Journal of Machine Learning Research 26 (2025) 1-31 Submitted 5/23; Published 5/25

Four Axiomatic Characterizations of the Integrated Gradients Attribution Method

Daniel Lundstrom DAN.DAVE.LUNDSTROM@GMAIL.COM
Department of Mathematics
University of Southern California
Los Angeles, CA 90007, USA

Meisam Razaviyayn RAZAVIYA@USC.EDU
Departments of Industrial and Systems Eng., Electrical Eng., and Computer Science
University of Southern California
Los Angeles, CA 90007, USA

Editor: Pradeep Ravikumar

Abstract

Deep neural networks have produced significant progress among machine learning models in terms of accuracy and functionality, but their inner workings are still largely unknown. Attribution methods seek to shine a light on these "black box" models by indicating how much each input contributed to a model's outputs. The Integrated Gradients (IG) method is a state of the art baseline attribution method in the axiomatic vein, meaning it is designed to conform to particular principles of attributions. We present four axiomatic characterizations of IG, establishing IG as the unique method satisfying four different sets of axioms.

Keywords: Machine Learning Explainability, Attribution Methods, Axiomatic Approach, Integrated Gradients, Path Methods

1. Introduction

Deep neural networks have revolutionized various fields of machine learning over the past decade, from computer vision to natural language processing. These models are often left unexplained, causing practitioners difficulties when troubleshooting training inference issues or poor performance. This can lead to a lack of user trust in the model and an inability to understand what features are important to a model's function. Various regulations have been proposed that would require that ML models be transparent in certain scenarios (House, 2022), (Commission, 2021), (Dorries, 2022).

Attribution methods, sometimes called salience maps in the context on computer vision, are a response to this issue, purporting to explain the working of a model by indicating which inputs are important to a model's output. One group of such methods, game-theoretic attribution methods, go about producing attributions in a principled way by stipulating axioms, or guiding principals, and proposing methods that conform to those principles. When axioms are posited, the possible forms of an attribution become constrained, possibly to a single, unique method.

This is the case with the Integrated Gradients method. Initially introduced and analyzed in Axiomatic Attributions for Neural Networks (Sundararajan et al., 2017), counterexamples to its uniqueness claims have since been provided by Lundstrom et al. (2022) and Lerma and Lucas (2021).

While the original uniqueness claim about IG is problematic, in this work, we show that IG uniqueness claims can be established rigorously via different axioms. We start by introducing different axioms common to game-theoretic attribution methods, namely, implementation invariance, linearity, dummy, and completeness. Then, using axioms, we establish the following characterizations.

1. Path methods can be characterized among attribution methods by the linearity, completeness, dummy, and non-decreasing positivity axioms.

2. IG can be characterized among monotone path methods by the symmetry-preserving and affine scale invariance axioms.

3. IG can be characterized among attribution methods by the linearity, affine scale invariance, completeness, non-decreasing positivity, and proportionality axioms.

4. IG can be characterized among attribution methods by the linearity, completeness, dummy, and symmetric-monotonicity axioms.

5. IG can be characterized among attribution methods by its action on monomials and the continuity of Taylor approximations for analytic functions axiom.

Furthermore, we show that IG attributions to neural networks with ReLU and max functions coincide with IG attributions to softplus approximations to such models. This establishes a sort of continuity of IG among softplus approximations.

2. Background

Many solutions have been proposed to help explain black box neural networks. Using the taxonomy of Linardatos et al. (2020), we can divide types of explainability methods into various overlapping categories. One approach is to make models intrinsically explainable (Chen et al., 2019)(Letham et al., 2015), while another method allows a user to choose model architecture and train the model then explains the model afterwards, called post-hoc explanations (Lundberg and Lee, 2017). Some methods are designed to be used for particular data type such as images (Smilkov et al., 2017) or language (Ventura et al., 2021). Some methods are designed for use on specific types of models (Vig, 2019), while others are model agnostic (Ribeiro et al., 2016). Some methods explain the models workings over an entire data set (Ibrahim et al., 2019), while others explain the model's actions with regard to a particular input (Zeiler and Fergus, 2014).

Attributions methods are post-hoc methods designed to explain a model's action on a specific input. One method, SHAP is a popular method based on the Shapley Value (Lundberg and Lee, 2017). Another, LIME, locally approximates a black-box model by a more interpretable model and uses that surrogate to explain the original model (Ribeiro et al., 2016). DeepLIFT back-propagates contributions through internal layers of a neural network by comparing a neuron activation to its reference value (Shrikumar et al., 2017). in Layer-Wise Relevance Propagation relevance scores are conservatively propagated backwards to the input (Binder et al., 2016). Grad-CAM averages gradient maps of intermediate layers to highlight important regions of the input space (Selvaraju et al., 2017). SmoothGrad takes the expected gradient of the model with respect to the input with added noise perturbing the input (Smilkov et al., 2017). The requirements of the above methods vary: some

methods require output gradients, some are designed for use on deep convolutional neural networks, others require gradients of internal neurons, others require full knowledge of model architecture and weights.

A particular subgroup of attributions apply methods from game-theoretic cost-sharing. These methods borrow from a developed set of literature which provides them with a strong theoretical background and established results. The previously mentioned SHAP method (Lundberg and Lee, 2017) is an import of the Shapley value cost-sharing method (Shapley and Shubik, 1971) into the ML attributions context. Likewise, the Integrated Gradient (Sundararajan et al., 2017) is an import of the Aumann-Shapley cost-sharing method (Aumann and Shapley, 1974) into the ML attributions context.

Various works have analyzed the Aumann-Shapley method, characterizing it as the unique method to satisfy a set of desirable properties, or axioms. Billera and Heath (1982) gave a characterization based on the idea of proportionality, while Mirman and Tauman (1982) and Samet and Tauman (1982) gave further characterization in a similar vein. McLean et al. (2004) showed a characterization based on the ideas of potential and consistency, while Calvo and Santos (2000) characterized the Aumann-Shapley method based on the idea of balanced contributions. Sprumont (2005) developed constraints around the merging or splitting agents to provide a characterization. Young (1985) provided a characterization using the principle of symmetric monotonicity, Monderer and Neyman (1988) developed another characterization based on potential, while Albizuri et al. (2014) developed a characterization based on both merging/splitting and monotonicity.

The Integrated Gradients was first introduced in Sundararajan et al. (2017) and a characterization was provided for it as well. This claim did not cite any characterizations of the Aumann Shapley, but used the idea of preserving symmetry. However, Lerma and Lucas (2021) and Lundstrom et al. (2022) critiqued various aspects of the uniqueness claim with counterexamples and issues with the proof methods. The issues cited in Lundstrom et al. (2022)'s criticism is that the ML context is significantly different than the cost-sharing context, causing unforeseen difficulties in applying results from one to another. Another characterization of IG was provide in Sundararajan and Najmi (2020), this time based on a cost-sharing result relying on the principle of proportionality. This proof was also criticized by Lundstrom et al. (2022), for the same reasons.

Multiple attribution methods were inspired by the integrated gradients, some by extending the method and others by augmenting it. Giving an abridged list: Erion et al. (2021) extended IG by taking the expectation of IG over various baseline values; Xu et al. (2020) augmented IG by using not a straight path, but one defined by deblurring the input image; Dhamdhere et al. (2018) augmented the IG integrand to derive a method of attributing to internal neurons, and Lundstrom et al. (2022) extended this method to attribution to image patches; and Pascal Sturmfels (2020) explored the effects of different baseline choices for IG.

3. Preliminaries

In this section, we cover preliminaries needed for our work.

3.1 Baseline Attribution Notations

We begin by establishing preliminary notions. For a, $b \in \mathbb{R}^n$, let $[a, b]$ denote the hyper-rectangle with opposite vertices a and b. Here $[a, b]$ represents the domain of input of a ML

model, such as a colored image. We denote the set of ML models of interest $\mathcal{F}$, with $F \in \mathcal{F}$ being some function $F : [a, b] \to \mathbb{R}$, e.g. a deep learning model. Here we only consider one output of a model, so that if a model reports a probability vector of scores from a softmax layer, for instance, we only consider one entry of the probability vector.

Throughout the paper x represents a general function input, $\bar{x}$ represents a particular model input who's components are to be attributed, and x' denotes a reference baseline input. A *baseline attribution method* (BAM) explains a model by assigning scores to the components of an input indicating its contribution to the output $F(\bar{x})$. We define a BAM as:

Definition 1 (Baseline Attribution Method) *Given an input $\bar{x} \in [a, b]$ and baseline $x' \in [a, b]$, $F \in \mathcal{F}(a, b)$, a baseline attribution method is any function of the form $A : D \to \mathbb{R}^n$, where $D \subseteq [a, b] \times [a, b] \times \mathcal{F}$.*

A BAM reports a vector, so that $A_i(\bar{x}, x', F)$ reports the contribution of the i^{th} component of $\bar{x}$ to the output $F(\bar{x})$, given the reference baseline input x'. Our definition of a BAM is broad enough that the attribution need not take into account the baseline value. In this technical sense, attributions that do not utilize a baseline, such as gradients or Smoothgrad (Smilkov et al., 2017) can be formulated as BAMs. In the literature, however, BAMs are a type of attribution that harnesses the baseline as a means to measure the contribution of $\bar{x}$ by comparison x'. Often a baseline x' is implicit for the model F or recommended by attribution practitioners, and we may drop writing x' if it is unnecessary. It is not guaranteed that a BAM is defined for any input, as we will see in section 3.3. When discussing a set of BAMs defined on a general domain, we may use the general notation D, while we use D_A to indicate the domain of a particular BAM A.

There are two particular BAM's defined on different function classes we will discuss. Define $\mathcal{F}^1(a, b)$ to be the set of real analytic functions on $[a, b]$, and define $\mathcal{A}^1$ to be the set of BAMs defined on $[a, b] \times [a, b] \times \mathcal{F}^1(a, b)$. We may write $\mathcal{F}^1$ if a, b is apparent.

The class of real analytic functions is well understood, but does not include many practical deep NNs, such as those which use the ReLU and max functions. To address these networks, define $\mathcal{F}^2(a, b)$, or $\mathcal{F}^2$ if a, b is apparent, to be the set of feed-forward neural networks with a finite number of nodes on $[a, b]$ composed of real-analytic layers and ReLU layers. This includes fully connected, skip, residual, max, and softmax layers, as well as activation functions like sigmoid, mish, swish, softplus, and leaky ReLU.

Formally, let $n_0, ..., n_m \in \mathbb{N}$, and for $1 \leq k \leq m$, let $F^k : \mathbb{R}^{n_{k-1}} \to \mathbb{R}^{n_k}$ denote a real-analytic function. Let $S^k : \mathbb{R}^{n_k} \to \mathbb{R}^{n_k}$ to be any function of the form $S^k(x) = (f_1^k(x_1), ..., f_{n_k}^k(x_n))$, where $f_i^k(x_k)$ is the identity mapping or the ReLU function. That is, S^k performs one of either a pass through or a ReLU on each component, and could perform different operations on different components. Each function in $\mathcal{F}^2$ takes the form:

$$F(x) = S^m \circ F^m \circ S^{m-1} \circ F^{m-1} \circ ... \circ S^2 \circ F^2 \circ S^1 \circ F^1(x),$$

where $\circ$ denotes function composition. Note that a multi-input max function can be formulated by a series of two-input max functions, and $\max(x, y) = \text{ReLU}(x - y) + y$. Thus neural networks with the max function can be reformulated using only the ReLU function, and $\mathcal{F}^2$ includes neural networks with the max function. Define $\mathcal{A}^2(D)$ (or $\mathcal{A}^2$) to be the set of BAMs defined on $D \subseteq [a, b] \times [a, b] \times (\mathcal{F}^1 \cup \mathcal{F}^2)$.

3.2 Axiomatic Approach

The previous definition of a BAM is very broad, and includes many BAMs that do not track the importance of inputs. The axiomatic approach to attribution methods is to stipulate properties that can be imposed on A, limiting its structure and ensuring it accurately tracks feature contribution. It is even possible that a set of axioms constrains attribution methods to the degree that only one method satisfies all of them. In this case, the set of axioms would characterize the attribution method. We move to review axioms common to the literature.

Our first axiom, *implementation invariance* (Sundararajan et al., 2017), can be stated as follows:

1. *Implementation Invariance*: A is not a function of model implementation, but solely a function of the mathematical mapping of the model's domain to the range.

This axiom stipulates that an attribution method be independent of the model's implementation. Otherwise, the values of the attribution may carry information about implementation aspects such as architecture. Many methods, such as Smoothgrad (Smilkov et al., 2017) and SHAP (Lundberg and Lee, 2017), satisfy implementation invarinace while Sundararajan et al. (2017) showed that DeepLIFT (Shrikumar et al., 2017) and Layer-Wise Relevance Propogation (Binder et al., 2016) do not satisfy it.

