, then*

$$\frac{|\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L)|}{|\Omega|} \leq \frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} |\mathsf{ndf}(\sigma_{l,i})|.$$

Proof. Observe that

$$\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L) \subseteq \bigcup_{(l,i) \in \mathsf{Idx}} \bigcup_{c \in A_{l,i}} B_{l,i}(c), \qquad |B_{l,i}(c)| \leq |\mathbb{M}|^{W-1}, \tag{9}$$

where $A_{l,i} \triangleq \mathsf{ndf}(\sigma_{l,i})$ and $B_{l,i}(c) \triangleq \{w \in \Omega \mid y_{l,i}(w) = c\}$. Here the first equation is by Lemma B.6, and the second equation is by Lemma B.7 (which is applicable since τ_l has bias parameters by assumption). Combining the above observations, we obtain the conclusion:

$$\frac{|\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L)|}{|\Omega|} \leq \sum_{(l,i) \in \mathsf{Idx}} \sum_{c \in A_{l,i}} \frac{|B_{l,i}(c)|}{|\Omega|} \leq \sum_{(l,i) \in \mathsf{Idx}} |\mathsf{ndf}(\sigma_{l,i})| \cdot \frac{|\mathbb{M}|^{W-1}}{|\mathbb{M}|^W},$$

where the two inequalities use Eq. (9). □

Remark B.9. Theorem 3.3 is a direct corollary of Theorem B.8 and Theorem 3.2 (which we prove in §C). □

B.5. Lemmas (Technical: Part 2)

Lemma B.10. *Let $l \in [L]$. Suppose that $\tau_l : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l} \to \mathbb{R}^{N_l}$ is well-structured biaffine. Then, for every $i \in [N_l]$, there is a partial map $\phi_{l,i} : [W_l] \rightharpoonup [N_{l-1}]$ and associated matrix $M \in \mathbb{R}^{N_{l-1} \times W_l}$ and constant $d \in \mathbb{R}$ such that*

$$y_{l,i}(w) = d + \sum_{j \in dom(\phi_{l,i})} z_{l-1,\phi_{l,i}(j)}(w) \cdot M_{\phi_{l,i}(j),j} \cdot w_{l,j}$$

and $M_{\phi_{l,i}(j),j} \neq 0$ for all $j \in dom(\phi_{l,i})$.

Proof. Let $l \in [L]$, $\tau_l : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l} \to \mathbb{R}^{N_l}$ be a well-structured biaffine function, and $i \in [N_l]$. Then, there is a matrix $M \in \mathbb{R}^{N_{l-1} \times W_l}$ and a constant $d \in \mathbb{R}$ such that $\tau_{l,i}(x, u) = x^\mathsf{T} M u + d$ for all (x, u) and each column of M has at most one non-zero entry. Define a partial map $\phi_{l,i} : [W_l] \rightharpoonup [N_{l-1}]$ as:

$$\phi_{l,i}(j) \triangleq \begin{cases} i' & \text{if } M_{i',j} \neq 0 \text{ for some } i' \in [N_{l-1}] \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Here $\phi_{l,i}$ is well-defined because $M_{-,j}$ contains at most one non-zero entry for all $j \in [W_l]$. We claim that $\phi_{l,i}$, M, and d satisfy the conditions in this lemma. First, by the definition of $\phi_{l,i}$, $M_{\phi_{l,i}(j),j} \neq 0$ for all $j \in dom(\phi_{l,i})$. Also, we have the desired equation as follows:

$$\begin{aligned} y_{l,i}(w) &= \tau_{l,i}(z_{l-1}(w), w_l) \\ &= d + (z_{l-1}(w)^\mathsf{T} M) \cdot w_l \\ &= d + (v_1, \ldots, v_{W_{l-1}})^\mathsf{T} \cdot w_l \\ &= d + \textstyle\sum_{j \in [W_{l-1}]} v_j \cdot w_{l,j} \\ &= d + \textstyle\sum_{j \in dom(\phi_{l,i})} z_{l-1,\phi_{l,i}(j)}(w) \cdot M_{\phi_{l,i}(j),j} \cdot w_{l,j}, \end{aligned}$$

where $v_j \in \mathbb{R}$ is defined as $v_j \triangleq z_{l-1,\phi_{l,i}(j)}(w) \cdot M_{\phi_{l,i}(j),j}$ if $j \in dom(\phi_{l,i})$, and $v_j \triangleq 0$ otherwise. Here the second line uses the definition of M and d, and the third and last lines use the definition of v_j. This concludes the proof. □

Lemma B.11. *For every $(l, i) \in \mathsf{Idx}$ and $c \in \mathbb{R}$, let $A_{l,i} \subseteq \mathbb{R}$ be any set and $B_{l,i}(c) \subseteq \mathbb{R}^W$ be the set $\{w \in \mathbb{R}^W \mid y_{l,i}(w) = c\}$. Suppose that for every $l \in [L]$, one of the following holds:*

(a) τ_l has bias parameters, or
(b) τ_l is well-structured biaffine.

In the case of (b), let $\phi_{l,i}$ be the partial map described in Lemma B.10 for all $i \in [N_l]$. Then,

$$\bigcup_{(l,i) \in \mathsf{Idx}} \bigcup_{c \in A_{l,i}} pbd(B_{l,i}(c)) \subseteq \bigcup_{(l,i) \in \mathsf{Idx}} \bigcup_{c' \in A'_{l,i}} B'_{l,i}(c'),$$

where $A'_{l,i} \subseteq \mathbb{R}$ and $B'_{l,i}(c') \subseteq \mathbb{R}^W$ are defined as

$$A'_{l,i} \triangleq \begin{cases} A_{l,i} & \textit{if } \tau_{l+1} \textit{ satisfies the condition (a) or } l = L \\ A_{l,i} \cup \mathsf{bdz}(\sigma_{l,i}) & \textit{if } \tau_{l+1} \textit{ satisfies the condition (b),} \end{cases}$$

$$B'_{l,i}(c') \triangleq \begin{cases} B_{l,i}(c') & \textit{if } \tau_l \textit{ satisfies the condition (a)} \\ B_{l,i}(c') \cap \bigcup_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) \neq 0\} & \textit{if } \tau_l \textit{ satisfies the condition (b).} \end{cases}$$

Proof. We claim that the following holds: for all $l \in [L]$, $i \in [N_l]$, and $c \in A'_{l,i}$,

$$pbd(B_{l,i}(c)) \subseteq \bigcup_{(l',i') \in \mathsf{Idx}} \bigcup_{c' \in A'_{l',i'}} B'_{l',i'}(c'). \tag{10}$$

This claim implies the conclusion because $A_{l,i} \subseteq A'_{l,i}$ for all $(l, i) \in \mathsf{Idx}$ (by the definition of $A'_{l,i}$). We prove the claim by induction on l.

Case $l = 1$**.** Let $i \in [N_l]$ and $c \in A'_{l,i}$. We prove Eq. (10) by case analysis on τ_l.

Subcase 1: τ_l satisfies the condition (a). In this subcase, Eq. (10) holds since

$$pbd(B_{l,i}(c)) \subseteq B_{l,i}(c) = B'_{l,i}(c), \qquad c \in A'_{l,i},$$

where the equality uses the definition of $B'_{l,i}$.

Subcase 2: τ_l satisfies the condition (b). In this subcase, we have

$$pbd(B_{l,i}(c)) = pbd\Big(\Big(B_{l,i}(c) \cap \bigcup_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) \neq 0\}\Big)$$

$$\cup \Big(B_{l,i}(c) \cap \bigcap_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) = 0\}\Big)$$
$$\subseteq pbd\Big(B_{l,i}(c) \cap \bigcup_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) \neq 0\}\Big)$$
$$\cup\, pbd\Big(B_{l,i}(c) \cap \bigcap_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) = 0\}\Big),$$

where the inclusion uses Lemma B.1. To prove Eq. (10), it suffices to show that the two terms in the last two lines are contained in the RHS of Eq. (10). The first term does so because

$$pbd\Big(B_{l,i}(c) \cap \bigcup_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) \neq 0\}\Big) = pbd(B'_{l,i}(c)) \subseteq B'_{l,i}(c), \qquad c \in A'_{l,i},$$

where the equality is by the definition of $B'_{l,i}$ and that τ_l does not have bias parameters. The second term is also contained in the RHS of Eq. (10) as follows. Let $S \triangleq \bigcap_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) = 0\}$, and $M \in \mathbb{R}^{N_{l-1} \times W_l}$ and $d \in \mathbb{R}$ be a matrix and a constant associated with $\phi_{l,i}$ that are described in Lemma B.10. Then,

$$B_{l,i}(c) \cap S = \begin{cases} S & \text{if } c = d \\ \emptyset & \text{if } c \neq d, \end{cases}$$

because $w \in S$ implies $y_{l,i}(w) = d + \sum_{j \in dom(\phi_{l,i})} z_{l-1,\phi_{l,i}(j)}(w) \cdot M_{\phi_{l,i}(j),j} \cdot w_{l,j} = d$ by Lemma B.10 (which is applicable since τ_l is well-structured biaffine by assumption). From this, we have

$$pbd\Big(B_{l,i}(c) \cap \bigcap_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) = 0\}\Big) = pbd(B_{l,i}(c) \cap S) \subseteq pbd(S) \cup pbd(\emptyset).$$

Hence, it suffices to show that $pbd(S)$ is contained in the RHS of Eq. (10) (since $pbd(\emptyset) = \emptyset$). Using $l = 1$, we obtain this:

$$pbd(S) \subseteq pbd(\mathbb{R}^W) \cup pbd(\emptyset) = \emptyset,$$

where the inclusion follows from $S \in \{\mathbb{R}^W, \emptyset\}$ which holds because $z_{l-1,\phi_{l,i}(j)}$ is a constant function for all $j \in [N_{l-1}]$ (by $l = 1$ and the assumption on z_0).

Case $l > 1$**.** Let $i \in [N_l]$ and $c \in A'_{l,i}$. We prove Eq. (10) in the exact same way as we did for the case $l = 1$. Note that the above proof for the previous case ($l = 1$) applies directly to the current case ($l > 1$), except for the following subclaim: if τ_l does not have bias parameters, then $pbd(S)$ is contained in the RHS of Eq. (10). This subclaim holds also for $l > 1$, as follows:

$$\begin{aligned}
pbd(S) &= pbd\Big(\bigcap_{j \in dom(\phi_{l,i})} \{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) = 0\}\Big) \\
&\subseteq \bigcup_{j \in dom(\phi_{l,i})} pbd\Big(\{w \in \mathbb{R}^W \mid z_{l-1,\phi_{l,i}(j)}(w) = 0\}\Big) \\
&= \bigcup_{j \in dom(\phi_{l,i})} pbd\Big(\big\{w \in \mathbb{R}^W \mid y_{l-1,\phi_{l,i}(j)}(w) \in pbd\big(\sigma^{-1}_{l-1,\phi_{l,i}(j)}(0)\big)\big\} \\
&\qquad\qquad \cup \big\{w \in \mathbb{R}^W \mid y_{l-1,\phi_{l,i}(j)}(w) \in int\big(\sigma^{-1}_{l-1,\phi_{l,i}(j)}(0)\big)\big\}\Big) \\
&\subseteq \bigcup_{j \in dom(\phi_{l,i})} pbd\Big(\big\{w \in \mathbb{R}^W \mid y_{l-1,\phi_{l,i}(j)}(w) \in pbd\big(\sigma^{-1}_{l-1,\phi_{l,i}(j)}(0)\big)\big\}\Big) \\
&\qquad\qquad \cup\, pbd\Big(\big\{w \in \mathbb{R}^W \mid y_{l-1,\phi_{l,i}(j)}(w) \in int\big(\sigma^{-1}_{l-1,\phi_{l,i}(j)}(0)\big)\big\}\Big) \\
&= \bigcup_{j \in dom(\phi_{l,i})} pbd\Big(\big\{w \in \mathbb{R}^W \mid y_{l-1,\phi_{l,i}(j)}(w) \in pbd\big(\sigma^{-1}_{l-1,\phi_{l,i}(j)}(0)\big)\big\}\Big) \\
&= \bigcup_{j \in dom(\phi_{l,i})} \bigcup_{b \in \mathsf{bdz}(\sigma_{l-1,\phi_{l,i}(j)})} pbd(B_{l-1,\phi_{l,i}(j)}(b)),
\end{aligned}$$

$$pbd(B_{l-1,\phi_{l,i}(j)}(b)) \subseteq \bigcup_{(l',i')\in \mathsf{Idx}} \bigcup_{c'\in A'_{l',i'}} B'_{l',i'}(c') \qquad \text{for all } j \in dom(\phi_{l,i}) \text{ and } b \in \mathsf{bdz}(\sigma_{l-1,\phi_{l,i}(j)}).$$

Here the first and second inclusions use Lemma B.1, and the second last equality uses that $y_{l-1,\phi_{l,i}(j)}$ is continuous (by Lemma A.9). The last equality uses $pbd(\sigma^{-1}_{l-1,\phi_{l,i}(j)}(0)) = \mathsf{bdz}(\sigma_{l-1,\phi_{l,i}(j)})$ (which holds since $\sigma_{l-1,\phi_{l,i}(j)}$ is continuous and the preimage of a closed set by a continuous map is closed), and the definition of $B_{l-1,\phi_{l,i}(j)}$. The last inclusion is by the induction hypothesis applied to $(l-1, j, b)$ for $j \in dom(\phi_{l,i})$ and $b \in \mathsf{bdz}(\sigma_{l-1,\phi_{l,i}(j)})$, together with $dom(\phi_{l,i}) \subseteq [N_{l-1}]$ and $\mathsf{bdz}(\sigma_{l-1,\phi_{l,i}(j)}) \subseteq A'_{l-1,\phi_{l,i}(j)}$ (which holds by the definition of $A'_{l-1,\phi_{l,i}(j)}$ with $l-1 \neq L$ and that τ_l does not have bias parameters). Hence, Eq. (10) holds for $l > 1$, and this concludes the proof. □

B.6. Theorem 4.2 (Main Lemmas)

Lemma B.12. *For every $l \in [L]$, suppose that τ_l satisfies either the condition (a) or (b) in Lemma B.11. Then,*

$$\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L) \subseteq \bigcup_{(l,i)\in \mathsf{Idx}} \bigcup_{c\in A_{l,i}} B_{l,i}(c),$$

where $A_{l,i} \subseteq \mathbb{R}$ and $B_{l,i}(c) \subseteq \Omega$ are defined as

$$A_{l,i} \triangleq \begin{cases} \mathsf{ndf}(\sigma_{l,i}) & \text{if } \tau_{l+1} \text{ satisfies the condition (a) or } l = L \\ \mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i}) & \text{if } \tau_{l+1} \text{ satisfies the condition (b),} \end{cases}$$

$$B_{l,i}(c) \triangleq \begin{cases} \{w \in \Omega \mid y_{l,i}(w) = c\} & \text{if } \tau_l \text{ satisfies the condition (a)} \\ \{w \in \Omega \mid y_{l,i}(w) = c \wedge \bigvee_{j\in dom(\phi_{l,i})} z_{l-1,\phi_{l,i}(j)}(w) \neq 0\} & \text{if } \tau_l \text{ satisfies the condition (b).} \end{cases}$$

Proof. We obtain the conclusion by chaining Lemma B.4, Lemma B.5, Lemma B.11 (which is applicable by assumption), and $\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L) = \big(\mathsf{ndf}_\mathbb{R}(z_L) \cup \mathsf{inc}_\mathbb{R}(z_L)\big) \cap \Omega$. □

Lemma B.13. *Let $(l, i) \in \mathsf{Idx}$ and $c \in \mathbb{R}$. Suppose that τ_l is well-structured biaffine. Consider $S = \{w \in \Omega \mid y_{l,i}(w) = c \wedge \bigvee_{j\in dom(\phi_{l,i})} z_{l-1,\phi_{l,i}(j)}(w) \neq 0\}$, where $\phi_{l,i}$ denotes the partial map described in Lemma B.10. Then,*

$$|S| \leq |\mathbb{M}|^{W-1}.$$

Proof. Suppose that τ_l is well-structured biaffine, and S is given as above. We make three observations. First,

$$\begin{aligned} S &= \big\{(u, v) \in \mathbb{M}^{W'} \times \mathbb{M}^{W-W'} \;\big|\; \big(\exists j \in dom(\phi_{l,i}).\, z_{l-1,\phi_{l,i}(j)}(u, 0, \ldots, 0) \neq 0\big) \wedge y_{l,i}(u, v) = c\big\} \\ &= \bigcup_{u\in U} \bigcup_{v\in \mathbb{M}^{W-W'}} \{(u, v) \mid y_{l,i}(u, v) = c\}, \end{aligned} \tag{11}$$

where the first line uses $W' \triangleq W_1 + \cdots + W_{l-1}$ and that z_{l-1} depends only on $w_1, \ldots, w_{l-1}$, and the second line uses $U \triangleq \{u \in \mathbb{M}^{W'} \mid \exists j \in dom(\phi_{l,i}).\, z_{l-1,\phi_{l,i}(j)}(u, 0, \ldots, 0) \neq 0\}$. Second, by Lemma B.10 (which is applicable since τ_l is well-structured biaffine by assumption), there are $M \in \mathbb{R}^{N_{l-1}\times W_l}$ and $d \in \mathbb{R}$ such that $M_{\phi_{l,i}(j),j} \neq 0$ for all $j \in \phi_{l,i}$, and

$$\begin{aligned} y_{l,i}(u, v) &= d + \sum_{j\in dom(\phi_{l,i})} z_{l-1,\phi_{l,i}(j)}(u, v) \cdot M_{\phi_{l,i}(j),j} \cdot v_j \\ &= d + \sum_{j\in dom(\phi_{l,i})} z_{l-1,\phi_{l,i}(j)}(u, 0, \ldots, 0) \cdot M_{\phi_{l,i}(j),j} \cdot v_j \end{aligned} \tag{12}$$

for all $(u, v) \in \mathbb{R}^{W'} \times \mathbb{R}^{W-W'}$, where the second equality uses that z_{l-1} depends only on u. Third, for any $u \in U$, the function $f_u : \mathbb{R}^{W-W'} \to \mathbb{R}$ defined by $f_u(v) \triangleq y_{l,i}(u, v) - c$ satisfies the preconditions of Lemma B.2 (after reordering the input variables of f_u) due to the following: $z_{l-1,\phi_{l,i}(j)}(u, 0, \ldots, 0) \neq 0$ for some $j \in dom(\phi_{l,i})$ since $u \in U$; and the coefficient of v_j in $f_u(v)$ is $z_{l-1,\phi_{l,i}(j)}(u, 0, \ldots, 0) \cdot M_{\phi_{l,i}(j),j} \neq 0$ by Eq. (12) and $M_{\phi_{l,i}(j),j} \neq 0$.