The next axiom, *linearity* (Sundararajan et al., 2017) (Sundararajan and Najmi, 2020) (Janizek et al., 2021), is given as,

2. *Linearity*: If $(\bar{x}, x', F)$, $(\bar{x}, x', G) \in D_A$, $\alpha, \beta \in \mathbb{R}$, then $(\bar{x}, x', \alpha F + \beta G) \in D_A$ and $A(\bar{x}, x', \alpha F + \beta G) = \alpha A(\bar{x}, x', F) + \beta A(\bar{x}, x', G)$.

The linearity axiom ensures that if F is a linear combination of other models, a weighted average of model outputs for example, then the attributions of F equals the average of the attributions to the sub-models. This imposes structure to the attributions outputs, so that if a model's outputs are scaled to give outputs twice as large for example, then the attributions are scaled as well.

We say that a function F does not vary in an input x_i if for every x in the domain of F, $G(t) := F(x_1, ..., x_{i-1}, t, x_{i+1}, ..., x_m)$ is a constant function. We denote that F does not vary in x_i by writing $\partial_i F \equiv 0$. With this definition we may state another axiom, *dummy*[1],

3. *Dummy*: If $(\bar{x}, x', F) \in D_A$ and $\partial_i F \equiv 0$, then $A_i(\bar{x}, x', F) = 0$.

Dummy ensures that whenever an input has no effect on the function, the attribution score is zero.

Another axiom, *completeness* (Sundararajan et al., 2017) (Sundararajan and Najmi, 2020) (Tsai et al., 2022), is given as,

4. *Completeness*: If $(\bar{x}, x', F) \in D_A$, then $\sum_{i=1}^{n} A_i(\bar{x}, x', F) = F(\bar{x}) - F(x')$.

Completeness grounds the meaning of the magnitude and sign of attributions. The magnitude of $A_i(\bar{x}, x', F)$ indicates that $\bar{x}_i$ contributed that quantity to the change in function value from $F(x')$ to $F(\bar{x})$. The sign of $A_i(\bar{x}, x', F)$ indicates whether $\bar{x}_i$ contributed to function increase or function decrease. Thus the attributions to each input give a complete account of function change, $F(\bar{x}) - F(x')$.

1. The dummy axiom here is called Sensitivity(b) in Sundararajan et al. (2017).

3.3 The Integrated Gradients

There is a particular form of baseline attribution method which satisfies axioms 1-4, called a path method. Define a path function as follows:

Definition 2 (Path Function) *A function $\gamma(\bar{x}, x', t) : [a,b] \times [a,b] \times [0,1] \to [a,b]$ is a path function if, for fixed $\bar{x}, x'$, $\gamma(t) := \gamma(\bar{x}, x', t)$ is a continuous piecewise smooth curve from x' to $\bar{x}$.*

We may drop both $\bar{x}$, x' when they are understood and write $\gamma(t)$. If we further suppose that $\frac{\partial F}{\partial x_i}(\gamma(t))$ exists almost everywhere[2], then the *path method* associated with γ can be defined as:

Definition 3 (Path Method) *Given the path function $\gamma(\cdot,\cdot,\cdot)$, the i^{th} component of the corresponding path method is defined as*

$$A_i^{\gamma}(\bar{x}, x', F) = \int_0^1 \frac{\partial F}{\partial x_i}(\gamma(\bar{x}, x', t)) \times \frac{\partial \gamma_i}{\partial t}(\bar{x}, x', t) dt, \tag{1}$$

where γ_i denotes the i^{th} entry of γ.

Path methods are well defined when ∇F exists and is continuous on $[a,b]$, however, this is not necessarily the case for common ML models, such as neural networks that use ReLU and max functions. For example, if $\bar{x} = (1,1)$, $x' = (0,0)$, we use could the straight line path $\gamma(t) = t(1,1)$. If $F(x) = \max(x_1, x_2)$, then $A^{\gamma}(\bar{x}, x', F)$ does not exist because the partial derivatives are undefined on every point on the path γ.

The *Integrated Gradients method* (Sundararajan et al., 2017) is the path method defined by the straight path from x' to $\bar{x}$, given as $\gamma(\bar{x}, x', t) = x' + t(\bar{x} - x')$, and takes the form:

Definition 4 (Integrated Gradients Method) *Given a function F and baseline x', the Integrated Gradients attribution of the i^{th} component of $\bar{x}$ is defined as*

$$IG_i(\bar{x}, x', F) = (\bar{x}_i - x'_i) \int_0^1 \frac{\partial F}{\partial x_i}(x' + t(\bar{x} - x'))dt \tag{2}$$

The Integrated Gradients method is the application of the Aumann Shapley cost-sharing method to the ML attributions context (Aumann and Shapley, 1974). We define $D_{\text{IG}} \subseteq [a,b] \times [a,b] \times (\mathcal{F}^1 \cup \mathcal{F}^2)$ to be the domain where IG is defined.

The Integrated Gradient method satisfies the four axioms stated above. We give a brief explanation of how it satisfies each. Assume here that $\gamma(t)$ is the straight line IG path.

- Implementation Invariance: IG only depends on ∇F, which is independent on the implementation of F.
- Linearity: For any index i we have:

$$\begin{aligned}
\text{IG}_i(\bar{x}, x', aF + bG) &= (\bar{x}_i - x'_i) \int_0^1 \frac{\partial (aF + bG)}{\partial x_i}(\gamma(t))dt \\
&= a(\bar{x}_i - x'_i) \int_0^1 \frac{\partial F}{\partial x_i}(\gamma(t))dt + b(\bar{x}_i - x'_i) \int_0^1 \frac{\partial G}{\partial x_i}(\gamma(t))dt \\
&= a\text{IG}_i(\bar{x}, x', G) + b\text{IG}_i(\bar{x}, x', G)
\end{aligned}$$

2. $\frac{\partial F}{\partial x_i}(\gamma(t))$ exists almost everywhere iff the Lebesgue measure of $\{t \in [0,1] : \frac{\partial F}{\partial x_i}(\gamma(t))$ does not exist.$\}$ is 0.

- Dummy: If $\partial_i F \equiv 0$ then $\mathrm{IG}_i(\bar{x}, x', F)$ integrates the zero function, and equals 0.

- Completeness: Letting " $\cdot$ " denote the inner product, we employ the fundamental theorem of line integrals to gain:

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{IG}_i(\bar{x}, x', F) &= \sum_{i=1}^{n} (\bar{x}_i - x'_i) \int_0^1 \frac{\partial F}{\partial x_i}(\gamma(t))dt \\ &= \int_0^1 \nabla F(\gamma(t)) \cdot \gamma'(t)dt \\ &= F(\bar{x}) - F(x')\end{aligned}$$

4. Path Method Characterization with NDP and Symmetry-Preserving

The first attempt at characterizing the Integrated Gradients was presented in Sundararajan et al. (2017). The general flow of the argument was 1) path methods uniquely satisfy a set of axioms, and 2) IG is the unique path method that satisfies certain extra properties. Later, Lundstrom et al. (2022) critiqued the first part of the argument with provided counterexamples, while Lerma and Lucas (2021) critiques and provided counterexamples to the second argument and adjusted some results. Here we present the current understanding of the argument and establish a characterization of IG among path methods by adding affine scale invariance, an axiom already seen in the literature.

4.1 Characterizing Ensembles of Monotone Path Methods

A path function $\gamma(t)$ from x' to $\bar{x}$ is called monotone if each component is monotone in t. We denote the set of monotone paths from x' to $\bar{x}$ by $\Gamma^m(\bar{x}, x')$. We say that F is non-decreasing from x' to $\bar{x}$ if $F(\gamma(t))$ is monotonically increasing in t for each $\gamma \in \Gamma^m(\bar{x}, x')$. We then define the axiom *Non-Decreasing Positivity* (NDP) as follows:

5. *Non-Decreasing Positivity*: If $(\bar{x}, x', F) \in D_A$ and F is non-decreasing from x' to $\bar{x}$ then $A(\bar{x}, x', F) \geq 0$.

If F is non-decreasing from x' to $\bar{x}$, each input of F does not cause a decrease if it moves closer to the input from the baseline. Thus, intuitively no component of $\bar{x}$ contributed to F decreasing by being at its input value rather than the baseline value. Because no input contributed to F decreasing in value, NDP asserts that those attributions should not be negative.

With NDP, we can characterize a sort of averaging of monotone path methods among all baseline attribution methods.[3]

Theorem 5 *(Lundstrom et al., 2022, Theorem 2) Suppose $A \in \mathcal{A}^1$. The following are equivalent:*

i. A satisfies completeness, linearity, dummy, and NDP.

3. For an account of the differences between theorem 5 here and proposition 2 in Sundararajan et al. (2017), see Lundstrom et al. (2022).

ii. There exists a family of probability measures $\mu^{\cdot,\cdot}$ indexed on $(\bar{x}, x') \in [a,b] \times [a,b]$, where $\mu^{\bar{x},x'}$ is a measure on $\Gamma^m(\bar{x}, x')$, such that

$$A(\bar{x}, x', F) = \int_{\Gamma^m(\bar{x},x')} A^\gamma(\bar{x}, x', F) d\mu^{\bar{x},x'}(\gamma)$$

Theorem 5 states that if $A \in \mathcal{A}^1$ is constrained according to the four axioms, then A is an expected value of path methods with monotone paths. We call this expected value of path methods an *ensemble of monotone path methods*, or more generally an ensemble of path methods if the expectation is not constrained to monotone paths. To present results for $\mathcal{F}^2$, we first give a result on the topology of NN models in $\mathcal{F}^2$:

Lemma 6 *(Lundstrom et al., 2022, Lemma 2) Suppose $F \in \mathcal{F}^2$. Then $[\bar{x}, x']$ can be partitioned into a nonempty region U and its boundary ∂U, where F is real-analytic on U, U is open with respect to the (usual) topology of the dimension of $[\bar{x}, x']$, and ∂U is measure 0.*

We now present a claim extending theorem 5 to functions in $\mathcal{F}^2$. Let U denote the set as described above in Lemma 6, and denote the set of points on the path γ by P^γ.

Theorem 7 *(Lundstrom et al., 2022, Theorem 3) Suppose $A \in \mathcal{A}^2$ is defined on $[a,b] \times [a,b] \times \mathcal{F}^1$ and some subset of $[a,b] \times [a,b] \times \mathcal{F}^2$, and satisfies completeness, linearity, dummy, and NDP. Let $\mu^{\cdot,\cdot}$ be the family of measures on monotone paths that defines A on $[a,b] \times [a,b] \times \mathcal{F}^1$ from Theorem 5, and let $(\bar{x}, x', F) \in [a,b] \times [a,b] \times \mathcal{F}^2$. If $A(\bar{x}, x', F)$ is defined, and for almost every path $\gamma \in \Gamma^m(\bar{x}, x')$ (according to $\mu^{\bar{x},x'}$), $\{t \in [0,1] : \gamma(t) \in \partial U\}$ is a null set w.r.t the Lebesgue measure on $\mathbb{R}$, then $A(\bar{x}, x', F)$ is equivalent to an ensemble of monotone path methods. Furthermore, this ensemble is defined with the same $\mu^{\cdot,\cdot}$ as Theorem 5.*

The above result answers two questions: 1) is A an ensemble of path methods when evaluating models in $\mathcal{F}^2$, and 2) is that ensemble the same ensemble that A uses to evaluate models in $\mathcal{F}^1$? The above theorem guarantees that when considering models in $\mathcal{F}^2$ which may not be differentiable on $[a,b]$, A is still an ensemble of path methods, and, in fact, is the same ensemble that define's A's action on models in $\mathcal{F}^1$. Thus Theorem 7 establishes that while ensembles of path methods uniquely satisfy a set of axioms for attributions in $\mathcal{A}^1$, they also satisfy these axioms for models in $\mathcal{F}^2$ when defined.

4.2 Characterizing IG Among Monotone Path Methods

Among ensembles of monotone path methods, another popular method exists: the Shapley value (Shapley and Shubik, 1971), (Lundberg and Lee, 2017). The Shapley value is obtained by considering average change in function value when a component's value is changed from x'_i to $\bar{x}_i$. Specifically, consider all possible ways that x' can transition to $\bar{x}$ by sequentially toggling each component from x'_i to $\bar{x}_i$. The Shapley value for $\bar{x}_i$ is the average change in function value over all possible transitions via toggling. This method can be formulated as an ensemble of $n!$ path methods. With speedups, calculating the Shapley value precisely is exponential in the number of inputs, and significant effort has been put into faster calculation via approximation (Chen et al., 2023). The Shapley value was criticized as potentially problematic compared to IG in the original IG paper (Sundararajan et al., 2017)[Remark 5].

As an alternative to this approach, the most computationally efficient ensemble would be an ensemble composed of a single-path method. It can be shown that IG is the unique path method that satisfies a couple of axioms.

The first axiom, *symmetry-preserving*, is given as:

6. *Symmetry-Preserving*: For a vector x and indices $1 \leq i,\ j \leq n$, define x^* by swapping the values of x_i and x_j. Now suppose that $\forall x \in [a, b]$, $F(x) = F(x^*)$. Then if $(\bar{x}, x', F) \in D_A$, $\bar{x}_i = \bar{x}_j$ and $x'_i = x'_j$, we have $A_i(\bar{x}, x', F) = A_j(\bar{x}, x', F)$.