By combining the above observations, we obtain the conclusion:

$$|S| = \Big| \bigcup_{u\in U} \bigcup_{v\in \mathbb{M}^{W-W'}} \{(u, v) \mid y_{l,i}(u, v) = c\} \Big|$$

$$
\begin{aligned}
&= \sum_{u \in U} \Big| \bigcup_{v \in \mathbb{M}^{W-W'}} \{(u,v) \mid y_{l,i}(u,v) = c\} \Big| \\
&= \sum_{u \in U} \big| \{v \in \mathbb{M}^{W-W'} \mid f_u(v) = 0\} \big| \\
&\leq |\mathbb{M}|^{W'} \cdot |\mathbb{M}|^{W-W'-1} = |\mathbb{M}|^{W-1},
\end{aligned}
$$

where the first line uses Eq. (11), the third line uses the definition of f_u, and the last line uses Lemma B.2 applied to f_u. □

B.7. Theorem 4.2 (Main Proof)

Theorem 4.2. *If τ_l either has bias parameters or is well-structured biaffine for all $l \in [L]$, then*

$$
\frac{|\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L)|}{|\Omega|} \leq \frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} \Big| \mathsf{ndf}(\sigma_{l,i}) \cup \big(\mathsf{bdz}(\sigma_{l,i}) \cap S_{l+1}\big) \Big|,
$$

where $S_l \subseteq \mathbb{R}$ is defined by

$$
S_l \triangleq \begin{cases} \emptyset & \text{if } l > L \text{ or } \tau_l \text{ has bias parameters} \\ \mathbb{R} & \text{otherwise.} \end{cases}
$$

Proof. Observe that

$$
\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L) \subseteq \bigcup_{(l,i) \in \mathsf{Idx}} \bigcup_{c \in A_{l,i}} B_{l,i}(c), \qquad |B_{l,i}(c)| \leq |\mathbb{M}|^{W-1}, \tag{13}
$$

where $A_{l,i} \subseteq \mathbb{R}$ and $B_{l,i}(c) \subseteq \Omega$ are defined as in Lemma B.12. Here the first equation is by Lemma B.12 and the second equation is by Lemmas B.7 and B.13, where these lemmas are applicable by the definition of $B_{l,i}(c)$ and because τ_l either has bias parameters or is well-structured biaffine (both by assumption). Observe further that

$$
A_{l,i} = \mathsf{ndf}(\sigma_{l,i}) \cup (\mathsf{bdz}(\sigma_{l,i}) \cap S_{l+1}) \tag{14}
$$

by the definition of $A_{l,i}$ and S_l, where S_l is defined in the statement of this theorem. Combining the above observations, we obtain the conclusion:

$$
\frac{|\mathsf{ndf}_\Omega(z_L) \cup \mathsf{inc}_\Omega(z_L)|}{|\Omega|} \leq \sum_{(l,i) \in \mathsf{Idx}} \sum_{c \in A_{l,i}} \frac{|B_{l,i}(c)|}{|\Omega|} \leq \sum_{(l,i) \in \mathsf{Idx}} \big| \mathsf{ndf}(\sigma_{l,i}) \cap (\mathsf{bdz}(\sigma_{l,i}) \cap S_{l+1}) \big| \cdot \frac{|\mathbb{M}|^{W-1}}{|\mathbb{M}|^W},
$$

where the first inequality is by Eq. (13) and the second inequality is by Eqs. (13) and (14). □

C. Upper Bounds on $|\mathsf{inc}_\Omega(z_L)|$

In the rest of the appendix, we use the following notation. For a vector $v \in \mathbb{R}^n$, $v_{a:b}$ denotes the vector $(v_a, \dots, v_b)$. For a matrix $M \in \mathbb{R}^{n\times m}$, $M_{a:b,\,c:d}$ denotes the matrix $(M_{i,j})_{a\le i\le b,\,c\le j\le d}$; $M_{a:b,\,c}$ denotes the vector $(M_{a,c}, \dots, M_{b,c})$; and $M_{*,\,c:d}$ denotes $M_{1:n,\,c:d}$ (and similarly for $M_{a:b,\,*}$ and $M_{*,\,c}$).

C.1. Lemmas (Basic)

Lemma C.1. *Let $n \in \mathbb{N}$. For each $j \in [n]$, let $f_j : \mathbb{R} \to \mathbb{R}$ and $\mathcal{A}_j$ be a finite cover of $\mathbb{R}$ (i.e., $\bigcup_{A\in\mathcal{A}_j} A = \mathbb{R}$ and $|\mathcal{A}_j| < \infty$). Consider $x \in \mathbb{R}$. Then, there is $\{x_i\}_{i\in\mathbb{N}} \subseteq (x, \infty)$ such that $\lim_{i\to\infty} x_i = x$ and for all $j \in [n]$,*

$$\{f_j(x_i) \mid i \in \mathbb{N}\} \subseteq A \qquad \text{for some } A \in \mathcal{A}_j.$$

Further, there is $\{x'_i\}_{i\in\mathbb{N}} \subseteq (-\infty, x)$ that satisfies the same conditions stated above.

Proof. Consider f_j, $\mathcal{A}_j$, and x stated above ($j \in [n]$). Let $x_i \triangleq x + 1/i$ for $i \in \mathbb{N}$. Then,

$$\{x_i\}_{i\in\mathbb{N}} \subseteq (x, \infty), \qquad \lim_{i\to\infty} x_i = x. \tag{15}$$

For each $(i, j) \in \mathbb{N} \times [n]$, let $A_{i,j} \in \mathcal{A}_j$ be the set satisfying $f_j(x_i) \in A_{i,j}$, and $A_i \triangleq (A_{i,1}, \dots, A_{i,n}) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$, where $A_{i,j}$ always exists since $\mathcal{A}_j$ is a cover of $\mathbb{R}$. Observe that since $|\mathcal{A}_1 \times \cdots \times \mathcal{A}_n| < \infty$ (by $|\mathcal{A}_j| < \infty$ and $n < \infty$) and $|\mathbb{N}| = \infty$, there must exist $\{k_i\}_{i\in\mathbb{N}} \subseteq \mathbb{N}$ such that

$$k_1 < k_2 < \cdots, \qquad A_{k_1} = A_{k_2} = \cdots. \tag{16}$$

We claim that $\{x_{k_i}\}_{i\in\mathbb{N}}$ satisfies the desired conditions. First, by Eq. (15) and $\lim_{i\to\infty} k_i = \infty$ (due to Eq. (16)), $\{x_{k_i}\}_{i\in\mathbb{N}} \subseteq (x, \infty)$ and $\lim_{i\to\infty} x_{k_i} = x$. Second, by Eq. (16), $\{f_j(x_{k_i}) \mid i \in \mathbb{N}\} \subseteq A_{k_1,j}$ for all $j \in [n]$. Hence, the claim holds and this concludes the proof. □

Lemma C.2. *Let $f, g : \mathbb{R} \to \mathbb{R}$ and $x \in \mathbb{R}$. Suppose that f and g are differentiable at x, and there is $\{x_i\}_{i\in\mathbb{N}} \subseteq \mathbb{R} \setminus \{x\}$ such that $\lim_{i\to\infty} x_i = x$ and $f(x_i) = g(x_i)$ for all $i \in \mathbb{N}$. Then,*

$$Df(x) = Dg(x).$$

Proof. Consider $f, g : \mathbb{R} \to \mathbb{R}$, $x \in \mathbb{R}$, and $\{x_i\}_{i\in\mathbb{N}} \subseteq \mathbb{R} \setminus \{x\}$ stated above. Then,

$$f(x) = \lim_{i\to\infty} f(x_i) = \lim_{i\to\infty} g(x_i) = g(x),$$

where the first and third equalities are by that f and g are continuous at x (as they are differentiable at x) and $x_i \to x$, and the second equality by that $f(x_i) = g(x_i)$ for all $i \in \mathbb{N}$. Using this, we obtain

$$Df(x) = \lim_{i\to\infty} \frac{f(x_i) - f(x)}{x_i - x} = \lim_{i\to\infty} \frac{g(x_i) - g(x)}{x_i - x} = Dg(x),$$

where the first and third equalities are by that f and g are differentiable at x, $x_i \to x$, and $x_i \neq x$ for all $i \in \mathbb{N}$, and the second equality by that $f(x_i) = g(x_i)$ for all $i \in \mathbb{N}$. This completes the proof. □

C.2. Lemmas (Technical: Part 1)

Definition C.3. Let $\gamma \in \Gamma$. Define $\mathcal{R}^\gamma_{\mathsf{cl}} \subseteq \mathbb{R}^W$ as

$$\mathcal{R}^\gamma_{\mathsf{cl}} \triangleq \bigcap_{(l,i)\in\mathsf{Idx}} \{w \in \mathbb{R}^W \mid y_{l,i}(w) \in cl(\mathcal{I}^{\gamma(l,i)}_{l,i})\}.$$

Note that when defining $\mathcal{R}^\gamma$ in Definition A.7, we used $\mathcal{I}^{\gamma(l,i)}_{l,i}$ instead of $cl(\mathcal{I}^{\gamma(l,i)}_{l,i})$.

Definition C.4. For $\gamma \in \Gamma$ and $l \in [L]$, define

$$\widetilde{\tau}_l : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l + W_{l+1} + \cdots + W_L} \to \mathbb{R}^{N_l} \times \mathbb{R}^{W_{l+1} + \cdots + W_L},$$

$$\widetilde{\sigma}_l, \widetilde{\sigma}_l^\gamma : \mathbb{R}^{N_l} \times \mathbb{R}^{W_{l+1}+\cdots+W_L} \to \mathbb{R}^{N_l} \times \mathbb{R}^{W_{l+1}+\cdots+W_L},$$
$$\widetilde{z}_l, \widetilde{z}_l^\gamma : \mathbb{R}^{N_{l-1}} \times \mathbb{R}^{W_l+W_{l+1}+\cdots+W_L} \to \mathbb{R}^{N_L}$$

as follows:

$$\begin{aligned}
&\widetilde{\tau}_l(x,u) \triangleq \big(\tau_l(x,u_1,\ldots,u_{W_l}), u_{W_l+1},\ldots,u_{W_l+W_{l+1}+\cdots+W_L}\big), \\
&\widetilde{\sigma}_l(x,u) \triangleq (\sigma_l(x),u), && \widetilde{\sigma}_l^\gamma(x,u) \triangleq (\sigma_l^\gamma(x),u), \\
&\widetilde{z}_l(x,u) \triangleq (\widetilde{z}_{l+1} \circ \widetilde{\sigma}_l \circ \widetilde{\tau}_l)(x,u) && \widetilde{z}_l^\gamma(x,u) \triangleq (\widetilde{z}_{l+1}^\gamma \circ \widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l)(x,u)
\end{aligned}$$

where $\widetilde{z}_{L+1}, \widetilde{z}_{L+1}^\gamma : \mathbb{R}^{N_L} \to \mathbb{R}^{N_L}$ are defined as the identity function.

Lemma C.5. *For all $l \in [L]$ and $\gamma \in \Gamma$, $\widetilde{z}_l$ is continuous and $\widetilde{z}_l^\gamma$ is differentiable.*

Proof. Since the proof is similar to that of Lemma A.9, we omit it. □

Lemma C.6. *Let $\gamma \in \Gamma$, $w = (w_1,\ldots,w_L) \in \mathcal{R}_{\mathsf{cl}}^\gamma$, $l \in [L]$, and $x = (z_{l-1}(w), w_l, \ldots, w_L)$. Then,*

$$\widetilde{\tau}_l(x) = \big(y_l(w), w_{l+1},\ldots,w_L\big), \qquad (\widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l)(x) = \big(z_l(w), w_{l+1},\ldots,w_L\big).$$

Proof. By the definition of $\widetilde{\tau}_l$ and $\widetilde{\sigma}_l^\gamma$, we get the conclusion:

$$\begin{aligned}
\widetilde{\tau}_l(x) &= \big(\tau_l(z_{l-1}(w), w_l), w_{l+1},\ldots,w_L\big) = \big(y_l(w), w_{l+1},\ldots,w_L\big), \\
(\widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l)(x) &= \big(\sigma_l^\gamma(y_l(w)), w_{l+1},\ldots,w_L\big) = \big(z_l(w), w_{l+1},\ldots,w_L\big),
\end{aligned}$$

where the last equality is by the observation that $\sigma_{l,i}^{\gamma(l,i)}(y_{l,i}(w)) = \sigma_{l,i}(y_{l,i}(w))$ for all $i \in [N_l]$. Here the observation holds because $\sigma_{l,i}^{\gamma(l,i)}$ and $\sigma_{l,i}$ coincide on $cl(\mathcal{I}_{l,i}^{\gamma(l,i)})$ (as they coincide on $\mathcal{I}_{l,i}^{\gamma(l,i)}$ and are both continuous) and $y_{l,i}(w) \in cl(\mathcal{I}_{l,i}^{\gamma(l,i)})$ (by $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$). □

Lemma C.7. *Let $\gamma \in \Gamma$ and $l \in [L]$. Then, for all $w = (w_1,\ldots,w_L) \in \mathbb{R}^W$,*

$$\widetilde{z}_l\big(z_{l-1}(w), w_l,\ldots,w_L\big) = z_L(w), \qquad \widetilde{z}_l^\gamma\big(z_{l-1}^\gamma(w), w_l,\ldots,w_L\big) = z_L^\gamma(w).$$

Proof. Let $\gamma \in \Gamma$. The proof is by induction on $l \in [L]$ (starting from $l = L+1$).

Case $l = L+1$. Since $\widetilde{z}_{L+1}$ and $\widetilde{z}_{L+1}^\gamma$ are identity functions, the desired equations clearly hold.

Case $l < L+1$. We obtain the first desired equation as follows:

$$\begin{aligned}
\widetilde{z}_l\big(z_{l-1}(w), w_l,\ldots,w_L\big) &= (\widetilde{z}_{l+1} \circ \widetilde{\sigma}_l \circ \widetilde{\tau}_l)\big(z_{l-1}(w), w_l,\ldots,w_L\big) \\
&= (\widetilde{z}_{l+1} \circ \widetilde{\sigma}_l)\big(\tau_l(z_{l-1}(w), w_l), w_{l+1},\ldots,w_L\big) \\
&= (\widetilde{z}_{l+1} \circ \widetilde{\sigma}_l)\big(y_l(w), w_{l+1},\ldots,w_L\big) \\
&= \widetilde{z}_{l+1}\big(\sigma_l(y_l(w)), w_{l+1},\ldots,w_L\big) \\
&= \widetilde{z}_{l+1}\big(z_l(w), w_{l+1},\ldots,w_L\big) \\
&= z_L(w),
\end{aligned}$$

where all but last lines use the definition of $\widetilde{z}_l$, $\widetilde{\tau}_l$, $\widetilde{\sigma}_l$, y_l, and z_l, and the last line uses induction hypothesis on $l+1$. We can obtain the second desired equation similarly, by using induction hypothesis on $l+1$ and the definition of $\widetilde{z}_l^\gamma$, $\widetilde{\tau}_l$, $\widetilde{\sigma}_l^\gamma$, y_l^γ, and z_l^γ. □

Lemma C.8. *Let $\gamma \in \Gamma$ and $l \in [L]$. Then, for all $w = (w_1,\ldots,w_L) \in \mathcal{R}^\gamma$,*

$$\widetilde{z}_l\big(z_{l-1}(w), w_l,\ldots,w_L\big) = \widetilde{z}_l^\gamma\big(z_{l-1}(w), w_l,\ldots,w_L\big).$$

Proof. By Lemma C.7, we have the conclusion as follows:

$$\widetilde{z}_l(z_{l-1}(w), w_l,\ldots,w_L) = z_L(w) = z_L^\gamma(w) = \widetilde{z}_l^\gamma(z_{l-1}(w), w_l,\ldots,w_L),$$

where the second equality is by Lemma A.11 with $w \in \mathcal{R}^\gamma$. □

C.3. Lemmas (Technical: Part 2)

Definition C.9. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $i \in [n]$. Define

$$D_i f : \mathbb{R}^n \to \mathbb{R}^m \cup \{\bot\}$$

be the partial derivative of f with respect to its i-th argument, where $\bot$ denotes non-differentiability. Hence, for any $x \in \mathbb{R}^n$ and $i \in [n]$, $Df(x) \neq \bot$ implies $D_i f(x) = (Df(x))_{1:m,\,i}$.

Lemma C.10. *Let $l \in [L]$, $w = (w_1, \dots, w_L) \in \mathbb{R}^W$, and $j \in [W]$ with $j > W_{<l}$, where $W_{<l} \triangleq W_1 + \cdots + W_{l-1}$. Suppose that $\widetilde{z}_l$ is differentiable with respect to its $(N_{l-1} + (j - W_{<l}))$-th argument at $(z_{l-1}(w), w_l, \dots, w_L)$, i.e., $D_{N_{l-1}+(j-W_{<l})}\widetilde{z}_l(z_{l-1}(w), w_l, \dots, w_L) \neq \bot$. Then, there are $\gamma \in \Gamma$ and $\{t_n\}_{n\in\mathbb{N}} \subseteq (v_j, \infty)$ satisfying the following conditions:*

- $w \in \mathcal{R}^\gamma_{\mathsf{cl}}$,
- $D_{N_{l-1}+(j-W_{<l})}\widetilde{z}_l\big(z_{l-1}(w), w_l, \dots, w_L\big) = D_{N_{l-1}+(j-W_{<l})}\widetilde{z}^\gamma_l\big(z_{l-1}(w), w_l, \dots, w_L\big)$,
- $\lim_{n\to\infty} t_n = v_j$, *and*
- $(v_1, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W) \in \mathcal{R}^\gamma$ *for all* $n \in \mathbb{N}$,

where $(v_1, \dots, v_W) \triangleq w$ denotes the scalar values of w (recall that $w_l \in \mathbb{R}^{W_l}$ is not scalar by definition). Further, there are $\gamma' \in \Gamma$ and $\{t'_n\}_{i\in\mathbb{N}} \subseteq (-\infty, v_j)$ that satisfy the same conditions stated above.

Proof. Consider $l \in [L]$, $w \in \mathbb{R}^W$, and $j \in [W]$ stated above. We show the existence of γ and $\{t_n\}_{n\in\mathbb{N}}$, and will omit the proof of the existence of γ' and $\{t'_n\}_{n\in\mathbb{N}}$ since the proof is almost identical.

First, we show that there is $\{t_n\}_{n\in\mathbb{N}} \subseteq (v_j, \infty)$ such that $\lim_{n\to\infty} t_n = v_j$ and

$$\big\{(v_1, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W) \;\big|\; n \in \mathbb{N}\big\} \subseteq \mathcal{R}^\gamma \qquad \text{for some } \gamma \in \Gamma. \tag{17}$$

By the definition of $\mathcal{R}^\gamma$, Eq. (17) is equivalent to the following: for all $(l, i) \in \mathsf{Idx}$,

$$\big\{f_{l,i}(t_n) \;\big|\; n \in \mathbb{N}\big\} \subseteq \mathcal{I}^k_{l,i} \qquad \text{for some } k \in [K_{l,i}],$$

where $f_{l,i} : \mathbb{R} \to \mathbb{R}$ is defined as $f_{l,i}(t) \triangleq y_{l,i}(v_1, \dots, v_{j-1}, t, v_{j+1}, \dots, v_W)$. Note that Lemma C.1 is applicable to $(f_{l,i}, \{\mathcal{I}^k_{l,i}\}_{k\in[K_{l,i}]}, v_j)$, since $\{\mathcal{I}^k_{l,i}\}_{k\in[K_{l,i}]}$ is a finite cover of $\mathbb{R}$ for all (l, i). Hence, by the lemma, there is $\{t_n\}_{n\in\mathbb{N}} \subseteq (v_j, \infty)$ such that $\lim_{n\in\infty} t_n = v_j$ and Eq. (17) holds with some $\gamma \in \Gamma$.

Next, we show that $w \in \mathcal{R}^\gamma_{\mathsf{cl}}$. By the definition of $\mathcal{R}^\gamma_{\mathsf{cl}}$, this is equivalent to $y_{l,i}(w) \in cl(\mathcal{I}^{\gamma(l,i)}_{l,i})$ for all $(l, i) \in \mathsf{Idx}$. To show this, let $(l, i) \in \mathsf{Idx}$. By Eq. (17) and the definition of $\mathcal{R}^\gamma$, we have

$$\big\{y_{l,i}(v_1, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W) \;\big|\; n \in \mathbb{N}\big\} \subseteq \mathcal{I}^{\gamma(l,i)}_{l,i}. \tag{18}$$

Using this, we obtain

$$y_{l,i}(w) = \lim_{n\to\infty} y_{l,i}(v_1, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W) \in cl(\mathcal{I}^{\gamma(l,i)}_{l,i}),$$

where the equality is from the continuity of $y_{l,i}$ (by Lemma A.9) and $\lim_{n\to\infty} t_n = v_j$ (by the above), and the inclusion is by Eq. (18). Hence, we have $w \in \mathcal{R}^\gamma_{\mathsf{cl}}$ as desired.