Symmetry-preserving requires "swappable" features with identical values to give identical attributions. This axiom was introduced in Sundararajan et al. (2017), but was criticized as insufficient to characterize IG among path methods in Lerma and Lucas (2021). In short, other path methods exist that take the straight line path as IG does when $x'_i = x'_j$, $\bar{x}_i = \bar{x}_j$, but deviate otherwise. These counter-examples exist because symmetry-preserving only makes requirements when $x'_i = x'_j$, $\bar{x}_i = \bar{x}_j$. To remedy this, we considered strengthening the symmetry axioms, but found it insufficient to characterize IG among path methods. See Appendix A for details.

The second axiom, *Affine Scale Invaraince*, is given as:

7. *Affine Scale Invariance (ASI)*: Let T be an arbitrary affine transformation of a single index, so that $T(x) := (x_1, ..., cx_i + d, ..., x_n)$ for constants $c \neq 0$, some index i. Then whenever $\bar{x}, x', T(\bar{x}), T(x') \in [a, b]$, we have $A(\bar{x}, x', F) = A(T(\bar{x}), T(x'), F \circ T^{-1})$.

This axiom can be justified by considering unit conversion. Suppose F is some machine learning model where input $\bar{x}_i$ is given in degrees Fahrenheit. T could be an affine transformation that converts the i^{th} input from Fahrenheit to Celcius, so that $F \circ T^{-1}$ is an adjusted model where $\bar{x}_i$ would be given in Celsius, converted to Fahrenheit, then input into the original model. Affine scale invariance would require that an attribution method A give the same attributions whether in Fahrenheit inputs, $(\bar{x}, x', F)$, or Celcius inputs, $(T(\bar{x}), T(x'), F \circ T^{-1})$. Note that the ASI definition implies a parallel property for affine transformations on multiple inputs by applying a sequence of affine transforms for single inputs.

It is interesting to note that ASI effectively means that the shape of a path for a path method stays the same regardless of the input or baseline values. Explicitly, suppose A^γ is a path method satisfying ASI. For any $\bar{x}$, x' such that $\forall i\ \bar{x}_i \neq x'_i$, there exists a unique affine transformation T such that $T(x') = 0$, $T(\bar{x}) = 1$, where by 0 and 1 we mean the vectors with entries that are all zero or one, respectively. Thus $A^\gamma(\bar{x}, x', F) = A^\gamma(T(\bar{x}), T(x'), F \circ T^{-1}) = A^\gamma(1, 0, F \circ T^{-1})$. The final expression uses the path $\gamma(1, 0, t)$, and ignores the form of the path for $\gamma(\bar{x}, x', t)$. This causes the path to keep the same shape, so that all paths are an affine stretching of the base path from $x' = 0$ to $\bar{x} = 1$.

With symmetry-preserving and ASI, IG can be characterized among monotone path methods:

Theorem 8 *(Symmetry-Preserving Path Method Characterization on $\mathcal{A}^2$) If $A \in \mathcal{A}^2(D_{IG})$ is a monotone path method satisfying ASI and symmetry-preserving, then it is the Integrated Gradients method.*

The proof of Theorem 8 is relegated to Appendix B.

5. Characterizing IG with ASI and Proportionality

A second attempt at characterizing the Integrated Gradients was presented in The Many Shapley Values for Model Explanation paper (Sundararajan et al., 2017), which was also critiqued by Lundstrom et al. (2022) later. Here we present the characterization.

The axiom of *proportionality* states,

8. *Proportionality*: If $0 \in [a, b]$ and there exists $G : [a, b] \to \mathbb{R}$ such that for all $x \in [a, b]$, $F(x) = G(\sum_i x_i)$, then there exists $c \in \mathbb{R}$ such that $A_i(\bar{x}, 0, F) = c\bar{x}_i$ for $1 \leq i \leq n$.

This axiom states that if F can be expressed as a function of the cumulative quantity, $\sum_i x_i$, then each attribution is proportional to its contribution to $\sum_i \bar{x}_i$, namely $\bar{x}_i$. This axiom originates from the context of cost-sharing (Friedman and Moulin, 1999), where each $\bar{x}_i$ may represent an investment. As an example, if the return on investment, $F(\bar{x})$, is a function of the cumulative dollars invested, $\sum_i \bar{x}_i$, then proportionality asserts that the payout to each investor should be proportional to the amount invested. This principle does not always apply in cost-sharing problems, as when different investors make different kinds of contributions to an investment. This principle is fitting, however, when all investments are of the same kind so that the payout is simply a function of the total investment. Admittedly, this axiom appears at first glance to be more sensible in the cost-sharing context than the ML attributions context, and depends on the application of interest.

With proportionality and ASI, we can characterize IG:

Theorem 9 *(Proportionality Characterization on $\mathcal{A}^2$) Suppose that $A \in \mathcal{A}^2(D_{IG})$. Then the following are equivalent:*

i. A satisfies linearity, ASI, completeness, NDP, and proportionality.

ii. A is the Integrated Gradients method.

The proof of Theorem 9 is deferred to Appendix C. Note that unlike theorem 8, which characterized IG among monotone path methods, this is a much broader characterization, establishing that of all BAMs in $\mathcal{A}^2$, only IG satisfies the given axioms.

6. Characterizing IG with Symmetric Monotonicity

We next present a characterization of IG employing the concept of *monotonicity*. The axiom of monotonicity can be stated as:

8a. *Monotonicity*: Suppose $F \in \mathcal{F}^1$. Then,

i. If $\bar{x}_i \neq x'_i$, then $\frac{\partial F}{\partial x_i}(x) \leq \frac{\partial G}{\partial x_i}(x)\ \forall x \in [\bar{x}, x']$ implies $\frac{A_i(\bar{x}, x', F)}{\bar{x}_i - x'_i} \leq \frac{A_i(\bar{x}, x', G)}{\bar{x}_i - x'_i}$.

ii. If $\bar{x}_i = x'_i$, then $A_i(\bar{x}, x', F) = 0$.

To explain i., the term $\frac{A_i(\bar{x}, x', F)}{\bar{x}_i - x'_i}$ is the per-unit attribution of $\bar{x}_i$. We take the contribution of $\bar{x}_i$ to the change in F, denoted $A_i(\bar{x}, x', F)$, and divide it by the total change in $\bar{x}_i$. If $\bar{x}_i$ contributed to F increasing, but $\bar{x}_i$ decreased from the baseline, then the per-unit attribution of $\bar{x}_i$ would be negative. As an example, suppose both derivatives are positive and $\bar{x}_i > x'_i$.

If increasing $\bar{x}_i$ causes at least as great an increase for G as it does for F, then according to monotonicity, the per-unit attribution of $\bar{x}_i$ should be at least as great for G as F.

Requirement ii. is the continuous extension of i. to the $\bar{x}_i = x'_i$ case under the assumption of completeness and dummy. To demonstrate this extension, let $F \in \mathcal{F}^1$, so that $c \leq \frac{\partial F}{\partial x_i} \leq d$ on the bounded domain $[a, b]$. Then, by completeness and dummy, $A_i(\bar{x}, x', cx_i) = c(\bar{x}_i - x'_i)$ and $A_i(\bar{x}, x', dx_i) = d(\bar{x}_i - x'_i)$. Thus $c = \frac{A_i(\bar{x}, x', cx_i)}{\bar{x}_i - x'_i} \leq \frac{A_i(\bar{x}, x', F)}{\bar{x}_i - x'_i} \leq \frac{A_i(\bar{x}, x', dx_i)}{\bar{x}_i - x'_i} = d$. Now, as $\bar{x}_i \to x'_i$, we have $A_i(\bar{x}, x', F) \to 0$.

With the idea of monotonicity, one can assert a similar principle to the comparison between different inputs with the axioms, *symmetric monotonicity*:

8b. *Symmetric Monotonicity*: Suppose $A \in \mathcal{A}^1$, F, $G \in \mathcal{F}^1$. Then:

i. If $\bar{x}_i \neq x'_i$ and $\bar{x}_j \neq x'_j$, then $\frac{\partial F}{\partial x_i}(x) \leq \frac{\partial G}{\partial x_j}(x)$ $\forall x \in [\bar{x}, x']$ implies $\frac{A_i(\bar{x}, x', F)}{\bar{x}_i - x'_i} \leq \frac{A_j(\bar{x}, x', G)}{\bar{x}_j - x'_j}$.

ii. If $\bar{x}_i = x'_i$, then $A_i(\bar{x}, x', F) = 0$.

Symmetric monotonicity enforces that the principle of monotonicity can be applied between different inputs. With symmetric monotonicity, we give the following characterization of IG among methods in $\mathcal{A}^1$:

Theorem 10 *(Symmetric Monotonicity Characterization on $\mathcal{A}^1$) Suppose that $A \in \mathcal{A}^1$. Then the following are equivalent:*

i. A satisfies completeness, dummy, linearity, and symmetric monotonicity.

ii. A is the Integrated Gradients method.

The proof of Theorem 10 is located in Appendix D.

To extend the results to $\mathcal{A}^2$, we consider two options. The first is to include NDP, and the second is to include a version of symmetric monotonicity that is formulated for functions that may not be differentiable. To do this, we replace the condition $\frac{\partial F}{\partial x_i}(x) \leq \frac{\partial G}{\partial x_j}(x)$ $\forall x \in [\bar{x}, x']$ with a condition applicable to non-differentiable functions.

Supposing F, $G \in \mathcal{F}^2$, we define the statement $\frac{\partial F}{\partial x_i}(x) \leq \frac{\partial G}{\partial x_j}(x)$ *locally approximately* to mean: $\forall x \exists \epsilon > 0$ such that if $|z| < \epsilon$ then $\frac{F(x_1, \ldots, x_i + z, \ldots, x_n) - F(x)}{z} \leq \frac{G(x_1, \ldots, x_j + z, \ldots, x_n) - G(x)}{z}$ whenever both terms exists. The above statement indicates we have something akin to $\frac{\partial F}{\partial x_i}(x) \leq \frac{\partial G}{\partial x_j}(x)$, using local secant approximations of the derivative. We now state $\mathcal{C}^0$-*symmetric monotonicity*, an adjustment to symmetric monotonicity for BAMs in $\mathcal{A}^2$:

8c. $\mathcal{C}^0$-*Symmetric Monotonicity*: Suppose $A \in \mathcal{A}^2(D_{\text{IG}})$, $(\bar{x}, x', F)$, $(\bar{x}, x', G) \in D_{\text{IG}}$. Then:

i. If $\bar{x}_i \neq x'_i$ and $\bar{x}_j \neq x'_j$, then $\frac{\partial F}{\partial x_i}(x) \leq \frac{\partial G}{\partial x_j}(x)$ locally approximately $\forall x \in [\bar{x}, x']$ implies $\frac{A_i(\bar{x}, x', F)}{\bar{x}_i - x'_i} \leq \frac{A_j(\bar{x}, x', G)}{\bar{x}_j - x'_j}$.

ii. If $\bar{x}_i = x'_i$, then $A_i(\bar{x}, x', F) = 0$.

We now extend the characterization of Theorem 10 for attributions in $\mathcal{A}^2$.

Theorem 11 *(Symmetric Monotonicity Characterization on $\mathcal{A}^2$) Suppose that $A \in \mathcal{A}^2(D_{IG})$. Then the following are equivalent:*

i. A satisfies completeness, dummy, linearity, symmetric monotonicity, and NDP.

ii. A satisfies completeness, dummy, linearity, and $\mathcal{C}^0$-symmetric monotonicity.

iii. A is the Integrated Gradients method.

The proof of Theorem 11 is located in Appendix E.

7. Characterization by the Attribution to Monomials

Another means of characterizing attribution methods is to begin with a principle of attributing to simple functions.[4] First, for $m \in \mathbb{N}_0^n$, define $[x]^m := x_1^{m_1} \cdots x_n^{m_n}$. Given a set baseline x' and $m \in \mathbb{N}_0^n$, we employ a slight abuse of terminology and define a monomial to be any function of the form $F(x) = [x - x']^m$. Now, consider a simple example function we would like to perform attribution on: $F(x_1, x_2) = (x_1 - x_1')^{100}(x_2 - x_2')$. The function F evaluated at $\bar{x} = (x_1' + 2, x_2' + 2)$ yields $F(\bar{x}) = 2^{100}2^1 = 2^{101}$. Now, considering methods that satisfy completeness, the attribution question is: how to distribute $F(\bar{x}) - F(x') = 2^{101}$ between x_1 and x_2?

One possibility is to consider x_1 and x_2 equal contributors, so that $A(\bar{x}, x', F) = (\frac{2^{101}}{2}, \frac{2^{101}}{2})$. This is in fact, what the Shapley value attribution. For any monomial $F(x) = [x - x']^m$, the Shapley value gives attributions equally to each input such that $m_i \neq 0$. This seems a naive attribution, given the structure of F.

Another means of attributing to the inputs of F is to consider the magnitude of m_i, the power of $\bar{x}_i$. Particularly, we could attribute to $\bar{x}_i$ proportionally to the number of times it is multiplied when evaluating $F(\bar{x})$. An attribution following this guideline would yield: $A((x_1' + 2, x_2' + 2), x', (x_1 - x_1')^{100}(x_2 - x_2')) = (\frac{100}{101}2^{101}, \frac{1}{101}2^{101})$, a result that appears equitable. In fact, this attribution coincides with the attribution of IG. For $m \in \mathbb{N}_0^n$ such that $m_i \neq 0$, we have

$$\begin{aligned}
&\mathrm{IG}_i(\bar{x}, x', [x - x']^m) \\
=&(\bar{x}_i - x_i') \int_0^1 \frac{\partial([x - x']^m)}{\partial x_i}(x' + t(\bar{x} - x'))dt \\
=&(\bar{x}_i - x_i') \int_0^1 m_i(t(\bar{x}_1 - x_1'))^{m_1} \cdots (t(\bar{x}_i - x_i'))^{m_i - 1} \cdots (t(\bar{x}_n - x_n'))^{m_n} dt \\
=&(\bar{x}_i - x_i') \int_0^1 m_i t^{\|m\|_1 - 1}(\bar{x}_1 - x_1')^{m_1} \cdots (\bar{x}_i - x_i')^{m_i - 1} \cdots (\bar{x}_n - x_n')^{m_n} dt \\
=&\frac{m_i}{\|m\|_1}[\bar{x} - x']^m
\end{aligned} \tag{3}$$

We may proceed from attributions on monomials to attributions on $\mathcal{F}^1$ by requiring a sort of continuity criteria. For $m \in \mathbb{N}_0^n$, define $[m]! := m_1! \cdots m_n!$, and define $D^m F = \frac{\partial^{\|m\|_1} F}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}$.

4. This concept was explored in Sundararajan et al. (2020) for an interactions method, and later by Lundstrom and Razaviyayn (2023) for IG gradient-based interactions methods. Here we state results from that paper for $\mathcal{A}^1$, and give a result on continuity into $\mathcal{A}^2$.

Recall that for $F \in \mathcal{F}^1$, the Taylor approximation of order l centered at x', denoted F_l, is given by:

$$T_l(x) = \sum_{m \in \mathbb{N}_0^n, \|m\|_1 \leq l} \frac{D^m(F)(x')}{[m]!}[x - x']^m \tag{4}$$

The Taylor approximation for analytic functions has the property that $D^m T_l$ uniformly converges to $D^m F$ for any $m \in \mathbb{N}_0^n$ and $x \in [a, b]$. Thus, it seems natural to require that any attribution $A \in \mathcal{A}^1$ satisfy $\lim_{l \to \infty} A(\bar{x}, x', T_l) = A(\bar{x}, x', F)$. This is the principle behind the axiom *Continuity of Taylor Approximation for Analytic Functions*, or what we may equivalently call the *continuity condition*, given below:

9 *Continuity of Taylor Approximation for Analytic Functions*: If $A \in \mathcal{A}^1$, $(\bar{x}, x', F) \in [a, b] \times [a, b] \times \mathcal{F}^1$, then $\lim_{l \to \infty} A(\bar{x}, x', T_l) = A(\bar{x}, x', F)$, where T_l is the l^{th} order Taylor approximation of F centered at x'.

We now give the characterization of IG according to its actions on monomials:

Theorem 12 *(Distribution of Monomials Characterization on $\mathcal{A}^1$)(Lundstrom and Razaviyayn, 2023, Corollary 3) Suppose $A \in \mathcal{A}^1$. Then the following are equivalent:*

i. A satisfies continuity of Taylor approximation for analytic functions and acts on monomials as:

$$A(\bar{x}, x', [x - x']^m) = \frac{m}{\|m\|_1} \times [\bar{x} - x']^m$$

ii. A is the Integrated Gradients method.

We may proceed from $\mathcal{F}^1$ to $\mathcal{F}^2$ by considering a means of approximating a feed-forward neural network by an analytic function. Suppose $F \in \mathcal{F}^2$ is a feed-forward neural network with ReLU and Max functions. Note that the multi-input max function can be formulated as a series of dual input max functions, and the dual input max function can be formulated as $\max(a, b) = \text{ReLU}(a - b) + b$. Thus we may formulate F using only the ReLU function. We may then define F_α to be the analytic approximation of F given by replacing all instances of ReLU in F with the parameterized softplus, $s_\alpha(z) = \frac{\ln(1+\exp(\alpha z))}{\alpha}$. We show in Appendix F that this softplus approximation uniformly converges to the function F.

Before we give our result, we first give a technical theorem on the topology of $[a, b]$ with respect softplus approximations of functions in $\mathcal{F}^2$. Let ∇F denote the gradient of F, and let λ denote the Lebesgue measure on $\mathbb{R}^n$. Then the following can be said about the topology of the domain of $F \in \mathcal{F}^2$:

Theorem 13 *For any $F \in \mathcal{F}^2$, there exists an open set $U \subseteq [a, b]$ such that $\lambda(U) = \lambda([a, b])$ and for each $x \in U$, the following hold:*

- *There exists an open set containing x, B_x, and real analytic function on $[a, b]$, H_x, such that $F \equiv H_x$ on B_x.*
- *$\nabla F(x)$ exists.*
- *$\nabla F_\alpha(x) \to \nabla F(x)$ as $\alpha \to \infty$.*

With this theorem, we give a result on IG's ability to uniquely extend to models in $\mathcal{F}^2$.

Corollary 14 *Let $(\bar{x}, x', F) \in \mathcal{F}^2$, and let U be the set as in Theorem 13. Let $\gamma(t) = x' + t(\bar{x} - x')$ and suppose $\lambda(\{t \in [0,1] : \gamma(t) \in U\}) = 1$. Then:*

$$\lim_{\alpha \to \infty} IG(\bar{x}, x', F_\alpha) = IG(\bar{x}, x', F)$$

Proofs of Theorem 13 and Corollary 14 are located in Appendices G and H, respectively.

8. Conclusion

We present a table of results, summarizing various characterizations of the Integrated Gradients method found in this paper.

Assumptions	Theorems 7 + 8	Theorem 9	Theorem 11 i.	Theorem 11 ii.	Theorem 12
Linearity (3.2)	x	x	x	x	x
Completeness (3.2)	x	x	x	x	-
Dummy (3.2)	x	-	x	x	-
NDP (4.1)	x	x	x	-	-
Path Method (3.3)	x	-	-	-	-
Symmetry-Preserving (4.2)	x	-	-	-	-
ASI (4.2)	x	x	-	-	-
Proportionality (5)	-	x	-	-	-
Symmetric Monotonicity (6)	-	-	x	x	-
Distribution of Monomials (7)	-	-	-	-	x
Continuity Condition (7)	-	-	-	-	x

Table 1: Each axiom and the section it is located in are listed under the "Assumptions" column. Under each column with a "Theorem" heading, the the set of axioms that characterized IG are marked. All results characterize IG among attributions in $\mathcal{A}^2$ except for Theorem 12, which characterized IG among attributions in $\mathcal{A}^1$. Also, note that Theorem 11 i. assumes Symmetric monotonicity for functions in $\mathcal{F}^1$, while Theorem 11 ii. assumes $\mathcal{C}^0$-symmetric monotonicity.

The axiomatic approach to attributions provides benefits, among which are: 1) identifying shortcomings or benefits of existing methods depending on whether they satisfy certain axioms, 2) providing guiding principles for the development of attribution methods, and 3) identifying methods that uniquely satisfy a set of desirable properties. The presented characterizations of the IG extend our knowledge of baseline attribution methods and lead to a definite and singular method that satisfies certain desirable properties. The community should consider these characterizations and their merit.

Even with these characterizations, admittedly it is unlikely that there is a single best attribution method. In cost-sharing, it has been established that no one method can satisfy all desirable properties,[5] and it is possible that this is the case for attribution methods

5. See Friedman and Moulin (1999, Lemma 4).

as well. Theorems 5 & 7 demonstrate that ensembles of path methods uniquely satisfy common and broadly used axioms. However, the community should consider the less common axioms that help characterize IG: Theorem 8 - being a symmetry preserving single path method, Theorem 9 - proportionality, Theorems 10 & 11 - symmetric monotonicity, and Theorem 12 - IG's distribution of monomials. In the author's opinion, these more defining axioms indicate contexts where the IG attribution method is preferable. We expect that there are other properties, suitable in other contexts, which may exclude IG and recommend another method.

9. Acknowledgment

This work was partially supported by a gift from the USC-Meta Center for Research and Education in AI and Learning.

References

M Josune Albizuri, H Díez, and A Sarachu. Monotonicity and the aumann–shapley cost-sharing method in the discrete case. *European Journal of Operational Research*, 238(2): 560–565, 2014.

Robert J. Aumann and Lloyd S. Shapley. *Values of Non-Atomic Games*. Princeton University Press, Princeton, NJ, 1974.

Louis J Billera and David C Heath. Allocation of shared costs: A set of axioms yielding a unique procedure. *Mathematics of Operations Research*, 7(1):32–39, 1982.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer, 2016.

Emilio Calvo and Juan Carlos Santos. A value for multichoice games. *Mathematical Social Sciences*, 40(3):341–354, 2000.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, pages 1–12, 2023.

European Commission. Proposal for an artificial intelligence act, 2021. 2021/0106(COD), Article 13, Accessed May 22, 2023.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron? *arXiv preprint arXiv:1805.12233*, 2018.

Nadine Dorries. Establishing a pro-innovation approach to regulating ai, 2022.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, pages 1–12, 2021.

Eric Friedman and Herve Moulin. Three methods to share joint costs or surplus. *Journal of economic Theory*, 87(2):275–312, 1999.

The White House. Blueprint for an ai bill of rights, 2022. Accessed May 22, 2023, Section: Notice and Explanation.

Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 279–287, 2019.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.*, 22:104–1, 2021.

Miguel Lerma and Mirtha Lucas. Symmetry-preserving paths in integrated gradients. *arXiv preprint arXiv:2103.13533*, 2021.

Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350 – 1371, 2015. doi: 10.1214/15-AOAS848. URL `https://doi.org/10.1214/15-AOAS848`.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Daniel Lundstrom and Meisam Razaviyayn. Distributing synergy functions: Unifying game-theoretic interaction methods for machine-learning explainability. *arXiv preprint arXiv:2305.03100*, 2023.

Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pages 14485–14508. PMLR, 2022.

Richard P McLean, Amit Pazgal, and William W Sharkey. Potential, consistency, and cost allocation prices. *Mathematics of Operations Research*, 29(3):602–623, 2004.

Leonard J Mirman and Yair Tauman. Demand compatible equitable cost sharing prices. *Mathematics of Operations Research*, 7(1):40–56, 1982.

Dov Monderer and Abraham Neyman. *Values of smooth nonatomic games: the method of multilinear approximation*. Cambridge University Press, Cambridge, 1988.

Su-In Lee Pascal Sturmfels, Scott Lundberg. Visualizing the impact of feature attribution baselines, 2020. URL `https://distill.pub/2020/attribution-baselines/`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Dov Samet and Yair Tauman. The determination of marginal cost prices under a set of axioms. *Econometrica: Journal of the Econometric Society*, pages 895–909, 1982.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Lloyd S Shapley and Martin Shubik. The assignment game i: The core. *International Journal of game theory*, 1(1):111–130, 1971.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Yves Sprumont. On the discrete version of the aumann–shapley cost-sharing method. *Econometrica*, 73(5):1693–1712, 2005.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pages 9259–9268. PMLR, 2020.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *arXiv preprint arXiv:2203.00870*, 2022.

Francesco Ventura, Salvatore Greco, Daniele Apiletti, and Tania Cerquitelli. Explaining the deep natural language processing by mining textual interpretable features. *arXiv preprint arXiv:2106.06697*, 2021.

Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019.

Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020.

H Peyton Young. Producer incentives in cost allocation. *Econometrica: Journal of the Econometric Society*, pages 757–765, 1985.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Appendix

Appendix A. Symmetry-Preserving Alone is Insufficient to Characterize IG Among Path Methods

Here we provide a counterexample to the claim that IG is the unique path method that satisfies symmetry-preserving. We also give a axiom that is stronger than symmetry-preserving, and show that this axiom is insufficient to characterize IG.

A.1 Another Path Method that Satisfies Symmetry-Preserving

Let $D = [0,1]^2$ and define $\gamma'(\bar{x}, x', t)$ element-wise as follows:

$$\gamma_i(\bar{x}, x', t) = x'_i + (\bar{x}_i - x'_i)t^{(\bar{x}_i - x'_i)^2}$$

Note that γ is monotonic. When $\bar{x}_1 = \bar{x}_2$, $x'_1 = x'_2$, then $\bar{x}_1 - x'_1 = \bar{x}_2 - x'_2$, and $\gamma_1(\bar{x}, x', t) = \gamma_2(\bar{x}, x', t)$. In this case γ is the straight path, and A^γ acts as IG. Thus, A^γ satisfies symmetry preserving. However, when $\bar{x}_1 - x'_1 \neq \bar{x}_2 - x'_2$, then the path γ differs from the straight line path, causing $A^\gamma \neq \text{IG}$.