Lastly, we show that $D_{N_{l-1}+(j-W_{<l})}\widetilde{z}_l(z_{l-1}(w), w_l, \dots, w_L) = D_{N_{l-1}+(j-W_{<l})}\widetilde{z}^\gamma_l(z_{l-1}(w), w_l, \dots, w_L)$. To do so, define $g, g^\gamma : \mathbb{R} \to \mathbb{R}^{N_L}$ as:

$$g(t) \triangleq \widetilde{z}_l\big(z_{l-1}(w), v_{W_{<l}+1}, \dots, v_{j-1}, t, v_{j+1}, \dots, v_W\big),$$
$$g^\gamma(t) \triangleq \widetilde{z}^\gamma_l\big(z_{l-1}(w), v_{W_{<l}+1}, \dots, v_{j-1}, t, v_{j+1}, \dots, v_W\big).$$

Using them, we obtain the desired equation as follows:

$$D_{N_{l-1}+(j-W_{<l})}\widetilde{z}_l\big(z_{l-1}(w), w_l, \dots, w_L\big) = Dg(v_j) = Dg^\gamma(v_j) = D_{N_{l-1}+(j-W_{<l})}\widetilde{z}^\gamma_l\big(z_{l-1}(w), w_l, \dots, w_L\big),$$

where the first and third equalities are by the definition of partial derivatives, and the second equality comes from Lemma C.2 applied to $(g, g^\gamma, v_j, \{t_n\}_{n\in\mathbb{N}})$. Here Lemma C.2 is applicable due to the following: g is differentiable at v_j

(as $Dg(v_j) = D_{N_{l-1}+(j-W_{<l})}\widetilde{z}_l(z_{l-1}(w), w_l, \dots, w_L) \neq \bot$ by assumption); g^γ is differentiable (as $\widetilde{z}_l^\gamma$ is differentiable by Lemma C.5); $\lim_{n\to\infty} t_n = v_j$ with $t_n \neq v_j$ (by the above); and $g(t_n) = g^\gamma(t_n)$ for all $n \in \mathbb{N}$ because

$$\begin{aligned}
g(t_n) &= \widetilde{z}_l\big(z_{l-1}(w), v_{W_{<l}+1}, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W\big) \\
&= \widetilde{z}_l\big(z_{l-1}(v_1, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W), v_{W_{<l}+1}, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W\big) \\
&= \widetilde{z}_l^\gamma\big(z_{l-1}(v_1, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W), v_{W_{<l}+1}, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W\big) \\
&= \widetilde{z}_l^\gamma\big(z_{l-1}(w), v_{W_{<l}+1}, \dots, v_{j-1}, t_n, v_{j+1}, \dots, v_W\big) = g^\gamma(t_n),
\end{aligned}$$

where the second and fourth lines use that z_{l-1} depends only on its first $W_{<l}$ arguments and $j > W_{<l}$, and the third line is by Lemma C.8 and Eq. (17). This completes the proof. □

Lemma C.11. *Let $w = (w_1, \dots, w_L) \in \mathbb{R}^W$ and $(l, i) \in \mathsf{Idx}$. Suppose the following hold: z_L is differentiable at w; τ_l has bias parameters; $\sigma_{l,i}$ is not differentiable at $y_{l,i}(w)$; and for all $\gamma_1, \gamma_2 \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_1} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_2}$, $D_i\widetilde{z}_{l+1}^{\gamma_1}(z_l(w), w_{l+1}, \dots, w_L) = D_i\widetilde{z}_{l+1}^{\gamma_2}(z_l(w), w_{l+1}, \dots, w_L)$. Then, for all $\gamma \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$,*

$$D_i\widetilde{z}_{l+1}^\gamma\big(z_l(w), w_{l+1}, \dots, w_L\big) = (0, \dots, 0).$$

Proof. Consider $w \in \mathbb{R}^W$ and $(l, i) \in \mathsf{Idx}$ satisfying the conditions in the lemma. First, we show that

$$D_{N_{l-1}+(W_l-N_l+i)}\widetilde{z}_l^\gamma(z_{l-1}(w), w_l, \dots, w_L) = D_i\widetilde{z}_{l+1}^\gamma(z_l(w), w_{l+1}, \dots, w_L) \cdot D\sigma_{l,i}^{\gamma(l,i)}(y_{l,i}(w)),$$

for any $\gamma \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$. To this end, we derive two derivatives:

$$\big(D\widetilde{\tau}_l(x)\big)_{*,\, N_{l-1}+(W_l-N_l+i)} \quad \text{and} \quad \big(D\widetilde{\sigma}_l^\gamma(x')\big)_{*,\, i}.$$

Since τ_l has bias parameters (by assumption), and by the definitions of $\widetilde{\tau}_l$ and $\widetilde{\sigma}_l^\gamma$, we have the following: for all $\gamma \in \Gamma$ and $i' \in [N_l + W_{l+1} + \dots + W_L]$, there is $\tau'_{l,i'} : \mathbb{R}^{N_{l-1}+(W_l-N_l)} \to \mathbb{R}$ such that

$$\widetilde{\tau}_{l,i'}(x) = \begin{cases} \tau'_{l,i'}(x_1, \dots, x_{N_{l-1}+(W_l-N_l)}) + x_{N_{l-1}+(W_l-N_l+i')} & \text{if } i' \leq N_l \\ x_{N_{l-1}+(W_l-N_l+i')} & \text{if } i' > N_l, \end{cases}$$

$$\widetilde{\sigma}_{l,i'}^\gamma(x') = \begin{cases} \sigma_{l,i'}^{\gamma(l,i')}(x'_{i'}) & \text{if } i' \leq N_l \\ x'_{i'} & \text{if } i' > N_l, \end{cases}$$

for all $x \in \mathbb{R}^{N_{l-1}+W_l+\dots+W_L}$ and $x' \in \mathbb{R}^{N_l+W_{l+1}+\dots+W_L}$. From this and $i \in [N_l]$, we obtain two derivatives:

$$\big(D\widetilde{\tau}_l(x)\big)_{*,\, N_{l-1}+(W_l-N_l+i)} = e_i, \qquad \big(D\widetilde{\sigma}_l^\gamma(x')\big)_{*,\, i} = e_i \cdot D\sigma_{l,i}^{\gamma(l,i)}(x'_i), \tag{19}$$

where $e_i \in \mathbb{R}^{N_l+W_{l+1}+\dots+W_L}$ denotes the standard unit vector with 1 at the i-th coordinate, $D\sigma_{l,i}^{\gamma(l,i)}(x'_i)$ is considered as a scalar value, and both equalities are by $i \in [N_l]$. Using this, we obtain the following equation for $x \triangleq (z_{l-1}(w), w_l, \dots, w_L)$ and for any $\gamma \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$:

$$\begin{aligned}
D_{N_{l-1}+(W_l-N_l+i)}\widetilde{z}_l^\gamma(x) &= \Big(D\widetilde{z}_l^\gamma(x)\Big)_{*,\, N_{l-1}+(W_l-N_l+i)} \\
&= \Big(D\widetilde{z}_{l+1}^\gamma((\widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l)(x)) \cdot D\widetilde{\sigma}_l^\gamma(\widetilde{\tau}_l(x)) \cdot D\widetilde{\tau}_l(x)\Big)_{*,\, N_{l-1}+(W_l-N_l+i)} \\
&= D\widetilde{z}_{l+1}^\gamma((\widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l)(x)) \cdot D\widetilde{\sigma}_l^\gamma(\widetilde{\tau}_l(x)) \cdot \Big(D\widetilde{\tau}_l(x)\Big)_{*,\, N_{l-1}+(W_l-N_l+i)} \\
&= D\widetilde{z}_{l+1}^\gamma((\widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l)(x)) \cdot \Big(D\widetilde{\sigma}_l^\gamma(\widetilde{\tau}_l(x))\Big)_{*,\, i} \\
&= \Big(D\widetilde{z}_{l+1}^\gamma((\widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l)(x))\Big)_{*,\, i} \cdot D\sigma_{l,i}^{\gamma(l,i)}(\widetilde{\tau}_{l,i}(x)) \\
&= D_i\widetilde{z}_{l+1}^\gamma(z_l(w), w_{l+1}, \dots, w_L) \cdot D\sigma_{l,i}^{\gamma(l,i)}(y_{l,i}(w)),
\end{aligned} \tag{20}$$

where the first two lines use $\widetilde{z}_l^\gamma = \widetilde{z}_{l+1}^\gamma \circ \widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l$ and that $\widetilde{z}_l^\gamma$, $\widetilde{z}_{l+1}^\gamma$, $\widetilde{\sigma}_l^\gamma$, and $\widetilde{\tau}_l$ are differentiable (by Lemma C.5), the fourth and fifth lines use Eq. (19), and the last line uses Lemma C.6 with $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$.

Next, we derive a sufficient condition for the conclusion by using Eq. (20) and applying Lemma C.10 to (l, w, j) with $j \triangleq W_{<l} + (W_l - N_l + i)$, where $W_{<l} \triangleq W_1 + \cdots + W_{l-1}$. Note that the lemma is applicable here due to the following: $W_{<l} < j \leq W$, because $W_l \geq N_l$ (as τ_l has bias parameters by assumption) and $1 \leq i \leq N_l$ (as $(l, i) \in \mathsf{Idx}$); and $\widetilde{z}_l$ is differentiable with respect to its $(N_{l-1} + (W_l - N_l + i))$-th argument at $(z_{l-1}(w), w_l \ldots, w_L)$, because

$$D_{N_{l-1}+(W_l-N_l+i)}\widetilde{z}_l(z_{l-1}(w), w_l, \ldots, w_L) = D_{W_{<l}+(W_l-N_l+i)}z_L(w) \neq \bot$$

where the equality follows from that $z_L(w') = \widetilde{z}_l(z_{l-1}(w'_1, \ldots, w'_{l-1}, 0, \ldots, 0), w'_l, \ldots, w'_L)$ for all $w' \in \mathbb{R}^W$ by Lemma C.7, and the inequality from that z_L is differentiable at w (by assumption). Let $(v_1, \ldots, v_W) \triangleq w$ be the scalar values of w. By applying Lemma C.10 to (l, w, j), it holds that there are $\gamma_+, \gamma_- \in \Gamma$, $\{t_n^+\}_{n\in\mathbb{N}} \subseteq (v_j, \infty)$, and $\{t_n^-\}_{n\in\mathbb{N}} \subseteq (-\infty, v_j)$ such that $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_+} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_-}$ and

$$D_{N_{l-1}+(W_l-N_l+i)}\widetilde{z}_l^{\gamma_+}(x) = D_{N_{l-1}+(W_l-N_l+i)}\widetilde{z}_l^{\gamma_-}(x),$$

$$\{(v_1, \ldots, v_{j-1}, t_n^+, v_{j+1}, \ldots, v_W)\}_{n\in\mathbb{N}} \subseteq \mathcal{R}^{\gamma_+}, \qquad \lim_{n\to\infty} t_n^+ = v_j, \tag{21}$$

$$\{(v_1, \ldots, v_{j-1}, t_n^-, v_{j+1}, \ldots, v_W)\}_{n\in\mathbb{N}} \subseteq \mathcal{R}^{\gamma_-}, \qquad \lim_{n\to\infty} t_n^- = v_j, \tag{22}$$

where $x \triangleq (z_{l-1}(w), w_l, \ldots, w_L)$. By the first line and Eq. (20) with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_+} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_-}$, we have

$$D_i\widetilde{z}_{l+1}^{\gamma_+}(x') \cdot D\sigma_{l,i}^{\gamma_+(l,i)}(y_{l,i}(w)) = D_i\widetilde{z}_{l+1}^{\gamma_-}(x') \cdot D\sigma_{l,i}^{\gamma_-(l,i)}(y_{l,i}(w)),$$

where $x' \triangleq (z_l(w), w_{l+1}, \ldots, w_L)$. From this, and since $D_i\widetilde{z}_{l+1}^\gamma(x')$ is the same for all $\gamma \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$ (by assumption), we immediately obtain the conclusion (i.e., $D_i\widetilde{z}_{l+1}^\gamma(x') = (0, \ldots, 0)$ for all $\gamma \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$) if the following holds:

$$D\sigma_{l,i}^{\gamma_+(l,i)}(y_{l,i}(w)) \neq D\sigma_{l,i}^{\gamma_-(l,i)}(y_{l,i}(w)). \tag{23}$$

Hence, to prove the conclusion, it suffices to show Eq. (23).

Finally, we prove Eq. (23) in two steps. We first show that there are $\delta^+, \delta^- > 0$ such that

$$\big(y_{l,i}(w), y_{l,i}(w) + \delta^+\big) \subseteq \mathcal{I}_{l,i}^{\gamma_+(l,i)}, \qquad \big(y_{l,i}(w) - \delta^-, y_{l,i}(w)\big) \subseteq \mathcal{I}_{l,i}^{\gamma_+(l,i)}. \tag{24}$$

Fix $j \triangleq W_{<l} + (W_l - N_l + i)$ and $(v_1, \ldots, v_W) \triangleq w$ as above. Observe that we have

$$\begin{aligned}
&y_{l,i}(v_1, \ldots, v_{j-1}, t_n^+, v_{j+1}, \ldots, v_W) \in \mathcal{I}_{l,i}^{\gamma_+(l,i)} \text{ for all } n \in \mathbb{N},\\
&y_{l,i}(v_1, \ldots, v_{j-1}, t_n^+, v_{j+1}, \ldots, v_W) > y_{l,i}(v_1, \ldots, v_W) = y_{l,i}(w) \text{ for all } n \in \mathbb{N},\\
&\lim_{n\to\infty} y_{l,i}(v_1, \ldots, v_{j-1}, t_n^+, v_{j+1}, \ldots, v_W) = y_{l,i}(v_1, \ldots, v_W) = y_{l,i}(w),
\end{aligned}$$

where the first line uses Eq. (21), the third line uses Eq. (21) and that $y_{l,i}$ is continuous (by Lemma A.9), and the second line uses the following and that $t_n^+ > v_j$ for all $n \in \mathbb{N}$: for all $t \in \mathbb{R}$,

$$y_{l,i}(v_1, \ldots, v_{j-1}, t, v_{j+1}, \ldots, v_W) = \tau'_{l,i}(z_{l-1}(w), v_{W_{<l}+1}, \ldots, v_{W_{<l}+(W_l-N_l)}) + t,$$

which holds since z_{l-1} depends only on its first $W_{<l}$ arguments, τ_l has bias parameters, and $j = W_{<l} + (W_l - N_l + i) > W_{<l}$. By these results, and since $\mathcal{I}_{l,i}^{\gamma_+(l,i)}$ is an interval, there is $\delta^+ > 0$ satisfying Eq. (24); similarly, there is $\delta^- > 0$ satisfying Eq. (24), due to Eq. (22) and $t_n^- < v_j$ for all n.

We next show that Eq. (23) indeed holds. By Eq. (24) and $\sigma_{l,i} = \sigma_{l,i}^k$ on $\mathcal{I}_{l,i}^k$ for all k, we have

$$\sigma_{l,i} = \sigma_{l,i}^{\gamma_+(l,i)} \text{ on } \big[y_{l,i}(w), y_{l,i}(w) + \delta^+\big), \qquad \sigma_{l,i} = \sigma_{l,i}^{\gamma_-(l,i)} \text{ on } \big(y_{l,i}(w) - \delta^-, y_{l,i}(w)\big],$$

where the inclusion of $y_{l,i}(w)$ is by that $\sigma_{l,i}$ and $\sigma_{l,i}^k$ are continuous for all k. From this, we have

$$D\sigma_{l,i}^{\gamma_+(l,i)}\big(y_{l,i}(w)\big) = \lim_{h\to 0^+} \frac{1}{h}\Big(\sigma_{l,i}(y_{l,i}(w) + h) - \sigma_{l,i}(y_{l,i}(w))\Big),$$

$$D\sigma_{l,i}^{\gamma_-(l,i)}\big(y_{l,i}(w)\big) = \lim_{h\to 0^-} \frac{1}{h}\Big(\sigma_{l,i}(y_{l,i}(w)+h) - \sigma_{l,i}(y_{l,i}(w))\Big).$$

Suppose here that Eq. (23) does not hold, i.e., $D\sigma_{l,i}^{\gamma_+(l,i)}(y_{l,i}(w)) = D\sigma_{l,i}^{\gamma_-(l,i)}(y_{l,i}(w))$. Then,

$$\lim_{h\to 0} \frac{1}{h}\Big(\sigma_{l,i}(y_{l,i}(w)+h) - \sigma_{l,i}(y_{l,i}(w))\Big) = D\sigma_{l,i}^{\gamma_+(l,i)}\big(y_{l,i}(w)\big) \neq \bot,$$

where the inequality is by that $\sigma_{l,i}^{\gamma_+(l,i)}$ is differentiable. This implies $D\sigma_{l,i}\big(y_{l,i}(w)\big) \neq \bot$, which contradicts to that $\sigma_{l,i}$ is non-differentiable at $y_{l,i}(w)$ (by assumption). Hence, Eq. (23) should hold. □

C.4. Theorem 3.2 (Main Lemmas)

Lemma C.12. *Let $w \in \mathbb{R}^W$ and $j \in [W]$. Suppose that z_L is differentiable with respect to its j-th argument at w (i.e., $D_j z_L(w) \neq \bot$). Then, there is $\gamma \in \Gamma$ such that $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$ and*

$$D_j z_L(w) = D_j z_L^\gamma(w).$$

Proof. Consider $w \in \mathbb{R}^W$ and $j \in [W]$ stated above. First, by Lemma C.7, and since $z_0 = z_0^\gamma$ is a constant function, we have $z_L(w') = \widetilde{z}_1(z_0(0,\dots,0), w')$ and $z_L^\gamma(w') = \widetilde{z}_1^\gamma(z_0(0,\dots,0), w')$ for all $w' \in \mathbb{R}^W$ and $\gamma \in \Gamma$. From this, we have

$$\begin{aligned} D_j z_L(w) &= D_{N_0+j}\widetilde{z}_1\big(z_0(0,\dots,0), w\big) = D_{N_0+j}\widetilde{z}_1\big(z_0(w), w\big), \\ D_j z_L^\gamma(w) &= D_{N_0+j}\widetilde{z}_1^\gamma\big(z_0(0,\dots,0), w\big) = D_{N_0+j}\widetilde{z}_1^\gamma\big(z_0(w), w\big) \quad \text{for all } \gamma \in \Gamma, \end{aligned}$$

where the second and fourth equalities follow from that z_0 is a constant function. Second, by Lemma C.10 applied to $(l = 1, w, j)$, there is $\gamma \in \Gamma$ such that

$$w \in \mathcal{R}_{\mathsf{cl}}^\gamma, \qquad D_{N_0+j}\widetilde{z}_1\big(z_0(w), w\big) = D_{N_0+j}\widetilde{z}_1^\gamma\big(z_0(w), w\big).$$

Here Lemma C.10 is applicable, because $D_{N_0+j}\widetilde{z}_1(z_0(w), w) = D_j z_L(w) \neq \bot$ (by the above and by assumption). From these results, there is $\gamma \in \Gamma$ such that $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$ and $D_j z_L(w) = D_j z_L^\gamma(w)$. □

Lemma C.13. *Let $w \in \mathbb{R}^W$. Suppose that the following hold:*

- *z_L is differentiable at w.*
- *For all $l \in [L]$, if τ_l does not have bias parameters, then $\sigma_{l,i}$ is differentiable at $y_{l,i}(w)$ for all $i \in [N_l]$.*

Then, for all $\gamma_1, \gamma_2 \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_1} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_2}$,

$$D z_L^{\gamma_1}(w) = D z_L^{\gamma_2}(w).$$

Proof. Let $w \in \mathbb{R}^W$. Consider the following claim: for all $l \in [L+1]$ and $\gamma_1, \gamma_2 \in \Gamma$, if $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_1} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_2}$, then

$$D\widetilde{z}_l^{\gamma_1}\big(z_{l-1}(w), w_l, \dots, w_L\big) = D\widetilde{z}_l^{\gamma_2}\big(z_{l-1}(w), w_l, \dots, w_L\big).$$

Note that the claim implies the conclusion: for any $\gamma_1, \gamma_2 \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_1} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_2}$,

$$D z_L^{\gamma_1}(w) = \Big(D\widetilde{z}_1^{\gamma_1}(z_0(0,\dots,0), w)\Big)_{*,\, N_0+1:N_0+W} = \Big(D\widetilde{z}_1^{\gamma_2}(z_0(0,\dots,0), w)\Big)_{*,\, N_0+1:N_0+W} = D z_L^{\gamma_2}(w),$$

where the first and third equalities follow from that $z_L^\gamma(w') = \widetilde{z}_1^\gamma(z_0(0,\dots,0), w')$ for all $\gamma \in \Gamma$ and $w' \in \mathbb{R}^W$ (by Lemma C.7 and since $z_0^\gamma = z_0$ is a constant function), and $\widetilde{z}_1^\gamma$ is differentiable for all $\gamma \in \Gamma$ (by Lemma C.5); and the second equality is by the claim for $l = 1$ and that z_0 is a constant function. We prove the claim by induction on l (starting from $L+1$).