A.2 Strong Symmetry-Preserving

Here we present an attempt to strengthen symmetry and show that multiple path methods satisfy the strengthened axiom. The axiom, *Strong Symmetry-Preserving*, extends the symmetry-preserving axiom to cases when $\bar{x}_i \neq \bar{x}_j$, $x'_i \neq x'_j$:

1. (Strong Symmetry-Preserving): For a vector x and indices $1 \leq i, j \leq n$, let x^* denote the vector x but with the i^{th} and j^{th} components swapped. Suppose F is symmetric in i and j, meaning that $F(x) = F(x^*)$ for all $x \in [a, b]$. Then $A_i(\bar{x}, x', F) = A_j(\bar{x}^*, x'^*, F)$.

Here we provide a counterexample to the claim that IG is the unique path method that satisfies strong symmetry-preserving.

Let $D = [0,2]^2$, and let F be symmetric in components 1 and 2. We define a path function $\gamma(t)$ that is equivalent to the IG path except when $\bar{x} = (2,1)$, $x' = (1,0)$ or $\bar{y} = \bar{x}^* = (1,2)$, $y' = x'^* = (0,1)$. For $\bar{x}$, x', let $\gamma(t)$ be the path that travels in straight lines along the course: $(1,0) \to (2,0) \to (2,1)$. Now for baseline $y' = (0,1)$ and input $\bar{y} = (1,2)$, let $\gamma(t)$ be the path that travels in straight lines along the course: $(0,1) \to (0,2) \to (1,2)$.

We then have $A_1^\gamma(\bar{x}, x', F) = F(2,0) - F(1,0) = F(0,2) - F(0,1) = A_2^\gamma(\bar{x}^*, x'^*, F)$, and likewise $A_2^\gamma(\bar{x}, x', F) = A_1^\gamma(\bar{x}^*, x'^*, F)$. Thus we have another strong symmetry-preserving path method that is not the IG path.

Appendix B. Proof of Theorem 8

Proof We present an adjusted version of the proof found in (Sundararajan et al., 2017, Theorem 1). Suppose that $A \in \mathcal{A}(D_{\text{IG}})$ is a monotone path method that satisfies symmetry preserving and affine scale invariance, with associated monotone path γ. First we establish a property of A. Suppose $\bar{x}_i = \bar{x}_j = 1$, $x'_i = x'_j = 0$ for some $i \neq j$. We will show that $\gamma_i(\bar{x}, x', t), \gamma_j(\bar{x}, x', t) = t$. We proceed by contradiction and suppose WLOG that there exists a t such that $\gamma_i(\bar{x}, x', t) > \gamma_j(\bar{x}, x', t)$. Let (t_a, t_b) be the maximal open set such that if $t \in (t_a, t_b)$ then $\gamma_i(\bar{x}, x', t) > \gamma_j(\bar{x}, x', t)$. If $\gamma_i(\bar{x}, x; , t_a) > \gamma_j(\bar{x}, x', t_a)$, then (t_a, t_b) is not maximal by continuity of $\gamma(\bar{x}, x', \cdot)$. So $\gamma_i(\bar{x}, x', t_a) = \gamma_j(\bar{x}, x', t_a)$, and by

similar argument $\gamma_i(\bar{x}, x;, t_b) = \gamma_j(\bar{x}, x', t_b)$. Define $g_a = \gamma_i(\bar{x}, x', t_a) = \gamma_j(\bar{x}, x', t_a)$, and $g_b = \gamma_i(\bar{x}, x', t_b) = \gamma_j(\bar{x}, x', t_b)$.

We now move to define a $(\bar{x}, x', F) \in D_{\text{IG}}$ where F is symmetric in x_i and x_j, but A does not give equal attributions to $\bar{x}_i$ and $\bar{x}_j$. Define $F \in \mathcal{F}^2$ by

$$F(x) = \text{ReLu}(x_i x_j - g_a^2) - \text{ReLu}(x_i x_j - g_b^2) + g_a^2$$

F can be written in a case-format as follows:

$$F(x) = \begin{cases} g_a^2 & \text{if } x_i x_j \leq g_a^2 \\ x_i x_j & \text{if } g_a^2 \leq x_i x_j \leq g_b^2 \\ g_b^2 & \text{if } g_b^2 \leq x_i x_j \end{cases} \tag{5}$$

It is easy to verify that $(\bar{x}, x', F) \in D_{\text{IG}}$ and is symmetric. Using the shorthand $\gamma(t) = \gamma(\bar{x}, x', t)$, note that for $t \in [t_a, t_b]$, we have $g_a \leq \gamma_i(t), \gamma_j(t) \leq g_b$ since the path is monotonic. Thus, for $\in [t_a, t_b]$, we have $F(\gamma(t)) = \gamma_i(t)\gamma_j(t)$, while for $t \notin (t_a, t_b)$, $F(\gamma(t))$ is constant. Now observe,

$$\begin{aligned} A_i^{\gamma}(\bar{x}, x', F) &= \int_0^1 \frac{\partial F}{\partial x_i}(\gamma(t)) \frac{d\gamma_i}{dt} dt \\ &= \int_{t_a}^{t_b} \gamma_j(t) \frac{d\gamma_i}{dt} dt \\ &< \int_{t_a}^{t_b} \gamma_i(t) \frac{d\gamma_i}{dt} dt \\ &= \int_{\gamma_i(t_a)}^{\gamma_i(t_b)} u du \\ &= \int_{\gamma_j(t_a)}^{\gamma_j(t_b)} u du \\ &= \int_{t_a}^{t_b} \gamma_j(t) \frac{d\gamma_j}{dt} dt \\ &< \int_{t_a}^{t_b} \gamma_i(t) \frac{d\gamma_j}{dt} dt \\ &= \int_0^1 \frac{\partial F}{\partial x_j}(\gamma(t)) \frac{d\gamma_j}{dt} dt \\ &= A_j^{\gamma}(\bar{x}, x', F) \end{aligned} \tag{6}$$

This is a contradiction, as A is symmetry preserving. Thus if $\bar{x}_i = \bar{x}_j = 1$, $x_i' = x_j' = 0$ for some $i \neq j$ then $\gamma_i(\bar{x}, x', t), \gamma_j(\bar{x}, x', t) = t$.

It is left to show that if $(\bar{x}, x', F) \in D_{\text{IG}}$ then $A(\bar{x}, x', F) = \text{IG}(\bar{x}, x', F)$. Define $I = \{i | \bar{x}_i \neq x_i')\}$ and $I^c = \{i | \bar{x}_i = x_i')\}$. For $i \in I^c$, $\gamma_i(\bar{x}, x', t) = \bar{x}_i = x_i'$, so that $\frac{d\gamma_i}{dt} = 0$ and $A_i(\bar{x}, x', F) = 0 = \text{IG}_i(\bar{x}, x', F)$. It is left to show that for any $i \in I$, $A_i(\bar{x}, x', F) = \text{IG}_i(\bar{x}, x', F)$.

Suppose $|I| = 1$. Note that by the fundamental theorem for line integrals, $\sum_k A_k(\bar{x}, x', F) = F(\bar{x}) - F(x')$ (proof identical to completeness proof for IG). Then for $i \in I$, we have $A_i(\bar{x}, x', F) = \sum_k A_k(\bar{x}, x', F) = F(\bar{x}) - F(x') = \sum_k \text{IG}_k(\bar{x}, x', F) = \text{IG}_i(\bar{x}, x', F)$.

Suppose instead $|I| \geq 2$, and let T be an affine transform such that $T(x') = 0$ and $T_i(\bar{x}) = \mathbf{1}_{i \in I}$, where $\mathbf{1}$ is the indicator function. Particularly, $T_i(x) = \frac{x - x'}{\bar{x} - x'}$ for $i \in I$,

and $T_i(x) = x - x'$ for $i \in I^c$. For $i \in I^c$, we have $\gamma_i(T(\bar{x}), T(x'), t) = T_i(x')$ since γ is a monotone path, and by extension, $T_i^{-1}(\gamma(T(\bar{x}), T(x'), t)) = x'_i = x'_i + t(\bar{x}_i - x'_i)$. For $i \in I$ we have $\gamma_i(T(\bar{x}), T(x'), t) = t$ by the property shown above, and by extension, $T_i^{-1}(\gamma(T(\bar{x}), T(x'), t)) = x'_i + t(\bar{x}_i - x'_i)$.

Thus $T^{-1}(\gamma(T(\bar{x}), T(x'), t)) = x' + t(\bar{x} - x')$. Then, for $i \in I$,

$$\begin{aligned}
A_i^\gamma(\bar{x}, x', F) &= A_i^\gamma(T(\bar{x}), T(x'), F \circ T^{-1}) \\
&= \int_0^1 \frac{\partial (F \circ T^{-1})}{\partial x_i}(\gamma(T(\bar{x}), T(x'), t)) \frac{d(\gamma_i(T(\bar{x}), T(x'), t))}{dt} dt \\
&= \int_0^1 \frac{\partial (F \circ T^{-1})}{\partial x_i}(\gamma(T(\bar{x}), T(x'), t)) \cdot 1 dt \\
&= \int_0^1 \frac{\partial F}{\partial x_i}(T^{-1}(\gamma(T(\bar{x}), T(x'), t))) \frac{\partial (T^{-1})_i}{\partial x_i}(\gamma(T(\bar{x}), T(x'), t)) dt \\
&= \int_0^1 \frac{\partial F}{\partial x_i}(x' + t(\bar{x} - x'))(\bar{x}_i - x'_i) dt \\
&= \mathrm{IG}_i(\bar{x}, x', F)
\end{aligned} \tag{7}$$

■

Appendix C. Proof of Theorem 9

We set out to establish the results for $\mathcal{A}^1$, then move to establish the results for $\mathcal{A}^2$.

C.1 Proof for $A \in \mathcal{A}^1$

Proof Here we present a proof along the lines of that found in Sundararajan and Najmi (2020). This proof was criticized in Lundstrom et al. (2022) and partially rectified, but we present the argument in full.

Suppose $A \in \mathcal{A}^2$.

(ii. $\Rightarrow$ i) IG satisfies linearity and completeness and proportionality because it is a path method. Suppose F is non-decreasing from x' to $\bar{x}$, then $(\bar{x}_i - x'_i)$ and $\frac{\partial F}{\partial x_i}(x'(t(\bar{x} - x'))$ do not have opposite signs, so

$$\mathrm{IG}_i(\bar{x}, x', F) = (\bar{x}_i - x'_i) \int_0^1 \frac{\partial F}{\partial x_i}(x' + t(\bar{x} - x'))dt \geq 0,$$

which shows IG satisfies NDP. Finally, let $F(x) = G(\sum_j x_j)$ and $x' = 0$. Then $\frac{\partial F}{\partial x_i}(\bar{x}) = G'(\sum_j \bar{x}_j)$, and

$$\mathrm{IG}_i(\bar{x}, x', F) = \bar{x}_i \int_0^1 G'(\sum_j \bar{x}_j)dt$$

Note that the integral is equivalent for any i, so we take $c = \int_0^1 G'(\sum_j x_j)dt$ to gain $\mathrm{IG}_i(\bar{x}, x', F) = cx_i$. Thus IG satisfies proportionality, and ii. $\Rightarrow$ i.

(i. $\Rightarrow$ ii.) Now suppose that A satisfies linearity, ASI, completeness, NDP, and proportionality. Let $\mathcal{A}^0$ denote the set of all BAMs such that 1) they are defined on analytic,

non-decreasing functions, 2) they are only defined for $x' = 0$, $\bar{x} \geq 0$, 3) the BAMs give non-negative attributions and 4) the BAMs satisfy completeness. By (Friedman and Moulin, 1999, Theorem 3), the only BAM in $\mathcal{A}^0$ to satisfy proportionality and ASI is the Integrated Gradients method.

Let $A \in \mathcal{A}^1$, and note that if $x' = 0$, $\bar{x} \geq 0$, F non-decreasing, then $A(1, 0, F) \geq 0$ by NDP. A also satisfies completeness by assumption. Thus if we let A' denote A with the requisite restriction of domains, then $A' \in \mathcal{A}^0$. Because A' satisfies ASI and proportionality, $A' = \text{IG}$ on this restricted domain.

Let $x' = 0$, and $\bar{x} = 1$, the vector of all ones. For any $F \in \mathcal{F}^1$, F is Lipschitz on bounded domain, and there exists $c \in \mathbb{R}^n$ such that $c \geq 0$, $F(x) + c^\intercal x$ is non-decreasing. Thus

$$\begin{aligned} A(1, 0, F(x)) &= A(1, 0, F(x) + c^\intercal x - c^\intercal x) \\ &= A(1, 0, F(x) + c^\intercal x) - A(1, 0, c^\intercal x) \\ &= \text{IG}(1, 0, F(x) + c^\intercal x) - \text{IG}(1, 0, c^\intercal x) \\ &= \text{IG}(1, 0, F(x)) \end{aligned}$$

We can then harness ASI as in Eq. 7 to get that $A(\bar{x}, x', F) = \text{IG}(\bar{x}, x', F)$ for any $\bar{x}$, x'. ■

C.2 Proof for $A \in \mathcal{A}^2(D_{\text{IG}})$

Proof Let $A \in \mathcal{A}^2(D_{\text{IG}})$. It is easy to show ii. $\Rightarrow$ i.. We turn to show i. $\Rightarrow$ ii..