Case $l = L+1$. The claim clearly holds, since $\widetilde{z}_{L+1}^\gamma$ is the identity function for all $\gamma \in \Gamma$.

Case $l < L+1$. To show the claim, we first analyze the derivatives mentioned in the claim. Let $\gamma \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^\gamma$, and consider any $x \in \mathbb{R}^{N_{l-1}+W_l+\dots+W_L}$ and $x' \in \mathbb{R}^{N_l+W_{l+1}+\dots+W_L}$. Recall the definition of $\widetilde{z}_l^\gamma$ and $\widetilde{\sigma}_l^\gamma$: for all $i \in [N_l + W_{l+1} + \dots + W_L]$,

$$\widetilde{z}_l^\gamma(x) = (\widetilde{z}_{l+1}^\gamma \circ \widetilde{\sigma}_l^\gamma \circ \widetilde{\tau}_l)(x), \qquad \widetilde{\sigma}_{l,i}^\gamma(x') = \begin{cases} \sigma_{l,i}^{\gamma(l,i)}(x'_i) & \text{if } i \leq N_l \\ x'_i & \text{if } i > N_l. \end{cases}$$

Since every function in the RHS of the above equation is differentiable (by Lemma C.5), the following hold for all $i \in [N_l + W_{l+1} + \cdots + W_L]$:

$$D\widetilde{z}_l^{\gamma}(x) = D\widetilde{z}_{l+1}^{\gamma}\big((\widetilde{\sigma}_l^{\gamma} \circ \widetilde{\tau}_l)(x)\big) \cdot D\widetilde{\sigma}_l^{\gamma}\big(\widetilde{\tau}_l(x)\big) \cdot D\widetilde{\tau}_l(x), \tag{25}$$

$$\big(D\widetilde{\sigma}_l^{\gamma}(x')\big)_{*,i} = \begin{cases} e_i \cdot D\sigma_{l,i}^{\gamma(l,i)}(x'_i) & \text{if } i \le N_l \\ e_i & \text{if } i > N_l, \end{cases}$$

where the first line uses the chain rule, $e_i \in \mathbb{R}^{N_l + W_{l+1} + \cdots + W_L}$ denotes the standard unit vector with 1 at the i-th coordinate, and $D\sigma_{l,i}^{\gamma(l,i)}(x'_i)$ is considered as a scalar value. By the second line, the following holds for all $i \in [N_l + W_{l+1} + \cdots + W_L]$:

$$\Big(D\widetilde{z}_{l+1}^{\gamma}((\widetilde{\sigma}_l^{\gamma} \circ \widetilde{\tau}_l)(x)) \cdot D\widetilde{\sigma}_l^{\gamma}(\widetilde{\tau}_l(x))\Big)_{*,i} = D_i\widetilde{z}_{l+1}^{\gamma}\big((\widetilde{\sigma}_l^{\gamma} \circ \widetilde{\tau}_l)(x)\big) \cdot \begin{cases} D\sigma_{l,i}^{\gamma(l,i)}\big(\widetilde{\tau}_{l,i}(x)\big) & \text{if } i \le N_l \\ 1 & \text{if } i > N_l. \end{cases} \tag{26}$$

We can further simplify the two term in the RHS when $x = (z_{l-1}(w), w_l, \ldots, w_L)$, as follows:

$$D\widetilde{z}_{l+1}^{\gamma}\big((\widetilde{\sigma}_l^{\gamma} \circ \widetilde{\tau}_l)(x)\big) = D\widetilde{z}_{l+1}^{\gamma}(x'), \quad D\sigma_{l,i}^{\gamma(l,i)}\big(\widetilde{\tau}_{l,i}(x)\big) = D\sigma_{l,i}^{\gamma(l,i)}\big(y_{l,i}(w)\big) \text{ for all } i \in [N_l], \tag{27}$$

where $x' \triangleq (z_l(w), w_{l+1}, \ldots, w_L)$ and both equalities are by Lemma C.6 with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma}$.

We now prove the claim. Let $\gamma_1, \gamma_2 \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_1} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_2}$, and fix $x \triangleq (z_{l-1}(w), w_l, \ldots, w_L)$ and $x' \triangleq (z_l(w), w_{l+1}, \ldots, w_L)$. By induction hypothesis on $l+1$, we obtain

$$D\widetilde{z}_{l+1}^{\gamma_1}(x') = D\widetilde{z}_{l+1}^{\gamma_2}(x'). \tag{28}$$

Since we want to show $D\widetilde{z}_l^{\gamma_1}(x) = D\widetilde{z}_l^{\gamma_2}(x)$, it suffices to show the following due to Eqs. (25)–(28): for all $i \in [N_l]$,

$$D_i\widetilde{z}_{l+1}^{\gamma_1}(x') \cdot D\sigma_{l,i}^{\gamma_1(l,i)}\big(y_{l,i}(w)\big) = D_i\widetilde{z}_{l+1}^{\gamma_1}(x') \cdot D\sigma_{l,i}^{\gamma_2(l,i)}\big(y_{l,i}(w)\big). \tag{29}$$

Let $i \in [N_l]$. We prove Eq. (29) by case analysis on i.

Subcase 1: $\sigma_{l,i}$ is non-differentiable at $y_{l,i}(w)$. To show Eq. (29), it suffices to show that

$$D_i\widetilde{z}_{l+1}^{\gamma_1}(x') = (0, \ldots, 0).$$

We obtain this equation by applying Lemma C.11 to $(w, (l,i), \gamma_1)$. Note that the lemma is applicable here because: z_L is differentiable at w (by assumption); $\sigma_{l,i}$ is non-differentiable at $y_{l,i}(w)$ and so τ_l has bias parameters (by assumption); $D_i\widetilde{z}_{l+1}^{\gamma}(x')$ is independent of γ for all $\gamma \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma}$ (by induction hypothesis on $l+1$); and $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_1}$.

Subcase 2: $\sigma_{l,i}$ is differentiable at $y_{l,i}(w)$. To show Eq. (29), it suffices to show that for all $j \in [2]$,

$$D\sigma_{l,i}^{\gamma_j(l,i)}\big(y_{l,i}(w)\big) = D\sigma_{l,i}\big(y_{l,i}(w)\big). \tag{30}$$

Let $j \in [2]$. If $y_{l,i}(w) \in \mathcal{I}_{l,i}^{\gamma_j(l,i)}$, then we obtain Eq. (30) as follows:

$$D\sigma_{l,i}^{\gamma_j(l,i)}\big(y_{l,i}(w)\big) = D^{\mathtt{AD}}\sigma_{l,i}\big(y_{l,i}(w)\big) = D\sigma_{l,i}\big(y_{l,i}(w)\big),$$

where the first equality holds because $y_{l,i}(w) \in \mathcal{I}_{l,i}^{\gamma_j(l,i)}$ and $\{(\sigma_{l,i}^k, \mathcal{I}_{l,i}^k)\}_{k \in [K_{l,i}]}$ defines $D^{\mathtt{AD}}\sigma_{l,i}$; and the second equality holds because $\sigma_{l,i}$ is differentiable at $y_{l,i}(w)$ and $D^{\mathtt{AD}}\sigma_{l,i}$ is an extended derivative of $\sigma_{l,i}$. If $y_{l,i}(w) \notin \mathcal{I}_{l,i}^{\gamma_j(l,i)}$, then we obtain Eq. (30) directly from Lemma C.2 applied to $(\sigma_{l,i}^{\gamma_j(l,i)}, \sigma_{l,i}, y_{l,i}(w))$. Note that the lemma is applicable here because: $\sigma_{l,i}^{\gamma_j(l,i)}$ and $\sigma_{l,i}$ are differentiable at $y_{l,i}(w)$; they coincide on $\mathcal{I}_{l,i}^{\gamma_j(l,i)}$; $y_{l,i}(w) \in cl(\mathcal{I}_{l,i}^{\gamma(l,i)})$ (by $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma}$); and $int(\mathcal{I}_{l,i}^{\gamma(l,i)}) \neq \emptyset$ (by $y_{l,i}(w) \notin \mathcal{I}_{l,i}^{\gamma(l,i)}$). Therefore, Eq. (29) holds and this completes the proof. □

C.5. Theorem 3.2 (Main Proof)

Theorem 3.2. *If z_L has bias parameters, then for all $w \in \mathbb{R}^W$ at which z_L is differentiable,*

$$D^{\mathtt{AD}}z_L(w) = Dz_L(w).$$

This implies that $|\mathsf{inc}_{\Omega}(z_L)| = 0$.

Proof. Let $w \in \mathbb{R}^W$ such that z_L is differentiable at w (i.e., $Dz_L(w) \neq \bot$). By Lemma A.8, there is (unique) $\gamma \in \Gamma$ such that $w \in \mathcal{R}^\gamma$. Using the γ, we obtain the conclusion:

$$\begin{aligned} Dz_L(w) &= \left[\, D_1 z_L(w) \;\middle|\; \cdots \;\middle|\; D_W z_L(w) \,\right] \\ &= \left[\, D_1 z_L^{\gamma_1}(w) \;\middle|\; \cdots \;\middle|\; D_W z_L^{\gamma_W}(w) \,\right] \quad \text{for some } \gamma_j \in \Gamma \text{ with } w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_j} \ (j \in [W]) \\ &= \left[\, D_1 z_L^{\gamma}(w) \;\middle|\; \cdots \;\middle|\; D_W z_L^{\gamma}(w) \,\right] \\ &= Dz_L^\gamma(w) = D^{\mathsf{AD}} z_L(w). \end{aligned}$$

Here the second line uses Lemma C.12 with that z_L is differentiable at w. The third line uses Lemma C.13 with the following: z_L is differentiable at w; τ_l has bias parameters for all $l \in [L]$ (by assumption); and $w \in \mathcal{R}^\gamma \subseteq \mathcal{R}_{\mathsf{cl}}^\gamma$ and $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_j}$ for all $j \in [W]$ (by the second line). The last line uses Lemma A.12 with $w \in \mathcal{R}^\gamma$. □

C.6. Lemmas (Technical: Part 3)

Lemma C.14. *Let z_L be a neural network, $w \in \mathbb{R}^W$, and $A \subseteq \mathsf{Idx}$. Suppose that $w \notin pbd(\{u \in \mathbb{R}^W \mid y_{l,i}(u) = c\})$ for all $(l,i) \in A$ and $c \in \mathsf{ndf}(\sigma_{l,i})$. Then, there is a neural network z'_L (which consists of τ'_l, $\sigma'_{l,i}$, y'_l, z'_l, and $D^{\mathsf{AD}}\sigma'_{l,i}$) satisfying the following conditions:*

① *$Dz'_L(w) = Dz_L(w)$, $D^{\mathsf{AD}} z'_L(w) = D^{\mathsf{AD}} z_L(w)$, and $\tau'_l = \tau_l$ for all $l \in [L]$.*
② *$y'_{l,i}(w) \notin \mathsf{ndf}(\sigma'_{l,i})$ for all $(l,i) \in A$.*

Proof. Consider the setup given above. Define a function f from neural networks to $\mathbb{N}$ as:

$$f(z'_L) \triangleq \left| \left\{ (l,i,c) \;\middle|\; (l,i) \in A,\ c \in \mathsf{ndf}(\sigma'_{l,i}),\ w \in int\big(\{u \in \mathbb{R}^W \mid y'_{l,i}(u) = c\}\big) \right\} \right|.$$

Note that $f(z_L) \in \mathbb{N}$ (i.e., $f(z_L) < \infty$), because $\sigma_{l,i}$ is continuous, piecewise-analytic and so $|\mathsf{ndf}(\sigma_{l,i})| < \infty$ for all $(l,i) \in \mathsf{Idx}$ (by Theorem A.4). The proof proceeds by induction on $f(z_L)$.

Case $f(z_L) = 0$. We claim that z_L satisfies ①-②. Clearly, it satisfies ①. Further, it also satisfies ②: by the assumption and $f(z_L) = 0$, we have that for all $(l,i) \in A$ and $c \in \mathsf{ndf}(\sigma_{l,i})$,

$$w \notin pbd\big(\{u \in \mathbb{R}^W \mid y_{l,i}(u) = c\}\big) \cup int\big(\{u \in \mathbb{R}^W \mid y_{l,i}(u) = c\}\big) = \{u \in \mathbb{R}^W \mid y_{l,i}(u) = c\},$$

which implies that $y_{l,i}(w) \neq c$. Hence, $y_{l,i}(w) \notin \mathsf{ndf}(\sigma_{l,i})$ for all $(l,i) \in A$, as desired.

Case $f(z_L) > 0$. Since $f(z_L) > 0$, there are $(l,i) \in A$ and $c \in \mathsf{ndf}(\sigma_{l,i})$ such that $w \in int(\{u \in \mathbb{R}^W \mid y_{l,i}(u) = c\})$. This implies that there is an open $U \subseteq \mathbb{R}^W$ such that $w \in U$ and $y_{l,i}(u) = c$ for all $u \in U$. Let z'_L be the exactly same neural network as z_L except that it uses different $\sigma'_{l,i}$ and $D^{\mathsf{AD}}\sigma'_{l,i}$:

$$\sigma'_{l,i}(x) \triangleq D^{\mathsf{AD}}\sigma_{l,i}(c) \cdot (x - c) + \sigma_{l,i}(c), \qquad D^{\mathsf{AD}}\sigma'_{l,i}(x) \triangleq D^{\mathsf{AD}}\sigma_{l,i}(c).$$

Note that $\sigma'_{l,i}$ and $D^{\mathsf{AD}}\sigma'_{l,i}$ satisfy the assumptions in §2.2–§2.3: the former is continuous and piecewise-analytic (since it is differentiable), and the latter is an extended derivative of the former (since the former is differentiable and $D^{\mathsf{AD}}\sigma'_{l,i} = D\sigma'_{l,i}$). Moreover, z'_L satisfies ① because $y_{l,i}(u) = c$ for all $u \in U$, and because $\sigma'_{l,i}(c) = \sigma_{l,i}(c)$ and $D^{\mathsf{AD}}\sigma'_{l,i}(c) = D^{\mathsf{AD}}\sigma_{l,i}(c)$. Further, we have that $w \notin pbd(\{u \in \mathbb{R}^W \mid y'_{l',i'}(u) = c'\})$ for all $(l',i') \in A$ and all $c' \in \mathsf{ndf}(\sigma'_{l',i'})$, and that

$$f(z'_L) = f(z_L) - 1,$$

where both results follow from $\mathsf{ndf}(\sigma'_{l,i}) = \emptyset$, $\mathsf{ndf}(\sigma'_{l',i'}) = \mathsf{ndf}(\sigma_{l',i'})$ for all $(l',i') \neq (l,i)$, and $y'_{l',i'} = y_{l',i'}$ on U for all $(l',i') \in \mathsf{Idx}$. Hence, we can apply induction to z'_L, and by induction hypothesis, there is a neural network z''_L such that (z''_L, z'_L) (instead of (z'_L, z_L)) satisfies ①-②. From this, and since (z'_L, z_L) satisfies ① (by the above), we conclude (z''_L, z_L) satisfies ①-②, as desired. □

Lemma C.15. *We have*

$$\mathsf{inc}_{\mathbb{R}}(z_L) \subseteq \bigcup_{(l,i) \in \mathsf{Idx}} \ \bigcup_{c \in \mathsf{ndf}(\sigma_{l,i}) \cap S_l} pbd\big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) = c\}\big),$$

where $S_l \subseteq \mathbb{R}$ is defined by $S_l \triangleq \emptyset$ if τ_l has bias parameters, and $S_l \triangleq \mathbb{R}$ otherwise.

Proof. Let $U \subseteq \mathbb{R}^W$ be the RHS of the above equation:

$$U \triangleq \bigcup_{l \in [L]: S_l = \mathbb{R}} \bigcup_{i \in [N_l]} \bigcup_{c \in \mathsf{ndf}(\sigma_{l,i})} pbd\big(\{w \in \mathbb{R}^W \mid y_{l,i}(w) = c\}\big).$$

Then, it suffices to show that for any $w \in \mathbb{R}^W$, $w \notin U$ implies $w \notin \mathsf{inc}_{\mathbb{R}}(z_L)$. Consider any $w \in \mathbb{R}^W$ with $w \notin U$. We want to show $w \notin \mathsf{inc}_{\mathbb{R}}(z_L)$. If z_L is not differentiable at w, then $w \notin \mathsf{inc}_{\mathbb{R}}(z_L)$ clearly holds by the definition of $\mathsf{inc}_{\mathbb{R}}(-)$. Hence, assume that z_L is differentiable at w. By the definition of $\mathsf{inc}_{\mathbb{R}}(-)$, it suffices to show the following:

$$D^{\mathtt{AD}} z_L(w) = D z_L(w). \tag{31}$$

We prove this in two steps.

Step 1. Since z_L at w does not satisfy the assumption of Lemma C.13 (which we will apply to show Eq. (31)), we construct another neural network z'_L that is identical to z_L nearby w while satisfying the assumption. To do so, we apply Lemma C.14 to (z_L, w, A) with $A \triangleq \{(l,i) \in \mathsf{Idx} \mid S_l = \mathbb{R}\}$. The lemma is applicable here, since $w \notin pbd(\{v \in \mathbb{R}^W \mid y_{l,i}(v) = c\})$ for all $(l,i) \in A$ and $c \in \mathsf{ndf}(\sigma_{l,i})$ (by $w \notin U$). Hence, by Lemma C.14, we get a neural network z'_L (which consists of τ'_l, $\sigma'_{l,i}$, y'_l, z'_l, and $D^{\mathtt{AD}}\sigma'_{l,i}$) satisfying the following conditions:

① $Dz'_L(w) = Dz_L(w)$, $D^{\mathtt{AD}} z'_L(w) = D^{\mathtt{AD}} z_L(w)$, and $\tau'_l = \tau_l$ for all $l \in [L]$.
② $y'_{l,i}(w) \notin \mathsf{ndf}(\sigma'_{l,i})$ for all $(l,i) \in A$.

Step 2. We now prove Eq. (31) based on z'_L. Let Γ', $\mathcal{R}'$, and $\mathcal{R}'_{\mathsf{cl}}$ be the counterparts of Γ, $\mathcal{R}$, and $\mathcal{R}_{\mathsf{cl}}$ for z'_L. Then, by Lemma A.8, there is $\gamma' \in \Gamma'$ such that $w \in \mathcal{R}'^{\gamma'}$. Using z'_L and γ', we obtain Eq. (31):

$$\begin{aligned}
Dz_L(w) &= Dz'_L(w) \\
&= \big[\, D_1 z'_L(w) \mid \cdots \mid D_W z'_L(w) \,\big] \\
&= \big[\, D_1 z'^{\gamma'_1}_L(w) \mid \cdots \mid D_W z'^{\gamma'_W}_L(w) \,\big] \quad \text{for some } \gamma'_j \in \Gamma' \text{ with } w \in \mathcal{R}'^{\gamma'_j}_{\mathsf{cl}} \ (j \in [W]) \\
&= \big[\, D_1 z'^{\gamma'}_L(w) \mid \cdots \mid D_W z'^{\gamma'}_L(w) \,\big] \\
&= D z'^{\gamma'}_L(w) = D^{\mathtt{AD}} z'_L(w) = D^{\mathtt{AD}} z_L(w).
\end{aligned}$$

Here the first and last lines use ① and Lemma A.12 with $w \in \mathcal{R}'^{\gamma'}$. The third line uses Lemma C.12 with that z'_L is differentiable at w (by ①). The fourth line uses Lemma C.13 with the following: z'_L is differentiable at w (by ①); for all $(l,i) \in \mathsf{Idx}$, if τ'_l does not have bias parameters, then $y'_{l,i}(w) \notin \mathsf{ndf}(\sigma'_{l,i})$, i.e., $\sigma'_{l,i}$ is differentiable at $y'_{l,i}(w)$ (by ① and ②); and $w \in \mathcal{R}'^{\gamma'} \subseteq \mathcal{R}'^{\gamma'}_{\mathsf{cl}}$ and $w \in \mathcal{R}'^{\gamma'_j}_{\mathsf{cl}}$ for all $j \in [W]$ (by the third line). □

C.7. Theorem 4.4 (Main Lemma)

Lemma C.16. *Suppose that for every $l \in [L]$, one of the following holds:*

(a) τ_l has bias parameters, or
(b) τ_l is well-structured biaffine.

In the case of (b), let $\phi_{l,i}$ be the partial map described in Lemma B.10 for all $i \in [N_l]$. Then,

$$\mathsf{inc}_{\Omega}(z_L) \subseteq \bigcup_{(l,i) \in \mathsf{Idx}} \bigcup_{c \in A_{l,i}} B_{l,i}(c),$$

where $A_{l,i} \subseteq \mathbb{R}$ and $B_{l,i}(c) \subseteq \Omega$ are defined as

$$A_{l,i} \triangleq \begin{cases} (\mathsf{ndf}(\sigma_{l,i}) \cap S_l) & \text{if } \tau_{l+1} \text{ satisfies the condition (a) or } l = L \\ (\mathsf{ndf}(\sigma_{l,i}) \cap S_l) \cup \mathsf{bdz}(\sigma_{l,i}) & \text{if } \tau_{l+1} \text{ satisfies the condition (b)}, \end{cases}$$

$$B_{l,i}(c) \triangleq \begin{cases} \{w \in \Omega \mid y_{l,i}(w) = c\} & \text{if } \tau_l \text{ satisfies the condition (a)} \\ \{w \in \Omega \mid y_{l,i}(w) = c \wedge \bigvee_{j \in dom(\phi_{l,i})} z_{l-1,\phi_{l,i}(j)}(w) \neq 0\} & \text{if } \tau_l \text{ satisfies the condition (b)}, \end{cases}$$

and $S_l \subseteq \mathbb{R}$ is defined as $S_l \triangleq \emptyset$ if τ_l has bias parameters, and $S_l \triangleq \mathbb{R}$ otherwise.

Proof. We obtain the conclusion by chaining Lemma C.15, Lemma B.11 (which is applicable by the assumption on τ_l), and $\mathsf{ndf}_{\Omega}(z_L) \cup \mathsf{inc}_{\Omega}(z_L) = \big(\mathsf{ndf}_{\mathbb{R}}(z_L) \cup \mathsf{inc}_{\mathbb{R}}(z_L)\big) \cap \Omega$. □

C.8. Theorem 4.4 (Main Proof)

Theorem 4.4. *If τ_l either has bias parameters or is well-structured biaffine for all $l \in [L]$, then*

$$\frac{|\mathsf{inc}_\Omega(z_L)|}{|\Omega|} \leq \frac{1}{|\mathbb{M}|} \sum_{(l,i)\in\mathsf{Idx}} \Big| \big(\mathsf{ndf}(\sigma_{l,i}) \cap S_l\big) \cup \big(\mathsf{bdz}(\sigma_{l,i}) \cap S_{l+1}\big) \Big|,$$

where $S_l \subseteq \mathbb{R}$ is defined by

$$S_l \triangleq \begin{cases} \emptyset & \text{if } l > L \text{ or } \tau_l \text{ has bias parameters} \\ \mathbb{R} & \text{otherwise.} \end{cases}$$

Proof. Observe that

$$\mathsf{inc}_\Omega(z_L) \subseteq \bigcup_{(l,i)\in\mathsf{Idx}} \bigcup_{c\in A_{l,i}} B_{l,i}(c), \qquad |B_{l,i}(c)| \leq |\mathbb{M}|^{W-1}, \tag{32}$$

where $S_l \subseteq \mathbb{R}$, $A_{l,i} \subseteq \mathbb{R}$ and $B_{l,i}(c) \subseteq \Omega$ for $l \in [L]$ are defined as in Lemma C.16. Here the first equation is by Lemma C.16 and the second equation is by Lemmas B.7 and B.13, where these lemmas are applicable by the definition of $B_{l,i}(c)$ and because τ_l either has bias parameters or is well-structured biaffine (by assumption). Observe further that

$$A_{l,i} = (\mathsf{ndf}(\sigma_{l,i}) \cap S_l) \cup (\mathsf{bdz}(\sigma_{l,i}) \cap S_{l+1}), \tag{33}$$

by the definition of $A_{l,i}$ and S_l, where we use $S_{L+1} \triangleq \emptyset$. Combining the above observations, we obtain the conclusion:

$$\frac{|\mathsf{inc}_\Omega(z_L)|}{|\Omega|} \leq \sum_{(l,i)\in\mathsf{Idx}} \sum_{c\in A_{l,i}} \frac{|B_{l,i}(c)|}{|\Omega|} \leq \sum_{(l,i)\in\mathsf{Idx}} \big|(\mathsf{ndf}(\sigma_{l,i}) \cap S_l) \cup (\mathsf{bdz}(\sigma_{l,i}) \cap S_{l+1})\big| \cdot \frac{|\mathbb{M}|^{W-1}}{|\mathbb{M}|^W},$$

where the first inequality uses Eq. (32) and the second inequality uses Eqs. (32) and (33). □

D. Lower Bounds on $|\mathsf{ndf}_\Omega(z_L)|$ and $|\mathsf{inc}_\Omega(z_L)|$

D.1. Theorem 3.4 (Main Proof)

Theorem 3.4. *For any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ with $1 \le |\mathbb{M}| < \infty$, $n \ge 2$, and $\alpha \le |\mathbb{M}|/(n-1)$, there is a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that satisfies*

$$\frac{|\mathsf{ndf}_\Omega(z_L)|}{|\Omega|} \ge \frac{1}{2} \cdot \frac{1}{|\mathbb{M}|} \sum_{(l,i)\in\mathsf{Idx}} |\mathsf{ndf}(\sigma_{l,i})|$$

and the following: z_L has bias parameters, it has $n+1$ neurons, and $|\mathsf{ndf}(\sigma_{1,i})| = \alpha$ for all $i \in [N_1]$.

Proof. Consider any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ that satisfy the assumption. We claim that there is a neural network z_L that has $L = 2$ layers, $N = n+1$ neurons, and $W = n+1$ parameters, and satisfies the given inequality.

We first define a few components to be used in the network. Let $\{x_1, \dots, x_\alpha\} \subseteq \mathbb{M}$ be distinct machine-representable numbers, and $h : \mathbb{R} \to \mathbb{R}$ be a continuous, piecewise-analytic function such that $\mathsf{ndf}(h) = \{x_1, \dots, x_\alpha\}$. Note that such x_j always exists since $|\mathbb{M}| \ge \alpha$ (by assumption). Using h, define a function $f : \mathbb{R}^W \to \mathbb{R}$ as

$$f(w) = w_{n+1} + \sum_{i\in[n]} h(w_i).$$

We assume here (and in the rest of the proof) that $w \in \mathbb{R}^W$ is represented as $w = (w_1, \dots, w_W)$ for $w_i \in \mathbb{R}$ (instead of $w = (w_{1,1}, w_{1,2}, \dots, w_{L,W_L})$ with $w_{l,j} \in \mathbb{R}$ as we assumed so far).

Given these, we construct a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that is essentially the same as f, as follows

$$\begin{aligned}
&z_0(w) = 0 \in \mathbb{R}, && \\
&y_1(w) = (w_1, \dots, w_n) \in \mathbb{R}^n, && z_1(w) = (h(w_1), \dots, h(w_n)) \in \mathbb{R}^n, \\
&y_2(w) = f(x) \in \mathbb{R}, && z_2(w) = f(x) \in \mathbb{R}.
\end{aligned}$$

Then, z_L has 2 layers, $n+1$ neurons, and $n+1$ parameters, and $|\mathsf{ndf}(\sigma_{1,i})| = |\mathsf{ndf}(h)| = \alpha$ for all i. Also, we can easily make all τ_l have bias parameters (e.g., by using $\tau_1(x, w_1, \dots, w_n) = (x + w_1, \dots, x + w_n)$). What remains is to prove that z_L satisfies the inequality in the conclusion. To do so, observe that

$$\begin{aligned}
\mathsf{ndf}_\Omega(z_L) &\supseteq \{w \in \Omega \mid w_i \in \mathsf{ndf}(h) \text{ for some } i \in [n]\} \\
&= \Omega \setminus \{w \in \Omega \mid w_i \notin \mathsf{ndf}(h) \text{ for all } i \in [n]\},
\end{aligned}$$

which follows from the definition of f and $\mathsf{ndf}(h) \subseteq \mathbb{M}$. From this, we have

$$\begin{aligned}
\frac{|\mathsf{ndf}_\Omega(z_L)|}{|\Omega|} &\ge \frac{1}{|\mathbb{M}|^{n+1}} \Big(|\mathbb{M}|^{n+1} - |\mathbb{M}| \cdot (|\mathbb{M}| - \alpha)^n \Big) = 1 - \Big(1 - \frac{\alpha}{|\mathbb{M}|}\Big)^n \\
&\ge 1 - \Big(1 - n\frac{\alpha}{|\mathbb{M}|} + \frac{1}{2}n(n-1)\Big(\frac{\alpha}{|\mathbb{M}|}\Big)^2\Big) = \frac{n\alpha}{|\mathbb{M}|}\Big(1 - \frac{n-1}{2}\frac{\alpha}{|\mathbb{M}|}\Big) \\
&\ge \frac{1}{2} \cdot \frac{n\alpha}{|\mathbb{M}|},
\end{aligned}$$

where the first inequality uses $\mathsf{ndf}(h) \subseteq \mathbb{M}$ and $|\mathsf{ndf}(h)| = \alpha$, the second inequality follows from $(1-x)^n \le 1 - nx + \frac{1}{2}n(n-1)x^2$ (for any $x \le 1$ and $n \in \mathbb{N}$) and $\alpha \le |\mathbb{M}|$, and the third inequality is by the assumption that $\alpha \le |\mathbb{M}|/(n-1)$. By combining this result and

$$\frac{1}{|\mathbb{M}|} \sum_{(l,i)\in\mathsf{Idx}} |\mathsf{ndf}(\sigma_{l,i})| = \frac{n\alpha}{|\mathbb{M}|},$$

we obtain the desired inequality. □

D.2. Theorem 4.3 (Main Proof)

Theorem 4.3. *For any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ with $1 \leq |\mathbb{M}| < \infty$, $n \geq 4$, and $\alpha \leq |\mathbb{M}|/(n-1)$, there is a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that satisfies*

$$\frac{|\mathsf{ndf}_\Omega(z_L)|}{|\Omega|} \geq \frac{1}{9} \cdot \frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} \Big| \mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i}) \Big|$$

and the following: (i) τ_l is well-structured biaffine without bias parameters for all $l < L$, and has bias parameters for $l = L$; (ii) z_L has $n+1$ neurons; and (iii) $|\mathsf{ndf}(\sigma_{1,i})| = \alpha$, $|\mathsf{bdz}(\sigma_{1,i})| = 0$ for all i. We get the same result for (i), (ii'), and (iii'): (ii') z_L has $2n+1$ neurons; and (iii') $|\mathsf{ndf}(\sigma_{1,i})| = 0$, $|\mathsf{bdz}(\sigma_{1,i})| = \alpha$ for all i.

Proof. We prove the two cases (one for (i), (ii), (iii), and the other for (i), (ii'), (iii')) as follows. Consider any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ that satisfy the assumption. Let $\{x_1, \ldots, x_\alpha\} \subseteq \mathbb{M}$ be distinct machine-representable numbers; such x_j always exists since $|\mathbb{M}| \geq \alpha$ (by assumption). In the rest of the proof, we assume that $w \in \mathbb{R}^W$ is represented as $w = (w_1, \ldots, w_W)$ for $w_i \in \mathbb{R}$, as in the proof of Theorem 3.4 (see §D.1).

First case. Let $W = n+1$ and $h : \mathbb{R} \to \mathbb{R}$ be a continuous, piecewise-analytic function such that $\mathsf{ndf}(h) = \{x_1, \ldots, x_\alpha\}$ and $h(x) > 0$ for all $x \in \mathbb{R}$. Using this h, define a function $f : \mathbb{R}^W \to \mathbb{R}$ as

$$f(w) = w_{n+1} + \sum_{i \in [n]} h(w_i).$$

We now construct a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that is essentially the same as f, as follows:

$$\begin{aligned}
&z_0(w) = 1 \in \mathbb{R}, && \\
&y_1(w) = (w_1, \ldots, w_n) \in \mathbb{R}^n, && z_1(w) = (h(w_1), \ldots, h(w_n)) \in \mathbb{R}^n, \\
&y_2(w) = f(x) \in \mathbb{R}, && z_2(w) = f(x) \in \mathbb{R}.
\end{aligned}$$

Then, z_L has $L = 2$ layers, $N = n+1$ neurons, and $W = n+1$ parameters, and $|\mathsf{ndf}(\sigma_{1,i})| = |\mathsf{ndf}(h)| = \alpha$ and $|\mathsf{bdz}(\sigma_{1,i})| = |\mathsf{bdz}(h)| = 0$ for all i. Also, we can easily make τ_l be well-structured biaffine without bias parameters for all $l < L$, and make τ_L have bias parameters (e.g., by using $\tau_1(x, w_1, \ldots, w_n) = (x \cdot w_1, \ldots, x \cdot w_n)$). This shows that (i), (ii), and (iii) are satisfied.

What remains is to prove that z_L satisfies the inequality in the conclusion. To do so, observe that

$$\mathsf{ndf}_\Omega(z_L) \supseteq \{w \in \Omega \mid w_i \in \mathsf{ndf}(h) \text{ for some } i \in [n]\},$$

which follows from the definition of f and $\mathsf{ndf}(h) \subseteq \mathbb{M}$. From this, we have

$$\frac{|\mathsf{ndf}_\Omega(z_L)|}{|\Omega|} \geq \frac{1}{2} \cdot \frac{n\alpha}{|\mathbb{M}|},$$

as shown in the proof of Theorem 3.4 (see §D.1). Here we used $\mathsf{ndf}(h) \subseteq \mathbb{M}$ and $|\mathsf{ndf}(h)| = \alpha$, as well as $\alpha \leq |\mathbb{M}|/(n-1)$ and $n \geq 2$ (by assumption). Further, observe that

$$\frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} \Big| \mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i}) \Big| = \frac{n\alpha + 1}{|\mathbb{M}|} = \left(1 + \frac{1}{n\alpha}\right) \cdot \frac{n\alpha}{|\mathbb{M}|} \leq \frac{3}{2} \cdot \frac{n\alpha}{|\mathbb{M}|},$$

where the inequality uses $n \geq 2$ and $\alpha \geq 1$ (by assumption). From these results, we obtain the desired inequality.

Second case. Let $W = n+2$ and $h : \mathbb{R} \to \mathbb{R}$ be an analytic function such that $h(x_j) = 0$ and $Dh(x_j) = 1$ for all $j \in [\alpha]$, and $|\mathsf{bdz}(h)| = \alpha$. We remark that such a function h always exists due to Hermite interpolation (Burden et al., 2015). Using this h, define a function $f : \mathbb{R}^W \to \mathbb{R}$ as

$$f(w) = w_{n+2} + \sum_{i \in [n]} \mathrm{ReLU}(h(w_i) \cdot w_{n+1}).$$

We now construct a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that is essentially the same as f, as follows:

$$\begin{aligned}
&z_0(w) = 1, && \\
&y_1(w) = (w_1, \dots, w_n), && z_1(w) = (h(w_1), \dots, h(w_n)), \\
&y_2(w) = (h(w_1) \cdot w_{n+1}, \dots, h(w_n) \cdot w_{n+1}), && z_2(w) = (\mathrm{ReLU}(h(w_1) \cdot w_{n+1}), \dots, \mathrm{ReLU}(h(w_n) \cdot w_{n+1})), \\
&y_3(w) = f(x), && z_3(w) = f(x).
\end{aligned}$$

Then, z_L has $L = 3$ layers, $N = 2n + 1$ neurons, and $W = n + 2$ parameters, and $|\mathsf{ndf}(\sigma_{1,i})| = |\mathsf{ndf}(h)| = 0$ and $|\mathsf{bdz}(\sigma_{1,i})| = |\mathsf{bdz}(h)| = \alpha$ for all i. Also, we can easily make τ_l be well-structured biaffine without bias parameters for all $l < L$, and make τ_L have bias parameters, as discussed above. This shows that (i), (ii'), and (iii') are satisfied.

What remains is to prove that z_L satisfies the inequality in the conclusion. To do so, observe that

$$\begin{aligned}
\mathsf{ndf}_\Omega(z_L) &\supseteq \{w \in \Omega \mid w_{n+1} \neq 0 \text{ and } w_i \in \mathsf{bdz}(h) \text{ for some } i \in [n]\} \\
&= \Omega \setminus \big(\{w \in \Omega \mid w_{n+1} = 0\} \cup \{w \in \Omega \mid w_i \notin \mathsf{bdz}(h) \text{ for all } i \in [n]\}\big),
\end{aligned}$$

which follows from the definition of f and $\mathsf{bdz}(h) \subseteq \mathbb{M}$. From this, we have

$$\begin{aligned}
\frac{|\mathsf{ndf}_\Omega(z_L)|}{|\Omega|} &\geq \frac{1}{|\mathbb{M}|^{n+2}} \Big(|\mathbb{M}|^{n+2} - |\mathbb{M}|^{n+1} - |\mathbb{M}|^2 \cdot (|\mathbb{M}| - \alpha)^n \Big) \\
&\geq \frac{1}{2} \cdot \frac{n\alpha}{|\mathbb{M}|} - \frac{1}{|\mathbb{M}|} = \Big(\frac{1}{2} - \frac{1}{n\alpha} \Big) \cdot \frac{n\alpha}{|\mathbb{M}|} \\
&\geq \frac{1}{4} \cdot \frac{n\alpha}{|\mathbb{M}|},
\end{aligned}$$

where the second inequality follows from an argument in the proof of Theorem 3.4 (see §D.1), and the third inequality uses $n \geq 4$ and $\alpha \geq 1$ (by assumption). Note that when proving the second inequality, we used $\mathsf{bdz}(h) \subseteq \mathbb{M}$ and $|\mathsf{bdz}(h)| = \alpha$, as well as $\alpha \leq |\mathbb{M}|/(n-1)$ and $n \geq 2$ (by assumption). Further, observe that

$$\frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} \big| \mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i}) \big| = \frac{n\alpha + n + 1}{|\mathbb{M}|} = \Big(1 + \frac{1}{\alpha} + \frac{1}{n\alpha} \Big) \cdot \frac{n\alpha}{|\mathbb{M}|} \leq \frac{9}{4} \cdot \frac{n\alpha}{|\mathbb{M}|},$$

where the inequality uses $n \geq 4$ and $\alpha \geq 1$ (by assumption). From these results, we obtain the desired inequality. □

D.3. Theorem 4.5 (Main Proof)

Theorem 4.5. *For any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ with $1 \leq |\mathbb{M}| < \infty$, $n \geq 4$, and $\alpha \leq |\mathbb{M}|/(n-1)$, there is a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that satisfies*

$$\frac{|\mathsf{inc}_\Omega(z_L)|}{|\Omega|} \geq \frac{1}{13} \cdot \frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} \Big| \mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i}) \Big|$$

and the following: (i) τ_l is well-structured biaffine without bias parameters for all $l < L$, and has bias parameters for $l = L$; (ii) z_L has $2n+1$ neurons; and (iii) $|\mathsf{ndf}(\sigma_{1,i})| = \alpha$, $|\mathsf{bdz}(\sigma_{1,i})| = 0$ for all i. We get the same result for (i), (ii'), and (iii'): (ii') z_L has $3n+1$ neurons; and (iii') $|\mathsf{ndf}(\sigma_{1,i})| = 0$, $|\mathsf{bdz}(\sigma_{1,i})| = \alpha$ for all i.