Suppose A satisfies linearity, ASI, completeness, NDP, and proportionality. Let $(\bar{x}, x', F) \in D_{\text{IG}}$, and choose a component i. By methods found in the proof of (Lundstrom et al., 2022, Theorem 2), there exists a sequence of functions F_m such that:

- F_m is analytic for all m.
- $\frac{\partial F_m}{\partial x_i} \leq \frac{\partial F}{\partial x_i}$ where $\frac{\partial F}{\partial x_i}$ exists.
- $\lim_{m\to\infty} \frac{\partial F_m}{\partial x_i} = \frac{\partial F}{\partial x_i}$ where $\frac{\partial F}{\partial x_i}$ exists.
- $|\frac{\partial F_m}{\partial x_i}| \leq k$ for all m.
- $F - F_m$ is non-decreasing from x' to $\bar{x}$ in i.

$F - F_m$ is Lipshitz because F, F_m are Lipshitz. Thus, for each m, there exists $c \in \mathbb{R}^n$ such that $c_i = 0$ and $F(x) - F_m(x) + c^\intercal x$ is non-decreasing from x' to $\bar{x}$. Since $c^\intercal x \in \mathcal{F}^1$, we apply previous results to gain $A_i(\bar{x}, x', c^\intercal x) = \text{IG}_i(\bar{x}, x', c^\intercal x) = 0$. Thus,

$$\begin{aligned} A_i(\bar{x}, x', F(x)) - A_i(\bar{x}, x', F_m(x)) &= A_i(\bar{x}, x', F(x)) - A_i(\bar{x}, x', F_m(x)) + A_i(\bar{x}, x', c^\intercal x) \\ &= A_i(\bar{x}, x', F(x) - F_m(x) + c^\intercal x) \\ &\geq 0 \end{aligned}$$

Thus we have $A_i(\bar{x}, x', F) \geq A_i(\bar{x}, x', F_m)$.

Now, because $(\bar{x}, x', F) \in D_{\text{IG}}$, $\int_0^1 \frac{\partial F}{\partial x_i}(x' + t(\bar{x} - x'))dt$ exists and $\frac{\partial F}{\partial x_i}$ exists almost everythere on the path $x' + t(\bar{x} - x')$. Employing DCT, we have:

$$\begin{aligned} A_i(\bar{x}, x', F) &\geq \lim_{m\to\infty} A_i(\bar{x}, x', F_m) \\ &= \lim_{m\to\infty} \mathrm{IG}_i(\bar{x}, x', F_m) \\ &= \lim_{m\to\infty} \int_0^1 (\bar{x}_i - x'_i) \frac{\partial F_m}{\partial x_i}(\gamma(t))dt \\ &= \int_0^1 (\bar{x}_i - x'_i) \frac{\partial F}{\partial x_i}(\gamma(t))dt \\ &= \mathrm{IG}_i(\bar{x}, x', F) \end{aligned}$$

We may also gain the reverse, $A_i(\bar{x}, x', F) \leq \mathrm{IG}_i(\bar{x}, x', F)$, using a similar method. Thus $A_i(\bar{x}, x', F) = \mathrm{IG}_i(\bar{x}, x', F)$, concluding the proof. ■

Appendix D. Proof of Theorem 10

Proof

ii. ⇒ i.) Let $A \in \mathcal{A}^1$ be the IG method, and let $(\bar{x}, x', F)$, $(\bar{x}, x', G) \in [a, b] \times [a, b] \times \mathcal{F}^1$. If $\bar{x}_i = x'_i$, then it is easy to confirm that $\mathrm{IG}(\bar{x}, x', F) = 0$. Suppose $\bar{x}_i \neq x'_i$, $\bar{x}_j \neq x'_j$. Then, supposing $\frac{\partial F}{\partial x_i} \leq \frac{\partial F}{\partial x_j}$, we have:

$$\begin{aligned} \frac{\mathrm{IG}_i(\bar{x}, x', F)}{\bar{x}_i - x'_i} &= \int_0^1 \frac{\partial F}{\partial x_i}(x' + t(\bar{x} - x'))dt \\ &\leq \int_0^1 \frac{\partial F}{\partial x_j}(x' + t(\bar{x} - x'))dt \\ &= \frac{\mathrm{IG}_j(\bar{x}, x', F)}{\bar{x}_j - x'_j} \end{aligned}$$

and IG satisfies symmetric monotonicity.

i. ⇒ ii.) The following proof is inspired by (Young, 1985, Theorem 1). We begin with an important lemma:

Lemma 15 *Let $A \in \mathcal{A}^1$ satisfy completeness, dummy, linearity, and symmetric monotonicity. Then $A(\bar{x}, x', [x - x']^m) = IG(\bar{x}, x', [x - x']^m)$, where $m \in \mathbb{N}_0^n$.*

Proof Let $A \in \mathcal{A}^1$ satisfy completeness, dummy, linearity, and symmetric monotonicity. Fix $\bar{x}$, x'. It is useful to note that $\mathrm{IG}_i(\bar{x}, x', [x - x']^m) = \frac{m_i}{\|m\|_1}[\bar{x} - x']^m$. We proceed by lexicographic induction on $m \in \mathbb{N}_0^n$. What we mean by $m' <_{\text{lex}} m$ is that $m'_i = m_i$ for $1 \leq i < k$, but $m'_k < m_k$.

Let $M \subseteq \mathbb{N}_0^n$ be the set of values of m for which $A(\bar{x}, x', [x-x']^m) = \mathrm{IG}(\bar{x}, x', [x-x']^m) = \frac{1}{\|m\|_1}(m_1, ..., m_n)[\bar{x} - x']^m$. Now, $A(\bar{x}, x', [x - x']^0) = 0 = \mathrm{IG}(\bar{x}, x', [x - x']^0)$ by dummy, so $(0, ..., 0) \in M$. Suppose instead that $\|m\|_0 = 1$, so that only $m_i \neq 0$. By dummy, $A_j(\bar{x}, x', [x - x']^m) = 0$ for $j \neq i$, and by completeness, $A_i(\bar{x}, x', [x - x']^m) = [\bar{x} - x']^m$. Thus $A(\bar{x}, x', [x - x']^m) = \mathrm{IG}(\bar{x}, x', [x - x']^m)$, and $\|m\|_0 = 1$ implies $m \in M$.

Suppose there exists some element in $\mathbb{N}_0^n$ that is not an element in M. Let m^* be the smallest such element. Define $S = \{1 \leq i \leq n : A_i(\bar{x}, x', [x - x']^{m^*}) \neq \mathrm{IG}_i(\bar{x}, x', [x - x']^{m^*})\}$.

By the above, we have that $\|m^*\|_0 \geq 2$. Note that if $i \in S$ then it must be that 1) $\bar{x}_i \neq x'_i$, for otherwise $A_i = 0 = \text{IG}_i$, and 2) $m^*_i > 0$.

Choose i to be the least element in S. A and IG must disagree in two or more components, for if they disagreed in exactly one component, then they could not both satisfy completeness. Thus $i < n$. Define $F(x) = [x - x']^{m^*}$ and define,

$$G(x) = \frac{m^*_i}{m^*_n + 1}(x_1 - x'_1)^{m^*_1} \cdots (x_i - x'_i)^{m^*_i - 1} \cdots (x_n - x'_n)^{m^*_n + 1}$$

Note $\frac{\partial F}{\partial x_i} = \frac{\partial G}{\partial x_n}$. Thus, we have by symmetric monotonicity:

$$\frac{A_i(\bar{x}, x', F)}{\bar{x}_i - x'_i} = \frac{A_n(\bar{x}, x', G)}{\bar{x}_n - x'_n}$$

Also note that $m^{**} = (m_1, ..., m_i - 1, ..., m_n + 1) < m^*$. Thus $m^{**} \notin M$, $A(\bar{x}, x', G) = \text{IG}(\bar{x}, x', G)$. We then have,

$$\begin{aligned}
\frac{A_i(\bar{x}, x', F)}{\bar{x}_i - x'_i} &= \frac{A_n(\bar{x}, x', G)}{\bar{x}_n - x'_n} \\
&= \frac{\text{IG}_n(\bar{x}, x', G)}{\bar{x}_n - x'_n} \\
&= \frac{m^*_i}{\|m\|_0}(x_1 - x'_1)^{m^*_1} \cdots (x_i - x'_i)^{m^*_i - 1} \cdots (x_n - x'_n)^{m^*_n} \\
&= \frac{m^*_i}{\|m\|_0} \frac{(x_1 - x'_1)^{m^*_1} \cdots (x_i - x'_i)^{m^*_i} \cdots (x_n - x'_n)^{m^*_n}}{\bar{x}_i - x'_i} \\
&= \frac{\text{IG}_i(\bar{x}, x', F)}{\bar{x}_i - x'_i}
\end{aligned}$$

This shows that $A_i(\bar{x}, x', F) = \text{IG}_i(\bar{x}, x', F)$ for $i < n$. By completeness, we have $A_n(\bar{x}, x', F) = \text{IG}_n(\bar{x}, x', F)$. Thus $m^* \in M$, a contradiction. Thus there is no element of $\mathbb{N}_0^n$ that is not an element of M, and $M = \mathbb{N}_0^n$ concluding the proof. ■

We now move to the main proof:

Let $A \in \mathcal{A}^1$ satisfy completeness, dummy, linearity, and symmetric monotonicity and let $F \in \mathcal{F}^1$. For any i such that $1 \leq i \leq n$, $\frac{\partial F}{\partial x_i}$ is analytic and by the Stone Weierstrass theorem, for any $\epsilon > 0$, there exists a polynomial, p, such that $|p(x) - \frac{\partial F}{\partial x_i}(x)| < \epsilon$ on $[a, b]$. Let p_m be a polynomial such that $|p_m(x) - \frac{\partial F}{\partial x_i}(x)| < \frac{1}{2m}$, and let P_m be any polynomial so that $\frac{\partial P_m}{\partial x_i} = p_m$. Note that $\frac{\partial (P_m - \frac{x_i}{m})}{\partial x_i} = p_m - \frac{1}{m} < \frac{\partial F}{\partial x_i}$.

Now assume that $\bar{x}_i - x'_i \geq 0$. By symmetric monotonicity we have $A_i(\bar{x}, x', P_m - \frac{x_i}{m}) \leq A_i(\bar{x}, x', F)$. Employing the dominated convergence theorem, we have:

$$\begin{aligned}
A_i(\bar{x}, x', F) &\geq \lim_{m\to\infty} A_i(\bar{x}, x', P_m - \frac{x_i}{m}) \\
&= \lim_{m\to\infty} \mathrm{IG}_i(\bar{x}, x', P_m - \frac{x_i}{m}) \\
&= \lim_{m\to\infty} (\bar{x}_i - x'_i) \int_0^1 \frac{\partial(P_m - \frac{x_i}{m})}{\partial x_i}(\gamma(t))dt \\
&= \lim_{m\to\infty} (\bar{x}_i - x'_i) \int_0^1 p_m(\gamma(t))dt - \frac{(\bar{x}_i - x'_i)}{m} \\
&= (\bar{x}_i - x'_i) \int_0^1 \frac{\partial F}{\partial x_i}(\gamma(t))dt \\
&= \mathrm{IG}_i(\bar{x}, x', F)
\end{aligned}$$

By considering $P_m + \frac{x_i}{m}$, we gain the opposite inequality, namely, $A_i(\bar{x}, x', F) \leq \mathrm{IG}_i(\bar{x}, x', F)$. This establishes that $A_i(\bar{x}, x', F) = \mathrm{IG}_i(\bar{x}, x', F)$.

The case where $\bar{x}_i - x'_i \leq 0$ follows a parallel proof. ■

Appendix E. Proof of Theorem 11

Proof (iii. $\Rightarrow$ ii.) Suppose $A \in \mathcal{A}^2(D_{\mathrm{IG}})$ is the IG method and $(\bar{x}, x', F) \in D_{\mathrm{IG}}$. It is well known that IG satisfies completeness, dummy, and linearity. If $\bar{x}_i = x'_i$, then it is easy to see that $\mathrm{IG}_i(\bar{x}, x', F) = 0$.

Suppose that $(\bar{x}, x', G) \in D_{\mathrm{IG}}$ as well, and that $\bar{x}_i \neq x'_i$, $\bar{x}_j \neq x'_j$. Furthermore, suppose that $\frac{\partial F}{\partial x_i} \leq \frac{\partial F}{\partial x_j}$ locally approximately. Because $(\bar{x}, x', F)$, $(\bar{x}, x', F) \in D_{\mathrm{IG}}$, $\frac{\partial F}{\partial x_i}$ and $\frac{\partial F}{\partial x_j}$ can be integrated along the path $\gamma(t) = x' + t(\bar{x} - x')$, implying that the measure of points on the path where $\frac{\partial F}{\partial x_i}$ and $\frac{\partial F}{\partial x_j}$ exist has full measure with respect to the Lebesgue measure on $\mathbb{R}$. Suppose x is one such point. Then $\lim_{z\to\infty} \frac{F(x_1,\ldots,x_i+z,\ldots,x_n)-F(x)}{z}$, $\lim_{z\to\infty} \frac{G(x_1,\ldots,x_j+z,\ldots,x_n)-G(x)}{z}$ both exist and, because $\frac{\partial F}{\partial x_i} \leq \frac{\partial F}{\partial x_j}$ locally approximately, $\frac{\partial F}{\partial x_i}(x) = \lim_{z\to\infty} \frac{F(x_1,\ldots,x_i+z,\ldots,x_n)-F(x)}{z} \leq \lim_{z\to\infty} \frac{G(x_1,\ldots,x_j+z,\ldots,x_n)-G(x)}{z} = \frac{\partial F}{\partial x_j}(x)$. Thus,

$$\begin{aligned}
\frac{\mathrm{IG}_i(\bar{x}, x', F)}{\bar{x}_i - x'_i} &= \int_0^1 \frac{\partial F}{\partial x_i}(x' + t(\bar{x} - x'))dt \\
&\leq \int_0^1 \frac{\partial F}{\partial x_j}(x' + t(\bar{x} - x'))dt \\
&= \frac{\mathrm{IG}_j(\bar{x}, x', F)}{\bar{x}_j - x'_j}
\end{aligned}$$

and IG satisfies $\mathcal{C}^0$-symmetric monotonicity.

ii. $\Rightarrow$ i.) Suppose $A \in \mathcal{A}^2(D_{\mathrm{IG}})$ satisfies completeness, dummy, linearity, and $\mathcal{C}^0$-symmetric monotonicity. A satisfies symmetric monotonicity for $F \in \mathcal{F}^1$ immediately by the definition of partial derivatives. Suppose that F is non-decreasing from x' to $\bar{x}$ and let $(\bar{x}, x', F) \in D_{\mathrm{IG}}$. If $\bar{x}_i = x'_i$, then $A_i(\bar{x}, x', F) = 0$. Suppose $\bar{x}_i > x'_i$. As previously observed, $\frac{\partial F}{\partial x_i}$ exists almost everywhere on the straight path $\gamma(t)$. Setting $G \equiv 0$, then

$0 = \frac{\partial G}{\partial x_i} \leq \frac{\partial F}{\partial x_i}$ almost approximately since F is non-decreasing from x' to $\bar{x}$ and $\bar{x}_i > x'_i$. Thus $0 = \frac{A_i(\bar{x},x',G)}{\bar{x}_i - x'_i} \leq \frac{A_i(\bar{x},x',F)}{\bar{x}_i - x'_i}$, and $0 \leq A_i(\bar{x}, x', F)$. If instead we assume that $\bar{x}_i < x'_i$, then $0 = \frac{\partial G}{\partial x_i} \geq \frac{\partial F}{\partial x_i}$, and $0 = \frac{\partial G}{\partial x_i} \leq -\frac{\partial F}{\partial x_i}$. Thus $0 = \frac{A_i(\bar{x},x',G)}{\bar{x}_i - x'_i} \leq \frac{A_i(\bar{x},x',-F)}{\bar{x}_i - x'_i}$, and $0 \leq A_i(\bar{x}, x', F)$. Thus, in any case, $A_i(\bar{x}, x', F) \geq 0$ for all i, and A satisfies NDP.

i. $\Rightarrow$ iii.) Suppose $A \in \mathcal{A}^2(D_{\text{IG}})$ satisfies completeness, dummy, linearity, and NDP. Let $(\bar{x}, x', F) \in D_{\text{IG}}$ and choose a component i. By methods found in the proof of Lundstrom et al. (2022, Theorem 2), there exists a sequence of functions F_m such that:

- F_m is analytic for all m.
- $\frac{\partial F_m}{\partial x_i} \leq \frac{\partial F}{\partial x_i}$ where $\frac{\partial F}{\partial x_i}$ exists.
- $\lim_{m\to\infty} \frac{\partial F_m}{\partial x_i} = \frac{\partial F}{\partial x_i}$ where $\frac{\partial F}{\partial x_i}$ exists.
- $|\frac{\partial F_m}{\partial x_i}| \leq k$ for all m.
- $F - F_m$ is non-decreasing from x' to $\bar{x}$ in i.

By NDP we have $A_i(\bar{x}, x', F - F_m) \geq 0$ and $A_i(\bar{x}, x', F) \geq A_i(\bar{x}, x', F_m)$. Since $F_m \in \mathcal{A}^1$, we have $A_i(\bar{x}, x', F_m) = \text{IG}_i(\bar{x}, x', F)$ by Theorem 10. Recalling that $\frac{\partial F}{\partial x_i}$ exists almost everywhere on IG's path, we employ the dominated convergence theorem to gain:

$$\begin{aligned}
A_i(\bar{x}, x', F) &\geq \lim_{m\to\infty} A_i(\bar{x}, x', F_m) \\
&= \lim_{m\to\infty} \text{IG}_i(\bar{x}, x', F_m) \\
&= \lim_{m\to\infty} (\bar{x}_i - x'_i) \int_0^1 \frac{\partial F_m}{\partial x_i}(\gamma(t))dt \\
&= (\bar{x}_i - x'_i) \int_0^1 \frac{\partial F}{\partial x_i}(\gamma(t))dt \\
&= \text{IG}_i(\bar{x}, x', F)
\end{aligned}$$

By a parallel method we can gain $A_i(\bar{x}, x', F) \leq \text{IG}_i(\bar{x}, x', F)$.

■

Appendix F. Softplus Approximations Converge Uniformly

Define S^k_α to be as S^k, but replace each ReLU function s in S^k with the parameterized softplus, s_α. Then the softplus approximation of F is given by:

$$F_\alpha(x) = S^m_\alpha \circ F^m \circ S^{m-1}_\alpha \circ F^{m-1} \circ ... \circ S^2_\alpha \circ F^2 \circ S^1_\alpha \circ F^1(x)$$

Lemma 16 *$F_\alpha \to F$ uniformly on U.*

Proof Begin proof by induction. For $k = 1$, it is easy to show that $s_\alpha \to s$ uniformly on $\mathbb{R}$, and thus, $S^1_\alpha \to S^1$ uniformly on $\mathbb{R}^n$. Thus, for any $\epsilon > 0$, an $A > 0$ may be chosen

such that for any $y \in \mathbb{R}^n$, $\alpha > A$ implies $\|S^1_\alpha(y) - S^1(y)\| < \epsilon$. Replace y with $F^1(x)$ to get $S^1_\alpha(F^1) \to S^1(F^1)$ uniformly.

Write $G^k := S^k \circ F^k \circ ... \circ S^1 \circ F^1(x)$ and $G^k_\alpha := S^k_\alpha \circ F^k \circ ... \circ S^1_\alpha \circ F^1(x)$, and suppose $G_\alpha \to G$ uniformly. It remains to be shown that $S^k_\alpha \circ F^k \circ G^k_\alpha \to S^k \circ F^k \circ G^k$ uniformly.

$$\begin{aligned}
&\|S^k_\alpha(F^k(G^k_\alpha(x))) - S^k(F^k(G^k(x)))\| \\
\leq &\|S^k_\alpha(F^k(G^k_\alpha(x))) - S^k_\alpha(F^k(G^k(x)))\| + \|S^k_\alpha(F^k(G^k(x))) - S^k(F^k(G^k(x)))\| \\
\leq &\|F^k(G^k_\alpha(x)) - F^k(G^k(x))\| + \|S^k_\alpha(F^k(G^k(x))) - S^k(F^k(G^k(x)))\|
\end{aligned}$$

Where the third line is because S^k_α is Lipschitz with Lipschitz constant ≤ 1.

Since G^k is analytic, it is bounded on U. Since G^k_α converges uniformly to G^k, it is bounded for large enough α. Let α_0 produce this bound, that is, if $\alpha > \alpha_0$, then $\max(\|G^k_\alpha(x)\|, \|G^k(x)\|) \leq C_1$ for any $x \in U$. Since F is analytic, it is Lipshitz on bounded domains. Thus, if $\alpha > \alpha_0$, then

$$\|F^k(G^k_\alpha(x)) - F^k(G^k(x))\| \leq C_2\|G^k_\alpha(x) - G^k(x)\|$$

Now, by uniform continuity of G^k_α and S^k_α, choose α_1 so that $\alpha > \alpha_1$ guarantees that $\|G^k_\alpha(x) - G^k(x)\| < \epsilon/2C_2$, and choose α_2 so that $\alpha > \alpha_2$ guarantees that $\|S^k_\alpha(F^k(G^k(x))) - S^k(F^k(G^k(x)))\| < \epsilon/2$. Then $\alpha > \max(\alpha_0, \alpha_1, \alpha_2)$ guarantees that

$$\begin{aligned}
&\|S^k_\alpha(F^k(G^k_\alpha(x))) - S^k(F^k(G^k(x)))\| \\
\leq &\|F^k(G^k_\alpha(x)) - F^k(G^k(x))\| + \|S^k_\alpha(F^k(G^k(x))) - S^k(F^k(G^k(x)))\| \\
\leq &C_2\|G^k_\alpha(x) - G^k(x)\| + \|S^k_\alpha(F^k(G^k(x))) - S^k(F^k(G^k(x)))\| \\
< &\epsilon/2 + \epsilon/2 = \epsilon
\end{aligned}$$

showing that $S^k_\alpha \circ F^k \circ G^k_\alpha \to S^k \circ F^k \circ G^k$ uniformly. ■

Appendix G. Proof of Theorem 13

G.1 Setup

Define S^k_α to be as S^k, but replace each ReLU function s in S^k with the parameterized softplus, s_α. Then the softplus approximation of F is given by:

$$F_\alpha(x) = S^m_\alpha \circ F^m \circ S^{m-1}_\alpha \circ F^{m-1} \circ ... \circ S^2_\alpha \circ F^2 \circ S^1_\alpha \circ F^1(x)$$

Also, for a funciton $G : \mathbb{R}^n \to \mathbb{R}^m$, define DF to be the Jacobian, so that if F_i is the i^{th} output of F, then $(DG)_{i,j} = \frac{\partial G_i}{\partial x_j}$.

G.2 Main Proof

First, we state an outline of the proof. We proceed by induction. In the non-trivial case with one-dimensional output, F^1 is not the zero function and S^1 is ReLU. In this case, $\{y \in U : F^1(y) \neq 0\}$ is open and has full measure. For any x in this set, we can compose F^1 with ReLU and get that $S^1 \circ F^1$ behaves like F^1 or the zero function locally. For each

x in this set, $D(S^1_\alpha \circ F^1)$ converges locally to DF^1 or 0 locally. In the multivariate case, each $(S \circ F^1)_i$ has a set with desired behaviors, so for any x in the intersection of such sets, $S \circ F^1$ has the desired behaviors. That set is open and has full measure.

For the induction step, we assume that G^k has the desired properties and want to show $S^{k+1} \circ F^{k+1} \circ G^k$ does as well. If G^k is equivalent to an analytic function in some neighborhood, so is $F^{k+1} \circ G^k$. An argument similar to the $k = 1$ step shows that for almost every x in our neighborhood, $S^{k+1} \circ F^{k+1} \circ G^k$ is equivalent to an analytic function in some new open neighborhood containing x, and $S^{k+1}_\alpha \circ F^{k+1} \circ G^k_\alpha$ converges. We then consider a collection of points $x \in U$ with the desirable properties, and a collection of open sets N_x containing them, where $S^{k+1} \circ F^{k+1} \circ G^k$ is locally equivalent to an analytic function on N_x. We show that $\cup_x N_x$ is open and has full measure.

Proof Let $F \in \mathcal{F}^1$. As before, write $G^k := S^k \circ F^k \circ ... \circ S^1 \circ F^1$ and $G^k_\alpha := S^k_\alpha \circ F^k \circ ... \circ S^1_\alpha \circ F^1$. Assume that there exists $U^* \subset U$ with same measure as U, and that $x \in U^*$ implies that exists an open region containing x, B_x, such that: 1) $G^k \equiv H_x$ on B_x, where H_x is a real-analytic function on U, 2) $DG^k(x)$ exists, and 3) $DG^k_\alpha(x) \to DG^k(x)$ as $\alpha \to \infty$. We want to show that there is a set analogous to U^* for $S^{k+1} \circ F^{k+1} \circ G^k$ and $S^{k+1}_\alpha \circ F^{k+1} \circ G^k_\alpha$. With this established, we will have gained a proof by induction. To explain, the above is the $k \to k+1$ step. By setting $k = 1$, and setting F^1, S^1 as the identity mappings, we will prove the $k = 1$ step, concluding the proof.