Proof. We prove the two cases (one for (i), (ii), (iii), and the other for (i), (ii'), (iii')) as follows. Consider any $\mathbb{M} \subseteq \mathbb{R}$ and $n, \alpha \in \mathbb{N}$ that satisfy the assumption. Let $\{x_1, \dots, x_\alpha\} \subseteq \mathbb{M}$ be distinct machine-representable numbers; such x_j always exists since $|\mathbb{M}| \geq \alpha$ (by assumption). In the rest of the proof, we assume that $w \in \mathbb{R}^W$ is represented as $w = (w_1, \dots, w_W)$ for $w_i \in \mathbb{R}$, as in the proof of Theorem 3.4 (see §D.1).

First case. Let $W = n + 1$. Without loss of generality, assume that α is even and $0 < x_1 < \dots < x_{\alpha/2}$; other cases can be handled in a similar way. Consider a continuous, piecewise-analytic function $h : \mathbb{R} \to \mathbb{R}$ that satisfies the following conditions: for all $j \in [\alpha/2]$, $h(x_j) = 1$ if j is odd, and $h(x_j) = 2$ if j is even; $\mathsf{ndf}(h) \cap (0, \infty) = \{x_1, \dots, x_{\alpha/2}\}$; h is piecewise linear, constant on $[x_{\alpha/2}, \infty)$, and even (i.e., $h(x) = h(-x)$ for all $x \in \mathbb{R}$). For this h, consider a (consistent) extended derivative $D^{\mathsf{AD}} h : \mathbb{R} \to \mathbb{R}$ that takes the slope of the right piece of the function at non-differentiable points: e.g.,

$D^{\mathtt{AD}}h(x_2) = (h(x_3) - h(x_2))/(x_3 - x_2)$ and $D^{\mathtt{AD}}h(-x_2) = (h(-x_1) - h(-x_2))/(-x_1 + x_2)$. Using this h, define a function $f : \mathbb{R}^W \to \mathbb{R}$ as

$$f(w) = w_{n+1} + \sum_{i \in [n]} h(w_i) - h(-w_i).$$

Then, by using a similar approach taken in the proof of Theorem 4.3 (see §D.2), we can construct a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that is essentially the same as f and satisfies the following: z_L has $L = 2$ layers, $N = 2n + 1$ neurons, and $W = n+1$ parameters (where $2n$ neurons are at layer 1 and 1 neuron is at layer 2); τ_l is well-structured biaffine without bias parameters for all $l < L$, and has bias parameters for $l = L$; and $|\mathsf{ndf}(\sigma_{1,i})| = |\mathsf{ndf}(h)| = \alpha$ and $|\mathsf{bdz}(\sigma_{1,i})| = |\mathsf{bdz}(h)| = 0$ for all i. This shows that (i), (ii), and (iii) are satisfied.

What remains is to prove that z_L satisfies the inequality in the conclusion. To do so, observe that

$$\mathsf{inc}_\Omega(z_L) \supseteq \{w \in \Omega \mid w_i \in \{x_1, \ldots, x_{\alpha/2}\} \text{ for some } i \in [n]\},$$

which follows from the definition of f and $\{x_1, \ldots, x_{\alpha/2}\} \subseteq \mathbb{M}$. From this, we have

$$\frac{|\mathsf{inc}_\Omega(z_L)|}{|\Omega|} \geq \frac{1}{4} \cdot \frac{n\alpha}{|\mathbb{M}|}$$

by a similar argument to that in the proof of Theorem 3.4 (see §D.1). Here we used $\{x_1, \ldots, x_{\alpha/2}\} \subseteq \mathbb{M}$ as well as $\alpha \leq |\mathbb{M}|/(n-1)$ and $n \geq 2$ (by assumption). Further, observe that

$$\frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} \left|\mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i})\right| = \frac{2n\alpha + 1}{|\mathbb{M}|} = \left(2 + \frac{1}{n\alpha}\right) \cdot \frac{n\alpha}{|\mathbb{M}|} \leq \frac{9}{4} \cdot \frac{n\alpha}{|\mathbb{M}|},$$

where the inequality uses $n \geq 4$ and $\alpha \geq 1$ (by assumption). From these results, we obtain the desired inequality.

Second case. Let $W = n + 2$ and $h : \mathbb{R} \to \mathbb{R}$ be an analytic function such that $h(x_j) = 0$ and $Dh(x_j) = 1$ for all $j \in [\alpha]$, and $|\mathsf{bdz}(h)| = \alpha$. Using this h, define a function $f : \mathbb{R}^W \to \mathbb{R}$ as

$$f(w) = w_{n+2} + \sum_{i \in [n]} \mathrm{ReLU}(h(w_i) \cdot w_{n+1}) - \mathrm{ReLU}(-h(w_i) \cdot w_{n+1}),$$

and let $D^{\mathtt{AD}}\mathrm{ReLU} = \mathbf{1}_{(0,\infty)}$. By using an approach similar to the above, we can construct a neural network $z_L : \mathbb{R}^W \to \mathbb{R}$ that is essentially the same as f and satisfies the following: z_L has $L = 3$ layers, $N = 3n+1$ neurons, and $W = n+2$ parameters (where n neurons are at layer 1, $2n$ neurons at layer 2, and 1 neuron at layer 3); τ_l is well-structured biaffine without bias parameters for all $l < L$, and has bias parameters for $l = L$; and $|\mathsf{ndf}(\sigma_{1,i})| = |\mathsf{ndf}(h)| = 0$ and $|\mathsf{bdz}(\sigma_{1,i})| = |\mathsf{bdz}(h)| = \alpha$ for all i. This shows that (i), (ii'), and (iii') are satisfied.

What remains is to prove that z_L satisfies the inequality in the conclusion. To do so, observe that

$$\mathsf{inc}_\Omega(z_L) \supseteq \{w \in \Omega \mid w_{n+1} \neq 0 \text{ and } w_i \in \mathsf{bdz}(h) \text{ for some } i \in [n]\},$$

which follows from the definition of f and $\mathsf{bdz}(h) \subseteq \mathbb{M}$. From this, we have

$$\frac{|\mathsf{inc}_\Omega(z_L)|}{|\Omega|} \geq \frac{1}{4} \cdot \frac{n\alpha}{|\mathbb{M}|},$$

as shown in the proof of Theorem 4.3 (see §D.2). Here we used $\mathsf{bdz}(h) \subseteq \mathbb{M}$ and $|\mathsf{bdz}(h)| = \alpha$, as well as $1 \leq \alpha \leq |\mathbb{M}|/(n-1)$ and $n \geq 4$ (by assumption). Further, observe that

$$\frac{1}{|\mathbb{M}|} \sum_{(l,i) \in \mathsf{Idx}} \left|\mathsf{ndf}(\sigma_{l,i}) \cup \mathsf{bdz}(\sigma_{l,i})\right| = \frac{n\alpha + 2n + 1}{|\mathbb{M}|} = \left(1 + \frac{2}{\alpha} + \frac{1}{n\alpha}\right) \cdot \frac{n\alpha}{|\mathbb{M}|} \leq \frac{13}{4} \cdot \frac{n\alpha}{|\mathbb{M}|},$$

where the inequality uses $n \geq 4$ and $\alpha \geq 1$ (by assumption). From these results, we obtain the desired inequality. □

E. Computation of Standard Derivatives

E.1. Lemmas (Basic)

Lemma E.1. *Let $f, f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}^{d'}$ ($n \in \mathbb{N}$), $x \in \mathbb{R}^d$, and $U \subseteq \mathbb{R}^d$ be an open neighborhood of x. Suppose that for all $y \in U$, $f(y) = f_i(y)$ for some $i \in [n]$. Also, assume that $f(x) = f_i(x)$ for all $i \in [n]$, and $Df_i(x) = Df_j(x) \neq \bot$ for all $i, j \in [n]$. Then,*

$$Df(x) = Df_i(x) \neq \bot \qquad \text{for all } i \in [n].$$

Proof. Consider the setup of the statement. By the assumption, it suffices to show that $Df(x) = Df_1(x)$, which is equivalent to the following: for all $\epsilon > 0$, there exists $\delta > 0$ such that for all $h \in \mathbb{R}^d$,

$$0 < \|h\| < \delta \quad \Longrightarrow \quad \frac{\|f(x+h) - f(x) - Df_1(x) \cdot h\|}{\|h\|} < \epsilon,$$

where $\|\cdot\|$ denotes the ℓ_2-norm. To show this, consider any $\epsilon > 0$. Since $Df_i(x) \neq \bot$ (by assumption), there is $\delta_i > 0$ for each $i \in [n]$ such that for all $h \in \mathbb{R}^d$,

$$0 < \|h\| < \delta_i \quad \Longrightarrow \quad \frac{\|f_i(x+h) - f_i(x) - Df_i(x) \cdot h\|}{\|h\|} = \frac{\|f_i(x+h) - f_1(x) - Df_1(x) \cdot h\|}{\|h\|} < \epsilon,$$

where the equality is by assumption. Choose $0 < \delta < \min\{\delta_i \mid i \in [n]\}$ such that $\{x + h \mid \|h\| < \delta\} \subseteq U$, which is possible because U is an open neighborhood of x. Then, for all $h \in \mathbb{R}^d$, $0 < \|h\| < \delta$ implies that

$$\frac{\|f(x+h) - f(x) - Df_1(x) \cdot h\|}{\|h\|} = \frac{\|f_j(x+h) - f_1(x) - Df_1(x) \cdot h\|}{\|h\|} < \epsilon$$

for some $j \in [n]$, where the equality is by assumption and $x + h \in U$ and the inequality is by $\delta < \delta_j$. This proves $Df(x) = Df_1(x)$ as desired. □

E.2. Lemmas (Technical: Part 1)

In this subsection, we formally define the partial derivative $\partial^{\mathtt{AD}} z_L / \partial z_{l,i} \in \mathbb{R}^{N_L}$ of z_L with respect to $z_{l,i}$ that *reverse-mode* automatic differentiation computes (as a byproduct of computing $D^{\mathtt{AD}} z_L$). To do so, we fix $l' \in [L]$ and $w' \in \mathbb{R}^W$, and define $\partial^{\mathtt{AD}} z_L / \partial z_{l',i} \in \mathbb{R}^{N_L}$ at w' ($i \in [N_{l'}]$) in a similar way we defined $D^{\mathtt{AD}} z_L$ in §A.3.

We first define a program $\mathtt{Q}$ (different from $\mathtt{P}$ in §A.3) that represents a function from $\mathbb{R}^{N_{l'}}$ to $\mathbb{R}$ as follows:

$$\mathtt{Q} ::= r \mid \mathtt{x}_i \mid f\,(\mathtt{Q}_1, \ldots, \mathtt{Q}_n)$$

where $r \in \mathbb{R}$, $i \in [N_{l'}]$, $f \in \{\tau_{l,i}, \sigma_{l,i} \mid (l,i) \in \mathsf{Idx}, l > l'\}$, and $n \in \mathbb{N}$. This definition says that a program $\mathtt{Q}$ can be either a real-valued constant r, a real-valued variable $\mathtt{x}_i$ denoting the neuron $z_{l',i}$, or the application of a function $f : \mathbb{R}^n \to \mathbb{R}$ to subprograms $\mathtt{Q}_1, \ldots, \mathtt{Q}_n$. We focus on particular programs $\mathtt{Q}_{y_{l,i}}$ and $\mathtt{Q}_{z_{l,i}}$ ($l > l'$) that represent the neurons $y_{l,i}$ and $z_{l,i}$ but as functions of the neurons $z_{l',1}, \ldots, z_{l',N_{l'}}$ (instead of functions of parameters $w_{1,1}, w_{1,2}, \ldots, w_{L,W_L}$). These programs are defined in a canonical way as follows:

$$\begin{aligned}
\mathtt{Q}_{y_{l,i}} &\triangleq \tau_{l,i}\,(\mathtt{Q}_{z_{l-1,1}}, \ldots, \mathtt{Q}_{z_{l-1,N_{l-1}}}, w'_{l,1}, \ldots, w'_{l,W_l})\,, \\
\mathtt{Q}_{z_{l,i}} &\triangleq \sigma_{l,i}\,(\mathtt{Q}_{y_{l,i}})\,,
\end{aligned}$$

where $\mathtt{Q}_{z_{l',i}} \triangleq \mathtt{x}_i$ for $i \in [N_{l'}]$ represents the projection function from $\mathbb{R}^{N_{l'}}$ to $\mathbb{R}$. Note that $w'_{l,j}$ in the above equation is not a variable but a constant, while $\mathtt{x}_i$ in the definition of $\mathtt{Q}_{z_{l',i}}$ is a variable.

Given a program $\mathtt{Q}$, we define the function $[\![\mathtt{Q}]\!] : \mathbb{R}^{N_{l'}} \to \mathbb{R}$ that $\mathtt{Q}$ represents, and the function $[\![\mathtt{Q}]\!]^{\mathtt{AD}} : \mathbb{R}^{N_{l'}} \to \mathbb{R}^{1 \times N_{l'}}$ that reverse-mode automatic differentiation computes for $\mathtt{Q}$ (as a byproduct of computing other derivatives):

$$\begin{aligned}
[\![r]\!](x) &\triangleq r, \\
[\![\mathtt{x}_i]\!](x) &\triangleq x_i, \\
[\![f\,(\mathtt{Q}_1, \ldots, \mathtt{Q}_n)\,]\!](x) &\triangleq f\big([\![\mathtt{Q}_1]\!](x), \ldots, [\![\mathtt{Q}_n]\!](x)\big),
\end{aligned}$$

$$
\begin{aligned}
[\![r]\!]^{\mathtt{AD}}(x) &\triangleq \mathbb{0}, \\
[\![\mathtt{x}_i]\!]^{\mathtt{AD}}(x) &\triangleq \mathbb{1}_i, \\
[\![f\,(\mathtt{Q}_1, \ldots, \mathtt{Q}_n)\,]\!]^{\mathtt{AD}}(x) &\triangleq D^{\mathtt{AD}} f\big([\![\mathtt{Q}_1]\!](x), \ldots, [\![\mathtt{Q}_n]\!](x)\big) \cdot \big[[\![\mathtt{Q}_1]\!]^{\mathtt{AD}}(x) \,/\, \cdots \,/\, [\![\mathtt{Q}_n]\!]^{\mathtt{AD}}(x)\big].
\end{aligned}
$$

Here $(x_1, \ldots, x_{N_{l'}}) \triangleq x$ denote the scalar values of x, the notation $\mathbb{0}, \mathbb{1}_i \in \mathbb{R}^{1 \times W}$ denote the zero matrix and the matrix whose entries are all zeros except for a single one at the i-th entry, $D^{\mathtt{AD}} f : \mathbb{R}^n \to \mathbb{R}^{1 \times n}$ denotes a "derivative" of f used by automatic differentiation, and $[M_1 \,/\, \cdots \,/\, M_n]$ denotes the matrix that stacks up matrices $M_1, \ldots, M_n$ vertically. Note that the definitions of $[\![\mathtt{Q}]\!]$ and $[\![\mathtt{Q}]\!]^{\mathtt{AD}}$ are almost the same as that of $[\![\mathtt{P}]\!]$ and $[\![\mathtt{P}]\!]^{\mathtt{AD}}$ in §A.3.

Using the above definitions, $\partial^{\mathtt{AD}} z_L / \partial z_{l',i}$ at w' for $i \in [N_{l'}]$ (i.e., the partial derivative of z_L with respect to $z_{l',i}$ at w' that reverse-mode automatic differentiation computes) can be defined as follows:

$$
\frac{\partial^{\mathtt{AD}} z_L}{\partial z_{l',i}} \text{ at } w' \triangleq \Big[[\![\mathtt{Q}_{z_{L,1}}]\!]^{\mathtt{AD}}(z_{l'}(w')) \,/\, \cdots \,/\, [\![\mathtt{Q}_{z_{L,N_L}}]\!]^{\mathtt{AD}}(z_{l'}(w'))\Big]_{1:N_L,\, i} \in \mathbb{R}^{N_L}.
$$

Lemma E.2 (shown below) shows that $\partial^{\mathtt{AD}} z_L / \partial z_{l,i}$ can be expressed in terms of $\widetilde{z}^{\gamma}_{l+1}$ (defined in §C), as $D^{\mathtt{AD}} z_L$ can be expressed in terms of z^{γ}_L (Lemma A.12). We will rely on this lemma in the rest of this section, when working with $\partial^{\mathtt{AD}} z_L / \partial z_{l,i}$.

Lemma E.2. *Let $\gamma \in \Gamma$. Then, for all $l \in [L]$ and $w \in \mathcal{R}^{\gamma}$,*

$$
\frac{\partial^{\mathtt{AD}} z_L}{\partial z_{l,i}} \text{ at } w = D_i \widetilde{z}^{\gamma}_{l+1}\big(z_l(w), w_{l+1}, \ldots, w_L\big).
$$

Proof. The proof is similar to Lemma A.12, except that it uses Lemma C.8 instead of Lemma A.11; thus, we omit it. □

E.3. Lemmas (Technical: Part 2)

Lemma E.3. *Let $w \in \mathbb{R}^W$. Suppose that for all $(l, i) \in \mathsf{Idx}$, $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$ implies that*

$$
D_i \widetilde{z}^{\gamma}_{l+1}(z_l(w), w_{l+1}, \ldots, w_L) = (0, \ldots, 0)
$$

for the $\gamma \in \Gamma$ with $w \in \mathcal{R}^{\gamma}$. Then, for all $l \in [L+1]$ and $\gamma_1, \gamma_2 \in \Gamma$ with $w \in \mathcal{R}^{\gamma_1}_{\mathsf{cl}} \cap \mathcal{R}^{\gamma_2}_{\mathsf{cl}}$,

$$
D\widetilde{z}^{\gamma_1}_l(z_{l-1}(w), w_l, \ldots, w_L) = D\widetilde{z}^{\gamma_2}_l(z_{l-1}(w), w_l, \ldots, w_L).
$$

Proof. The proof is similar to that of Lemma C.13, except that this lemma assumes that certain partial derivatives are all zero while Lemma C.13 derives this assumption (in addition to proving the conclusion of this lemma). Let $w \in \mathbb{R}^W$ that satisfies the assumption of this lemma. The proof proceeds by induction on l (starting from $l = L+1$).

Case $l = L+1$. In this case, $\widetilde{z}^{\gamma}_{L+1}$ is the identity function for all $\gamma \in \Gamma$. Hence, the conclusion clearly holds.

Case $l < L+1$. For simple notation, let $x \triangleq (z_{l-1}(w), w_l, \ldots, w_L)$ and $x' \triangleq (z_l(w), w_{l+1}, \ldots, w_L)$. Observe that the following hold for any $\gamma \in \Gamma$ with $w \in \mathcal{R}^{\gamma}_{\mathsf{cl}}$, due to Eqs. (25)–(27) in the proof of Lemma C.13:

$$
\begin{aligned}
D\widetilde{z}^{\gamma}_l(x) &= D\widetilde{z}^{\gamma}_{l+1}\big((\widetilde{\sigma}^{\gamma}_l \circ \widetilde{\tau}_l)(x)\big) \cdot D\widetilde{\sigma}^{\gamma}_l\big(\widetilde{\tau}_l(x)\big) \cdot D\widetilde{\tau}_l(x), \\
\Big(D\widetilde{z}^{\gamma}_{l+1}((\widetilde{\sigma}^{\gamma}_l \circ \widetilde{\tau}_l)(x)) \cdot D\widetilde{\sigma}^{\gamma}_l(\widetilde{\tau}_l(x))\Big)_{*,\, i} &= D_i \widetilde{z}^{\gamma}_{l+1}(x') \cdot \begin{cases} D\sigma^{\gamma(l,i)}_{l,i}\big(y_{l,i}(w)\big) & \text{if } i \le N_l \\ 1 & \text{if } i > N_l. \end{cases}
\end{aligned}
$$

Using this observation, we prove the conclusion for l. Let $\gamma_1, \gamma_2 \in \Gamma$ with $w \in \mathcal{R}^{\gamma_1}_{\mathsf{cl}} \cap \mathcal{R}^{\gamma_2}_{\mathsf{cl}}$. We want to show $D\widetilde{z}^{\gamma_1}_l(x) = D\widetilde{z}^{\gamma_2}_l(x)$. By induction hypothesis on $l+1$, we obtain $D\widetilde{z}^{\gamma_1}_{l+1}(x') = D\widetilde{z}^{\gamma_2}_{l+1}(x')$. From this and the above equation, it suffices to show the following claim for all $i \in [N_l]$:

$$
D_i \widetilde{z}^{\gamma_1}_{l+1}(x') \cdot D\sigma^{\gamma_1(l,i)}_{l,i}\big(y_{l,i}(w)\big) = D_i \widetilde{z}^{\gamma_1}_{l+1}(x') \cdot D\sigma^{\gamma_2(l,i)}_{l,i}\big(y_{l,i}(w)\big).
$$

Let $i \in [N_l]$. We prove this claim by case analysis on i.

Subcase 1: $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$. Observe that for the $\gamma \in \Gamma$ with $w \in \mathcal{R}^\gamma$, we have

$$D_i \widetilde{z}_{l+1}^{\gamma_1}(x') = D_i \widetilde{z}_{l+1}^{\gamma}(x') = (0, \ldots, 0),$$

where the first equality is by induction hypothesis on $l+1$ with $w \in \mathcal{R}^\gamma \subseteq \mathcal{R}_{\mathsf{cl}}^\gamma$, and the second equality by assumption with $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$. This directly implies the claim.

Subcase 2: $y_{l,i}(w) \notin \mathsf{ndf}(\sigma_{l,i})$. To show the claim, it suffices to show that for all $j \in [2]$,

$$D\sigma_{l,i}^{\gamma_j(l,i)}\big(y_{l,i}(w)\big) = D\sigma_{l,i}\big(y_{l,i}(w)\big).$$

This is exactly the same as Eq. (30) in the proof of Lemma C.13, and we can prove this in the exact same way as before. Therefore, the claim holds and this completes the proof. □

Lemma E.4. *Let $w \in \mathbb{R}^W$. Suppose that for all $\gamma_1, \gamma_2 \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_1} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_2}$,*

$$Dz_L^{\gamma_1}(w) = Dz_L^{\gamma_2}(w).$$

Then, z_L is differentiable at w.

Proof. Consider the setup of this lemma. To apply Lemma E.1, we show the following claims for $\Gamma_w \triangleq \{\gamma \in \Gamma \mid w \in \mathcal{R}_{\mathsf{cl}}^\gamma\}$:

(i) For some open neighborhood $U \subseteq \mathbb{R}^W$ of w, if $w' \in U$, then $z_L(w') = z_L^\gamma(w')$ for some $\gamma \in \Gamma_w$.
(ii) $z_L(w) = z_L^\gamma(w)$ for all $\gamma \in \Gamma_w$.
(iii) $Dz_L^{\gamma_1}(w) = Dz_L^{\gamma_2}(w) \neq \bot$ for all $\gamma_1, \gamma_2 \in \Gamma_w$.

If these claims hold, then Lemma E.1 implies that $Dz_L(w) \neq \bot$ (i.e., z_L is differentiable at w). So what remains is to show these claims. First, (iii) follows from the assumption of this lemma and that z_L^γ is analytic for all $\gamma \in \Gamma$. Second, (ii) follows from Lemma C.6. Finally, (i) holds as follows. Consider any $\gamma \in \Gamma \setminus \Gamma_w$. Then, by $w \notin \mathcal{R}_{\mathsf{cl}}^\gamma$ and the definition of $\mathcal{R}_{\mathsf{cl}}^\gamma$, there is $(l,i) \in \mathsf{Idx}$ such that $y_{l,i}(w) \in A$ and $A \cap \mathcal{I}_{l,i}^{\gamma(l,i)} = \emptyset$ for some open $A \subseteq \mathbb{R}$. Since $y_{l,i}$ is continuous and $\mathcal{R}^\gamma \subseteq y_{l,i}^{-1}(\mathcal{I}_{l,i}^{\gamma(l,i)})$, the set $U^\gamma \triangleq y_{l,i}^{-1}(A)$ is an open neighborhood of w such that $U^\gamma \cap \mathcal{R}^\gamma = \emptyset$. We now define

$$U \triangleq \bigcap_{\gamma \in \Gamma \setminus \Gamma_w} U^\gamma.$$

Then, because Γ is finite, U is an open neighborhood of w such that $U \cap \bigcup_{\gamma \in \Gamma \setminus \Gamma_w} \mathcal{R}^\gamma = \emptyset$. Using this, we obtain (i) as follows: for any $w' \in U$, we have $w' \notin \bigcup_{\gamma \in \Gamma \setminus \Gamma_w} \mathcal{R}^\gamma$ and so $w' \in \mathcal{R}^\gamma$ for some $\gamma \in \Gamma_w$ (by Lemma A.8); this implies that $z_L(w') = z_L^\gamma(w')$ (by Lemma A.11). This completes the proof. □

E.4. Theorems 3.5 and 4.6 (Main Lemmas)

Lemma E.5. *Let $w \in \mathbb{R}^W$. Suppose that the following holds:*

- *For all $(l,i) \in \mathsf{Idx}$, $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$ implies that $D_i \widetilde{z}_{l+1}^\gamma(z_l(w), w_{l+1}, \ldots, w_L) = \vec{0}$ for the $\gamma \in \Gamma$ with $w \in \mathcal{R}^\gamma$.*

Then, we have the following:

- *$w \notin \mathsf{ndf}_{\mathbb{R}}(z_L)$ (i.e., z_L is differentiable at w).*

Proof. Consider the setup of this lemma. For all $\gamma_1, \gamma_2 \in \Gamma$ with $w \in \mathcal{R}_{\mathsf{cl}}^{\gamma_1} \cap \mathcal{R}_{\mathsf{cl}}^{\gamma_2}$,

$$Dz_L^{\gamma_1}(w) = D\widetilde{z}_1^{\gamma_1}(z_0(w), w_1, \ldots, w_L) = D\widetilde{z}_1^{\gamma_2}(z_0(w), w_1, \ldots, w_L) = Dz_L^{\gamma_2}(w),$$

where the first and last equalities are by Lemma C.7, and the second equality is by Lemma E.3. Then, by applying Lemma E.4, we obtain that z_L is differentiable at w, as desired. □

Lemma E.6. *Let $w \in \mathbb{R}^W$. Suppose that the following hold:*

- *$w \notin \mathsf{ndf}_{\mathbb{R}}(z_L)$ (i.e., z_L is differentiable at w).*
- *For all $(l,i) \in \mathsf{Idx}$, $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$ implies that τ_l has bias parameters.*

Then, we have the following:

- $w \notin \mathsf{ndf}_{\mathbb{R}}(z_L) \cup \mathsf{inc}_{\mathbb{R}}(z_L)$ *(i.e.,* $D^{\mathtt{AD}} z_L(w) = Dz_L(w) \neq \bot$*).*
- *For all* $(l, i) \in \mathsf{Idx}$*,* $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$ *implies that* $D_i \widetilde{z}^{\gamma}_{l+1}(z_l(w), w_{l+1}, \ldots, w_L) = \vec{0}$ *for the* $\gamma \in \Gamma$ *with* $w \in \mathcal{R}^{\gamma}$*.*

Proof. Consider the setup in the statement. By exactly following the proof of Theorem 3.2 (given in §C.5) under this setup, we obtain the conclusion of Theorem 3.2: $D^{\mathtt{AD}} z_L(w) = Dz_L(w)$, which implies the first conclusion of this lemma. Moreover, the second conclusion was already shown in the proof of Lemma C.13 (which has the same assumption as this lemma), especially in Subcase 1 of Case $l < L + 1$ in the proof. This completes the proof. □

E.5. Theorems 3.5 and 4.6 (Main Proofs)

Theorem 3.5. *If* z_L *has bias parameters, then the following are equivalent for all* $w \in \mathbb{R}^W$*.*

- z_L *is non-differentiable at* w*.*
- $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$ *and* $\partial^{\mathtt{AD}} z_L / \partial z_{l,i} \neq \vec{0}$ *at* w *for some* $(l, i) \in \mathsf{Idx}$*.*

Proof. Let $w \in \mathbb{R}^W$. Suppose that z_L has bias parameters. Then, by Lemmas E.5 and E.6, the following are equivalent:

(i) $w \notin \mathsf{ndf}_{\mathbb{R}}(z_L)$ (i.e., z_L is differentiable at w).
(ii) For all $(l, i) \in \mathsf{Idx}$, $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$ implies that $D_i \widetilde{z}^{\gamma}_{l+1}(z_l(w), w_{l+1}, \ldots, w_L) = \vec{0}$ for the $\gamma \in \Gamma$ with $w \in \mathcal{R}^{\gamma}$.

By taking the negation of (i)-(ii) and applying Lemma E.2 to (ii), we obtain the conclusion. □

Theorem 4.6. *Let* $w \in \mathbb{R}^W$*. If* $y_{l,i}(w) \notin \mathsf{ndf}(\sigma_{l,i})$ *for all* $(l, i) \in \mathsf{Idx}$ *such that* τ_l *does not have bias parameters or* $\partial^{\mathtt{AD}} z_L / \partial z_{l,i} \neq \vec{0}$ *at* w*, then*

$$D^{\mathtt{AD}} z_L(w) = Dz_L(w) \neq \bot.$$

Proof. Let $w \in \mathbb{R}^W$. Suppose that it satisfies the given assumption, which is equivalent to the following by Lemma E.2:

For all $(l, i) \in \mathsf{Idx}$, $y_{l,i}(w) \in \mathsf{ndf}(\sigma_{l,i})$ implies that
(i) τ_l has bias parameters, and
(ii) $D_i \widetilde{z}^{\gamma}_{l+1}(z_l(w), w_{l+1}, \ldots, w_L) = \vec{0}$ for the $\gamma \in \Gamma$ with $w \in \mathcal{R}^{\gamma}$.

First, by Lemma E.5 with (ii), we have

(iii) $w \notin \mathsf{ndf}_{\mathbb{R}}(z_L)$ (i.e., z_L is differentiable at w).

Next, by Lemma E.6 with (i) and (iii), we have the conclusion:

$w \notin \mathsf{ndf}_{\mathbb{R}}(z_L) \cup \mathsf{inc}_{\mathbb{R}}(z_L)$ (i.e., $D^{\mathtt{AD}} z_L = Dz_L(w) \neq \bot$). □

F. Computation of Clarke Subderivatives

F.1. Lemmas (Basic)

Definition F.1. Let $A \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$ (where A does not need to contain x). For $B \subseteq \mathbb{R}^n$, we say that A *has B-directions around x* if for all $b \in B$, there is $\delta > 0$ such that $\{x + tb \mid t \in (0, \delta)\} \subseteq A$. We say that A *has sufficient directions around x* if A has B-directions around x for some $B \subseteq \mathbb{R}^n$ with $span(B) = \mathbb{R}^n$, where $span(B) \triangleq \{\sum_{i=1}^{k} t_i b_i \mid k \in \mathbb{N}, t_i \in \mathbb{R}, b_i \in B\}$ denotes the span of B.

Lemma F.2. *Let $A \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$.*

1. *If $x \in int(A)$, then A has $\mathbb{R}^n$-directions around x.*
2. *Let $\alpha \in \{\pm 1\}$, $\epsilon \in \mathbb{R}_{>0} \cup \{\infty\}$, and $f : \mathbb{R}^n \to \mathbb{R}$. If f is differentiable at x and*

$$A = \{y \in \mathbb{R}^n \mid \alpha \cdot (f(y) - f(x)) \in (0, \epsilon)\},$$

then A has B-directions around x for

$$B \triangleq \{y \in \mathbb{R}^n \mid \alpha \cdot (Df(x) \cdot y) \in (0, \infty)\}.$$

3. *Let $A', B \subseteq \mathbb{R}^n$. If A has B-directions around x and $A \subseteq A'$, then A' has B-directions around x.*
4. *Let $A', B, B' \subseteq \mathbb{R}^n$. If A has B-directions around x and A' has B'-directions around x, then $(A \cap A')$ has $(B \cap B')$-directions around x.*
5. *If A has B-directions around x for some nonempty, open $B \subseteq \mathbb{R}^n$, then A has sufficient directions around x.*

Proof. The proofs of (1), (3), and (4) are straightforward, so we omit them.

Proof of (2). Consider the setup stated above. Assume that f is differentiable at x, and let $b \in B$. We want to show there is $\delta > 0$ such that $\{x + tb \mid t \in (0, \delta)\} \subseteq A$. We show this when $\alpha = 1$; we omit the case when $\alpha = -1$, as the proof is similar. Observe that since f is differentiable at x, there is $\delta' > 0$ such that for all $h \in \mathbb{R}^n$,

$$0 < \|h\| < \delta' \implies \frac{|f(x+h) - f(x) - Df(x) \cdot h|}{\|h\|} < \frac{Df(x) \cdot b}{2\|b\|}, \tag{34}$$

where $\|\cdot\|$ denotes the ℓ_2-norm. Here we used $Df(x) \cdot b > 0$ and $\|b\| > 0$, which hold by $b \in B$ and the definition of B.

We claim that $\{x + tb \mid t \in (0, \delta)\} \subseteq A$ holds for the following choice of δ:

$$\delta \triangleq \min\left\{\frac{\delta'}{\|b\|}, \frac{2\epsilon}{3(Df(x) \cdot b)}\right\} > 0.$$

To show this, consider any $t \in (0, \delta)$. It suffices to show $x + tb \in A$. Observe that for $h = tb$, we have $0 < \|h\| = \|tb\| < \delta\|b\| \leq (\delta'/\|b\|) \cdot \|b\| = \delta'$. Hence, Eq. (34) implies that

$$\begin{gathered}\big|f(x+tb) - f(x) - Df(x) \cdot (tb)\big| < \|tb\| \cdot \tfrac{1}{2\|b\|}(Df(x) \cdot b) = \tfrac{1}{2}(Df(x) \cdot b)t, \\ 0 < \tfrac{1}{2}(Df(x) \cdot b)t < f(x+tb) - f(x) < \tfrac{3}{2}(Df(x) \cdot b)t < \epsilon,\end{gathered}$$

where the second line uses $Df(x) \cdot b > 0$ and $t < \delta \leq \frac{2}{3}\epsilon/(Df(x) \cdot b)$. From this, and by the definition of A, we have $x + tb \in A$ as desired.

Proof of (5). This follows from the fact that the span of any nonempty, open set in $\mathbb{R}^n$ is $\mathbb{R}^n$. □

Lemma F.3. *Let $f, g : \mathbb{R}^n \to \mathbb{R}^m$, $A \subseteq \mathbb{R}^n$, and $x \in \mathbb{R}^n$. Suppose that $f = g$ on $A \cup \{x\}$, A has sufficient directions around x, and f and g are differentiable at x. Then,*

$$Df(x) = Dg(x).$$

Proof. Consider the setup stated above. Since A has sufficient directions around x, there is $B \subseteq \mathbb{R}^n$ such that A has B-directions around x and $span(B) = \mathbb{R}^n$. We claim that $Df(x) \cdot b = Dg(x) \cdot b$ for all $b \in B$. Note that this claim implies the conclusion: by the claim and $span(B) = \mathbb{R}^n$, we have $Df(x) \cdot v = Dg(x) \cdot v$ for all $v \in \mathbb{R}^n$, and so $Df(x) = Dg(x)$.

We now prove the above claim. Let $b \in B$. Note that it suffices to show:

$$\|(Df(x) - Dg(x)) \cdot b\| < \|b\|\epsilon \qquad \text{for all } \epsilon > 0,$$

since this implies $(Df(x) - Dg(x)) \cdot b = 0$, where $\|\cdot\|$ denotes the ℓ_2-norm. Let $\epsilon > 0$. Since f and g are differentiable at x, there is $\delta > 0$ such that for any $t \in (0, \delta)$,

$$\frac{\|f(x+tb) - f(x) - Df(x) \cdot (tb)\|}{\|tb\|} < \frac{\epsilon}{2}, \qquad \frac{\|g(x+tb) - g(x) - Dg(x) \cdot (tb)\|}{\|tb\|} < \frac{\epsilon}{2}. \tag{35}$$

Also, since $b \in B$, there is $\delta' > 0$ such that $\{x + tb \mid t \in (0, \delta')\} \subseteq A$. Fix $t \triangleq \min\{\delta, \delta'\}/2 > 0$. Then, we obtain the desired equation based on this t:

$$\begin{aligned}
\big\|\big(Df(x) - Dg(x)\big) \cdot b\big\| &= \frac{1}{t}\big\|\big(Df(x) - Dg(x)\big) \cdot (tb)\big\| \\
&= \frac{1}{t}\big\|\big(f(x+tb) - f(x) - Df(x) \cdot (tb)\big) - \big(f(x+tb) - f(x) - Dg(x) \cdot (tb)\big)\big\| \\
&= \frac{1}{t}\big\|\big(f(x+tb) - f(x) - Df(x) \cdot (tb)\big) - \big(g(x+tb) - g(x) - Dg(x) \cdot (tb)\big)\big\| \\
&= \frac{1}{t}\Big(\big\|f(x+tb) - f(x) - Df(x) \cdot (tb)\big\| + \big\|g(x+tb) - g(x) - Dg(x) \cdot (tb)\big\|\Big) \\
&= \frac{1}{t} \cdot \Big(\frac{\epsilon}{2} + \frac{\epsilon}{2}\Big)\|tb\| = \|b\|\epsilon,
\end{aligned}$$

where the third line uses that $f = g$ on $A \cup \{x\}$ (by assumption) and $x + tb \in A$ (by $t < \delta'$), and the last line uses Eq. (35) (by $t < \delta$). □

Lemma F.4. *Let $n \in \mathbb{N}$, $\{d_i \in \mathbb{N}\}_{i \in [n]}$ such that $d_1 < \cdots < d_n$, and $\{f_i : \mathbb{R}^{d_i - 1} \to \mathbb{R}\}_{i \in [n]}$. Then, for any $c \in \mathbb{R}^n$, there is $u \in \mathbb{R}^{d_n}$ such that*

$$f_i(u_1, \ldots, u_{d_i - 1}) + u_{d_i} = c_i \qquad \textit{for all } i \in [n].$$

Proof. The proof proceeds by induction on n.

Case $n = 1$. For any $c \in \mathbb{R}$, $u \triangleq (0, \ldots, 0, c - f_1(0, \ldots, 0)) \in \mathbb{R}^{d_1}$ satisfies the desired equation.

Case $n > 1$. Let $c \in \mathbb{R}^n$. By induction hypothesis on $n - 1$, there is $v \in \mathbb{R}^{d_{n-1}}$ such that $f_i(v_1, \ldots, v_{d_i - 1}) + v_{d_i} = c_i$ for all $i \in [n - 1]$. Define $u \triangleq (v, 0, \ldots, 0, c_n - f_n(v, 0, \ldots, 0)) \in \mathbb{R}^{d_n}$. Then, u satisfies the desired equations, since $(u_1, \ldots, u_{d_{n-1}}) = v$ by $d_{n-1} < d_n$. □

F.2. Lemmas (Technical)

In the following subsections, we consider a piecewise-C^1 (not piecewise-differentiable) representation of each $\sigma_{l,i}$, using the same notation in the previous sections. Formally, we make the following definitions.

Definition F.5. For each $(l, i) \in \mathsf{Idx}$, let

$$\{(\mathcal{I}^k_{l,i}, \sigma^k_{l,i})\}_{k \in [K_{l,i}]}$$

be a piecewise-C^1 representation of $\sigma_{l,i} : \mathbb{R} \to \mathbb{R}$ that defines $D^{\mathtt{AD}}\sigma_{l,i}$, where $K_{l,i} \in \mathbb{N}$, $\mathcal{I}^k_{l,i} \subseteq \mathbb{R}$, and $\sigma^k_{l,i} : \mathbb{R} \to \mathbb{R}$. We assume that the representation satisfies:

$$\bigcup_{k \in [K_{l,i}]} bd(\mathcal{I}^k_{l,i}) = \bigcup_{k \in [K_{l,i}]} pbd(\mathcal{I}^k_{l,i}) = \mathsf{ncdf}(\sigma_{l,i}),$$

where $\mathsf{ncdf}(f) \subseteq \mathbb{R}$ denotes the set of real numbers at which $f : \mathbb{R} \to \mathbb{R}$ is not continuously differentiable. If $D^{\mathtt{AD}}\sigma_{l,i}$ is consistent, we further assume that the representation satisfies the following:

$$int(\mathcal{I}^k_{l,i}) \neq \emptyset \text{ for all } k \in [K_{l,i}].$$

Note that such a representation always exists by Theorem A.4. Based on these new representations $\{(\mathcal{I}^k_{l,i}, \sigma^k_{l,i})\}_{k \in [K_{l,i}]}$, we define Γ, $\mathcal{R}^\gamma$, y^γ_l, z^γ_l, and σ^γ_l for $\gamma \in \Gamma$ and $l \in [L]$, as we defined them in §A.2; we omit their definitions here. □

Since we consider a piecewise-C^1 (not piecewise-differentiable) representation of $\sigma_{l,i}$, we have Lemma F.6 (shown below) that is stronger than Lemma A.9. Moreover, Lemmas A.8 and A.10–A.12 continue to hold under the new representations; the proofs are exactly the same as before, so we omit them.

Lemma F.6. *For all $l \in [L]$ and $\gamma \in \Gamma$, y_l and z_l are continuous, and y_l^γ and z_l^γ are C^1.*

Proof. The continuity of y_l and z_l follows directly from that $\tau_{l'}$, $\pi_{l'}$, and $\sigma_{l',i'}$ are continuous for all $(l', i') \in \mathsf{Idx}$. Similarly, the continuous differentiability of y_l^γ and z_l^γ follows directly from that $\tau_{l'}$, $\pi_{l'}$, and $\sigma_{l',i'}^{k'}$ are C^1 for all $(l', i') \in \mathsf{Idx}$ and $k' \in [K_{l',i'}]$. □

F.3. Theorems 3.6 and 4.7 (Main Lemmas)

Lemma F.7. *Let $\gamma \in \Gamma$ and $w \in \mathcal{R}^\gamma$. Suppose that for all $l \in [L]$, if τ_l does not have bias parameters, then $y_{l,i}(w) \notin pbd(\mathcal{I}_{l,i}^{\gamma(l,i)})$ for all $i \in [N_l]$. Also, assume that $D^{\mathsf{AD}}\sigma_{l,i}$ is consistent for all $(l, i) \in \mathsf{Idx}$. Then,*

$$int(\mathcal{R}^\gamma) \text{ has sufficient directions around } w.$$

Proof. First, observe that

$$\begin{aligned}
int(\mathcal{R}^\gamma) &= int\Big(\bigcap_{(l,i)\in\mathsf{Idx}} \big\{w' \in \mathbb{R}^W \mid y_{l,i}(w') \in \mathcal{I}_{l,i}^{\gamma(l,i)}\big\}\Big) \\
&= int\Big(\bigcap_{(l,i)\in\mathsf{Idx}} \big\{w' \in \mathbb{R}^W \mid y_{l,i}^\gamma(w') \in \mathcal{I}_{l,i}^{\gamma(l,i)}\big\}\Big) \\
&= \bigcap_{(l,i)\in\mathsf{Idx}} int\Big(\big\{w' \in \mathbb{R}^W \mid y_{l,i}^\gamma(w') \in \mathcal{I}_{l,i}^{\gamma(l,i)}\big\}\Big) \\
&\supseteq \bigcap_{(l,i)\in\mathsf{Idx}} A_{l,i} \qquad \text{for } A_{l,i} \triangleq \big\{w' \in \mathbb{R}^W \mid y_{l,i}^\gamma(w') \in int(\mathcal{I}_{l,i}^{\gamma(l,i)})\big\},
\end{aligned}$$

where the second line uses Lemma A.10, the third line uses that $int(U \cap V) = int(U) \cap int(V)$ for any $U, V \subseteq \mathbb{R}^n$, and the fourth line uses that $int(f^{-1}(U)) \supseteq f^{-1}(int(U))$ for any $U \subseteq \mathbb{R}^m$ and continuous $f : \mathbb{R}^n \to \mathbb{R}^m$. Note that $A_{l,i}$ is open, since $int(\mathcal{I}_{l,i}^{\gamma(l,i)})$ is open and $y_{l,i}^\gamma$ is continuous (by Lemma F.6).

Next, we show that it suffices to find some $B_{l,i} \subseteq \mathbb{R}^W$ for every $(l, i) \in \mathsf{Idx}$ such that

(i) $A_{l,i}$ has $B_{l,i}$-directions around w, and
(ii) $\bigcap_{(l,i)\in\mathsf{Idx}} B_{l,i}$ is nonempty and open.

Suppose that there are such $B_{l,i}$'s. By applying Lemma F.2-(4) to (i), we have

$$\textstyle\bigcap_{(l,i)\in\mathsf{Idx}} A_{l,i} \text{ has } \bigcap_{(l,i)\in\mathsf{Idx}} B_{l,i}\text{-directions around } w.$$

By applying Lemma F.2-(3) to the above and $\bigcap_{(l,i)\in\mathsf{Idx}} A_{l,i} \subseteq int(\mathcal{R}^\gamma)$, we have

$$int(\mathcal{R}^\gamma) \text{ has } \textstyle\bigcap_{(l,i)\in\mathsf{Idx}} B_{l,i}\text{-directions around } w.$$

By applying Lemma F.2-(5) to the above and (ii), we obtain the desired conclusion:

$$int(\mathcal{R}^\gamma) \text{ has sufficient directions around } w.$$

What remains is to show that there is $B_{l,i}$ satisfying (i) and (ii). We claim that the $B_{l,i}$ defined below satisfies (i) and (ii):

$$B_{l,i} = \begin{cases} \mathbb{R}^W & \text{if } w \in A_{l,i} \\ \{v \in \mathbb{R}^W \mid \alpha_{l,i} \cdot (Dy_{l,i}^\gamma(w) \cdot v) \in (0, \infty)\} & \text{if } w \notin A_{l,i}, \end{cases}$$

where $\alpha_{l,i} \in \{\pm 1\}$ is defined as

$$\alpha_{l,i} = \begin{cases} 1 & \text{if } w \notin A_{l,i} \text{ and } y_{l,i}^\gamma(w) = \inf \mathcal{I}_{l,i}^{\gamma(l,i)} \\ -1 & \text{if } w \notin A_{l,i} \text{ and } y_{l,i}^\gamma(w) = \sup \mathcal{I}_{l,i}^{\gamma(l,i)}. \end{cases}$$

Before proving (i) and (ii), we point out that $B_{l,i}$ is well-defined. In particular, $Dy^{\gamma}_{l,i}(w)$ exists since $y^{\gamma}_{l,i}$ is differentiable (by Lemma F.6); and $\alpha_{l,i}$ is well-defined (i.e., the cases in the definition of $\alpha_{l,i}$ covers all possible cases) since $w \notin A_{l,i}$ implies

$$y^{\gamma}_{l,i}(w) \in pbd(\mathcal{I}^{\gamma(l,i)}_{l,i}) = \{\inf \mathcal{I}^{\gamma(l,i)}_{l,i}, \sup \mathcal{I}^{\gamma(l,i)}_{l,i}\}. \tag{36}$$

Here the equality comes from that $\mathcal{I}^{\gamma(l,i)}_{l,i}$ is an interval in $\mathbb{R}$, and the inclusion comes from:

$$y^{\gamma}_{l,i}(w) = y_{l,i}(w), \qquad y^{\gamma}_{l,i}(w) \notin int(\mathcal{I}^{\gamma(l,i)}_{l,i}), \qquad y_{l,i}(w) \in \mathcal{I}^{\gamma(l,i)}_{l,i}, \tag{37}$$

where the first equation is by Lemma A.11 and $w \in \mathcal{R}^{\gamma}$, the second equation by $w \notin A_{l,i}$, and the third equation by $w \in \mathcal{R}^{\gamma}$.

We now prove that the $B_{l,i}$ defined above satisfies (i) and (ii).

Proof of (i). Consider $(l,i) \in \mathsf{Idx}$. If $w \in A_{l,i}$, then $A_{l,i}$ has $\mathbb{R}^W$-directions around w by Lemma F.2-(1), since $w \in int(A_{l,i}) = A_{l,i}$ (as $A_{l,i}$ is open); hence, (i) holds for this case. For the other case, suppose that $w \notin A_{l,i}$. Let $\epsilon_{l,i} \in \mathbb{R} \cup \{\infty\}$ be the length of the interval $\mathcal{I}^{\gamma(l,i)}_{l,i}$. Then,

$$\epsilon_{l,i} > 0, \qquad A_{l,i} = \{v \in \mathbb{R}^W \mid \alpha_{l,i} \cdot (y^{\gamma}_{l,i}(v) - y^{\gamma}_{l,i}(w)) \in (0, \epsilon_{l,i})\}.$$

Here the former holds, since we have $int(\mathcal{I}^{\gamma(l,i)}_{l,i}) \neq \emptyset$ (by Definition F.5) and that $D^{\mathsf{AD}}\sigma_{l,i}$ is consistent (by assumption). The latter holds, since $int(\mathcal{I}^{\gamma(l,i)}_{l,i})$ is either $(y^{\gamma}_{l,i}(w), y^{\gamma}_{l,i}(w) + \epsilon_{l,i})$ or $(y^{\gamma}_{l,i}(w) - \epsilon, y^{\gamma}_{l,i}(w))$ by $w \notin A_{l,i}$ and Eq. (36). By these two observations, and since $y^{\gamma}_{l,i}$ is differentiable, Lemma F.2-(2) is applicable to $(A_{l,i}, B_{l,i}, w)$ and directly implies (i).

Proof of (ii). First, $\bigcap_{(l,i)\in\mathsf{Idx}} B_{l,i}$ is open as desired, since every $B_{l,i}$ is open and Idx is finite. Second, we show that $\bigcap_{(l,i)\in\mathsf{Idx}} B_{l,i}$ is nonempty. Let $\mathsf{Idx}' \triangleq \{(l,i) \in \mathsf{Idx} \mid w \notin A_{l,i}\}$. By the definition of $B_{l,i}$, what we want to show is that for some $v' \in \mathbb{R}^W$,

$$\alpha_{l,i} \cdot (Dy^{\gamma}_{l,i}(w) \cdot v') > 0 \qquad \text{for all } (l,i) \in \mathsf{Idx}'.$$

Since $\alpha_{l,i} \neq 0$ for all $(l,i) \in \mathsf{Idx}'$, it suffices to show that for some $v' \in \mathbb{R}^W$,

$$Dy^{\gamma}_{l,i}(w) \cdot v' = \alpha_{l,i} \qquad \text{for all } (l,i) \in \mathsf{Idx}'. \tag{38}$$

To prove this, we analyze the above equation as follows. Consider any $(l,i) \in \mathsf{Idx}'$. Then, we have $w \notin A_{l,i}$, which implies $y_{l,i}(w) \in pbd(\mathcal{I}^{\gamma(l,i)}_{l,i})$ by Eqs. (36) and (37). From this, τ_l has bias parameters (by assumption). So, for all $v = (v_1, \ldots, v_W) \in \mathbb{R}^W$,

$$\begin{aligned} y^{\gamma}_{l,i}(v) &= \tau_{l,i}\big(z^{\gamma}_{l-1}(v), \pi_l(v)\big) \\ &= \tau_{l,i}\big(z^{\gamma}_{l-1}(v_1, \ldots, v_{W'}, 0, \ldots, 0), (v_{W'+1}, \ldots, v_{W'+W_l})\big) \\ &= \tau'_{l,i}\big(z^{\gamma}_{l-1}(v_1, \ldots, v_{W'}, 0, \ldots, 0), (v_{W'+1}, \ldots, v_{W'+(W_l-N_l)}\big) + v_{W'+(W_l-N_l+i)}, \end{aligned} \tag{39}$$

where the second line uses $W' \triangleq W_1 + \cdots + W_{l-1}$ and the fact that z^{γ}_{l-1} depends only on the parameters of $\tau_1, \ldots, \tau_{l-1}$, and the third line uses that τ_l has bias parameters. Let $\psi_{l,i} \triangleq W' + (W_l - N_l + i)$. Since the first term in Eq. (39) does not depend on $v_{\psi_{l,i}}, \ldots, v_W$, the following holds for all $j \geq \psi_{l,i}$:

$$\big(Dy^{\gamma}_{l,i}(w)\big)_j = \begin{cases} 1 & \text{if } j = \psi_{l,i} \\ 0 & \text{if } j > \psi_{l,i}. \end{cases}$$

From this, the following holds for all $v \in \mathbb{R}^W$:

$$Dy^{\gamma}_{l,i}(w) \cdot v = \sum_{j\in[W]} \big(Dy^{\gamma}_{l,i}(w)\big)_j \cdot v_j = f_{l,i}(v_1, \ldots, v_{\psi_{l,i}-1}) + v_{\psi_{l,i}},$$

where $f_{l,i} : \mathbb{R}^{\psi_{l,i}-1} \to \mathbb{R}$ is defined as $f_{l,i}(u) \triangleq \sum_{j\in[\psi_{l,i}-1]} \big(Dy^{\gamma}_{l,i}(w)\big)_j \cdot u_j$. Hence, what we planned to show (i.e., Eq. (38) holds for some $v' \in \mathbb{R}^W$) is equivalent to the following: for some $v' \in \mathbb{R}^W$,

$$f_{l,i}(v'_1, \ldots, v'_{\psi_{l,i}-1}) + v'_{\psi_{l,i}} = \alpha_{l,i} \qquad \text{for all } (l,i) \in \mathsf{Idx}'. \tag{40}$$

Since $\psi_{l,i} \neq \psi_{l',i'}$ for any $(l,i) \neq (l',i')$, Lemma F.4 implies that there is $v' \in \mathbb{R}^W$ satisfying Eq. (40). This proves (ii), and concludes the proof. □

Lemma F.8. *Let $\gamma \in \Gamma$ and $w \in \mathcal{R}^\gamma$. Suppose that $int(\mathcal{R}^\gamma)$ has sufficient directions around w. Then,*

$$D^{\mathtt{AD}} z_L(w) = \begin{cases} Dz_L(w) & \text{if } Dz_L(w) \neq \bot \\ \lim_{n\to\infty} Dz_L(w'_n) \text{ for some } w'_n \to w & \text{if } Dz_L(w) = \bot. \end{cases}$$

Proof. Let $\gamma \in \Gamma$ and $w \in \mathcal{R}^\gamma$ such that $int(\mathcal{R}^\gamma)$ has sufficient directions around w. By Lemmas A.11 and A.12,

$$z_L(w') = z_L^\gamma(w') \quad \wedge \quad D^{\mathtt{AD}} z_L(w') = Dz_L^\gamma(w') \qquad \text{for all } w' \in \mathcal{R}^\gamma. \tag{41}$$

We prove the conclusion for each of the two cases: $Dz_L(w) \neq \bot$ and $Dz_L(w) = \bot$.

Case 1: $Dz_L(w) \neq \bot$ (i.e., z_L is differentiable at w). We want to show

$$D^{\mathtt{AD}} z_L(w) = Dz_L(w).$$

This holds as follows:

$$D^{\mathtt{AD}} z_L(w) = Dz_L^\gamma(w) = Dz_L(w),$$

where the first equality is by Eq. (41), and the second equality follows directly from Lemma F.3 applied to $(z_L^\gamma, z_L, \mathcal{R}^\gamma, w)$. Here Lemma F.3 is applicable since its preconditions are satisfied: z_L^γ is differentiable at w (by Lemma F.6); z_L is differentiable at w (by assumption); $z_L^\gamma = z_L$ on $int(\mathcal{R}^\gamma) \cup \{w\}$ (by Eq. (41)); and $int(\mathcal{R}^\gamma)$ has sufficient directions around w (by assumption).

Case 2: $Dz_L(w) = \bot$ (i.e., z_L is not differentiable at w). We want to show:

$$D^{\mathtt{AD}} z_L(w) = \lim_{n\to\infty} Dz_L(w'_n) \qquad \text{for some } w'_n \to w. \tag{42}$$

Since $int(\mathcal{R}^\gamma)$ has sufficient directions around w (by assumption), there is $\{w'_n \in int(\mathcal{R}^\gamma)\}_{n\in\mathbb{N}}$ such that $w'_n \to w$. We show that these w'_n satisfy Eq. (42) as follows:

$$D^{\mathtt{AD}} z_L(w) = Dz_L^\gamma(w) = \lim_{n\to\infty} Dz_L^\gamma(w'_n) = \lim_{n\to\infty} Dz_L(w'_n),$$

where the first equality is by Eq. (41), the second equality uses that Dz_L^γ is continuous (by Lemma F.6), and the third equality uses that $Dz_L^\gamma(w'_n) = Dz_L(w'_n)$ for all n (since $w'_n \in int(\mathcal{R}^\gamma)$ and $z_L^\gamma = z_L$ on $\mathcal{R}^\gamma$ by Eq. (41)). This concludes the proof. □

F.4. Theorems 3.6 and 4.7 (Main Proofs)

Theorem 3.6. *If z_L has bias parameters and $D^{\mathtt{AD}}\sigma_{l,i}$ is consistent for all $(l,i) \in \mathsf{Idx}$, then for all $w \in \mathbb{R}^W$,*

$$D^{\mathtt{AD}} z_L(w) = \begin{cases} Dz_L(w) & \text{if } Dz_L(w) \neq \bot \\ \lim_{n\to\infty} Dz_L(w'_n) \text{ for some } w'_n \to w & \text{if } Dz_L(w) = \bot. \end{cases}$$

This implies that $D^{\mathtt{AD}} z_L$ is a Clarke subderivative of z_L.

Proof. This theorem is a special case of Theorem 4.7; we omit the proof. □

Theorem 4.7. *Let $w \in \mathbb{R}^W$ and assume that $D^{\mathtt{AD}}\sigma_{l,i}$ is consistent for all $(l,i) \in \mathsf{Idx}$. If $y_{l,i}(w) \notin \mathsf{ncdf}(\sigma_{l,i})$ for all $(l,i) \in \mathsf{Idx}$ such that τ_l does not have bias parameters, then*

$$D^{\mathtt{AD}} z_L(w) = \begin{cases} Dz_L(w) & \text{if } Dz_L(w) \neq \bot \\ \lim_{n\to\infty} Dz_L(w'_n) & \\ \quad \text{for some } w'_n \to w & \text{if } Dz_L(w) = \bot \end{cases}$$

and so $D^{\mathtt{AD}} z_L(w)$ is a Clarke subderivative of z_L at w.

Proof. Let $w \in \mathbb{R}^W$ that satisfies the assumption in the statement. By Lemma A.8, there is $\gamma \in \Gamma$ such that $w \in \mathcal{R}^\gamma$. Note that Lemma F.7 is applicable to (γ, w) because: $D^{\mathtt{AD}}\sigma_{l,i}$ is consistent for all $(l,i) \in \mathsf{Idx}$ (by assumption); and for all $l \in [L]$, if τ_l does not have bias parameters, then $y_{l,i}(w) \notin \mathsf{ncdf}(\sigma_{l,i})$ and so $y_{l,i}(w) \notin pbd(\mathcal{I}_{l,i}^{\gamma(l,i)})$ for all $i \in [N_l]$, where the former follows from the assumption and the latter from $pbd(\mathcal{I}_{l,i}^{\gamma(l,i)}) \subseteq \mathsf{ncdf}(\sigma_{l,i})$ (by Definition F.5). Hence, Lemma F.7 implies that $int(\mathcal{R}^\gamma)$ has sufficient directions around w, which subsequently implies the conclusion by Lemma F.8. □