First, let us consider the case where F^{k+1}, S^{k+1} output in one dimension. Let $x \in U^*$, and suppose $G^k \equiv H_x$ on B_x. Then $F^{k+1} \circ G^k$ is analytic on B_x, since compositions of real analytic function are real analytic.

Case 1: Consider the case where $\lambda(\{y \in B_x : G^k(y) = 0)\} > 0$. Then $G^k \equiv 0$ and $S^{k+1} \circ F^{k+1} \circ G^k$ is constant on B_x. In this case, the derivative of $S^{k+1} \circ F^{k+1} \circ G^k$ exists everywhere on B_x, and is equal to zero. Now, for $y \in B_x$, we have

$$\begin{aligned}
\lim_{\alpha\to\infty} \nabla(S^{k+1}_\alpha \circ F^{k+1} \circ G^k_\alpha)(y) &= \lim_{\alpha\to\infty} \sum_{j=1}^{n_k} \frac{dS^{k+1}_\alpha}{d(F^{k+1} \circ G^k)}(F^{k+1}(G^k_\alpha(y))) \\
&\qquad \times \frac{\partial F^{k+1}}{\partial G^k_{\alpha,j}}(G^k_\alpha(y)) \times \nabla G^k_{\alpha,j}(y) \\
&= 0 \\
&= \nabla(S^{k+1} \circ F^{k+1} \circ G^k)(y)
\end{aligned}$$

where the 0 comes from the fact that $|\frac{dS^{k+1}_\alpha}{d(F^{k+1} \circ G^k)}| \leq 1$, $\frac{\partial F^{k+1}}{\partial G^k_{\alpha,j}}$ is bounded for a bounded domain (which it is), and $\nabla G^k_{\alpha,j}(y) \to 0$ for each j. Thus $S^{k+1} \circ F^{k+1} \circ G^k$, $S^{k+1}_\alpha \circ F^{k+1} \circ G^k_\alpha$ have properties 1-3 of the theorem on the set B_x.

Case 2: Consider instead the case where G^k is not the zero function, but S^k is the identity mapping. Then $F^{k+1} \circ G^k$ is analytic on B_x and so is $S^{k+1} \circ F^{k+1} \circ G^k$, and the

derivative exists on B_x. Now, for $y \in B_x$, we have

$$\begin{aligned}
\lim_{\alpha\to\infty} \nabla(S_\alpha^{k+1} \circ F^{k+1} \circ G_\alpha^k)(y) &= \lim_{\alpha\to\infty} \nabla(F^{k+1} \circ G_\alpha^k)(y) \\
&= \lim_{\alpha\to\infty} \sum_{j=1}^{n_k} \frac{\partial F^{k+1}}{\partial G_{\alpha,j}^k}(G_\alpha^k(y)) \times \nabla G_{\alpha,j}^k(y) \\
&= \lim_{\alpha\to\infty} \sum_{j=1}^{n_k} \frac{\partial F^{k+1}}{\partial G_j^k}(G_\alpha^k(y)) \times \nabla G_{\alpha,j}^k(y) \qquad (8) \\
&= \sum_{j=1}^{n_k} \frac{\partial F^{k+1}}{\partial G_j^k}(G^k(y)) \times \nabla G_j^k(y) \\
&= \nabla(F^{k+1} \circ G^k)(y)
\end{aligned}$$

To explain the fourth line, $\nabla G_{\alpha,j}^k(y)$ converges pointwise by assumption. Also, $\frac{\partial F^{k+1}}{\partial G_j^k}$ is Lipschitz continuous in a bounded domain and $G_\alpha^k(y)$ converges uniformly. Thus each term converges pointwise. Thus $S^{k+1} \circ F^{k+1} \circ G^k$, $S_\alpha^{k+1} \circ F^{k+1} \circ G_\alpha^k$ have properties 1-3 of the theorem on the set B_x.

Case 3: Consider the case where G^k is not the zero function and S^{k+1} is the ReLU function. Then $F^{k+1} \circ G^k$ is analytic and either the zero function or not on B_x.

Case 3.1: Consider the subcase where $F^{k+1} \circ G^k \equiv 0$ on B_x. Then $S^{k+1} \circ F^{k+1} \circ G^k \equiv 0$ on B_x, is differentiable on B_x, and the derivative is the zero function. Then for $y \in B_x$, we have

$$\begin{aligned}
\lim_{\alpha\to\infty} \nabla(S_\alpha^{k+1} \circ F^{k+1} \circ G_\alpha^k)(y) &= \lim_{\alpha\to\infty} \frac{dS_\alpha^{k+1}}{d(F^{k+1} \circ G_\alpha^k)}(F^{k+1}(G_\alpha^k(y))) \times \nabla(F^{k+1} \circ G_\alpha^k)(y) \\
&= \lim_{\alpha\to\infty} \frac{dS_\alpha^{k+1}}{d(F^{k+1} \circ G_\alpha^k)}(F^{k+1}(G_\alpha^k(y))) \times \nabla(F^{k+1} \circ G^k)(y) \\
&= \lim_{\alpha\to\infty} \frac{dS_\alpha^{k+1}}{d(F^{k+1} \circ G^k)}(F^{k+1}(G_\alpha^k(y))) \times 0 \\
&= \nabla(S^{k+1} \circ F^{k+1} \circ G^k)(y)
\end{aligned}$$

where the third line is because $\frac{dS_\alpha^{k+1}}{d(F^{k+1} \circ G^k)}$ is bounded and $\nabla(F^{k+1} \circ G_\alpha^k) \to \nabla(F^{k+1} \circ G^k)$ on B_x by Eq. (8). Thus in this subcase, $S^{k+1} \circ F^{k+1} \circ G^k$, $S_\alpha^{k+1} \circ F^{k+1} \circ G_\alpha^k$ have properties 1-3 of the theorem on the set B_x.

Case 3.2: Instead consider the subcase where $F^{k+1} \circ G^k$ is a non-constant function on B_x. We have $\lambda(\{z \in B_x : F^{k+1} \circ G^k(z) = 0\}) = 0$.

Case 3.2.1: Suppose $F^{k+1} \circ G^k(x) > 0$. Because $F^{k+1} \circ G^K$ is continuous, there exists an open set B_x' containing x where $F^{k+1} \circ G^k > 0$, and that on such a set, $S^{k+1} \circ F^{k+1} \circ G^k \equiv F^{k+1} \circ G^k$. Then,

$$\begin{aligned}
\lim_{\alpha\to\infty} \nabla(S_\alpha^{k+1} \circ F^{k+1} \circ G_\alpha^k)(y) &= \lim_{\alpha\to\infty} \frac{dS_\alpha^{k+1}}{d(F^{k+1} \circ G^k)}(F^{k+1}(G_\alpha^k(y))) \times \nabla(F^{k+1} \circ G_\alpha^k)(y) \\
&= 1 \times \nabla(F^{k+1} \circ G^k)(y) \\
&= \nabla(S^{k+1} \circ F^{k+1} \circ G^k)(y)
\end{aligned}$$

Case 3.2.2: Suppose $F^{k+1} \circ G^k(x) < 0$. Because $F^{k+1} \circ G^K$ is continuous, there exists an open set B'_x containing x where $F^{k+1} \circ G^k < 0$, and that on such a set,, $S^{k+1} \circ F^{k+1} \circ G^k \equiv 0$. Then,

$$\begin{aligned}\lim_{\alpha\to\infty} \nabla(S^{k+1}_\alpha \circ F^{k+1} \circ G^k_\alpha)(y) &= \lim_{\alpha\to\infty} \frac{dS^{k+1}_\alpha}{d(F^{k+1} \circ G^k)}(F^{k+1}(G^k_\alpha(y))) \times \nabla(F^{k+1} \circ G^k_\alpha)(y) \\ &= 0 \times \nabla(F^{k+1} \circ G^k)(y) \\ &= \nabla(S^{k+1} \circ F^{k+1} \circ G^k)(y)\end{aligned}$$

Case 3.2.3: Suppose $F^{k+1} \circ G^k(x) = 0$. In this case, we do not define a B'_x set. We remind the reader that if $x \in U^*$ is a case 3.2.3 point, then $\lambda(\{z \in B_x : F^{k+1} \circ G^k(z) = 0\}) = 0$

Thus we have established in the one-dimensional output case that for each $x \in U^*$ that is not a case 3.2.3 point, there exists an open neighborhood containing x where properties 1-3 hold.

Now consider the multivariate case. Define $K \subset U^*$ as the set of points in U^* that are case 3.2.3 points for at least one output of $S^{k+1} \circ F^{k+1} \circ G^k$. Let $x \in U^* \setminus K$. Let $B'_{x,i}$ correspond to the open set containing x where properties 1-3 hold when we only consider the output $(S^{k+1} \circ F^{k+1} \circ G^k)_i$. Then properties 1-3 hold on $\cap_i B'_{x,i}$ for each output of $(S^{k+1} \circ F^{k+1} \circ G^k)_i$ and $(S^{k+1}_\alpha \circ F^{k+1} \circ G^k_\alpha)_i$. Thus properties 1-3 hold for $S^{k+1} \circ F^{k+1} \circ G^k$ and $S^{k+1}_\alpha \circ F^{k+1} \circ G^k_\alpha$ on $\cap_i B^*_{x,i}$. Thus we have established in the multivariate case the following: for each $x \in U^* \setminus K$, there exists an open neighborhood containing x, B'_x, where properties 1-3 hold for $S^{k+1} \circ F^{k+1} \circ G^k$ and $S^{k+1}_\alpha \circ F^{k+1} \circ G^k_\alpha$.

We now move to show that $\lambda(K) = 0$, which will conclude the proof. Let K_i denote the set of case 3.2.3 points for the i^{th} output of $S^{k+1} \circ F^{k+1} \circ G^k$. Since $K = \cup_i K_i$, it suffices to show $\lambda(K_i) = 0$. Let $x \in K_i$ for some i. Then x is a case 3.2.3 point for the output of $(S^{k+1} \circ F^{k+1} \circ G^k)_i$. According to our assumption, there exists a B_x containing x where properties 1-3 hold for G^k, G^k_α on B_x. Note that $\cup_{x\in K_i} B_x$ is an open cover, and has a countable subcover $\cup_{j\in\mathbb{N}} B_{x_j}$, where each x_j is a case 3.2.3 point for $(S^{k+1}F^{k+1} \circ G^k)_i$. Because $K_i \subseteq \cup_{x\in K_i} B_x$, we also have $K_i \subseteq \cup_{j\in\mathbb{N}} B_{x_j}$. Now,

$$\begin{aligned}K_i &= K_i \cap (\cup_{j\in\mathbb{N}} B_{x_j}) \\ &= \cup_{j\in\mathbb{N}} (B_{x_j} \cap K_i)\end{aligned}$$

Now, if $x \in K_i$, then $(F^{k+1} \circ G^k)_i(x) = 0$ by virtue of being a case 3.2.3 point. Also, it has been established that for case 3.2.3 points, $\lambda(\{z \in B_x : (F^{k+1} \circ G^k)_i(z) = 0\}) = 0$. Thus $\lambda(B_{x_j} \cap K_i) \leq \lambda(\{z \in B_{x_j} : (F^{k+1} \circ G^k)_i(z) = 0\}) = 0$. Thus, K_i is a countable union of sets of measure zero, and is thus measure zero. ■

Appendix H. Proof of Corollary 14

Proof Let $F \in \mathcal{F}^2$, and let U be the set as in Theorem 13. Let $\gamma(t)$ be the uniform speed path from x' to $\bar{x}$ and suppose $\lambda(\{t \in [0,1] : \gamma(t) \in U\}) = 1$, where m is the Lebesgue measure on $\mathbb{R}$. By Theorem 13, we have $\nabla F_\alpha(\gamma(t)) \to \nabla F(\gamma(t))$ for almost every t in $[0,1]$. Suppose ∇F_α is bounded on U for large enough α. Let a_n be any sequence such that $a_n \to \infty$. Choose any index i, and by employing the dominated convergence theorem we

gain:

$$\begin{aligned}\lim_{n\to\infty} \mathrm{IG}_i(\bar{x}, x', F_{a_n}) &= \lim_{n\to\infty} (\bar{x}_i - x'_i) \int_0^1 \frac{\partial F_{a_n}}{\partial x_i}(\gamma(t))dt \\ &= (\bar{x}_i - x'_i) \int_0^1 \frac{\partial F}{\partial x_i}(\gamma(t))dt \\ &= \mathrm{IG}_i(\bar{x}, x', F)\end{aligned}$$

Since $\lim_{n\to\infty} \mathrm{IG}_i(\bar{x}, x', F_{a_n}) = \mathrm{IG}_i(\bar{x}, x', F)$ for any sequence a_n, we have $\lim_{\alpha\to\infty} \mathrm{IG}_i(\bar{x}, x', F_\alpha) = \mathrm{IG}_i(\bar{x}, x', F)$.

We now turn to show that ∇F_α is bounded for large enough α. Using the notation introduced in Theorem 13, note that:

$$\nabla F_\alpha = DS_\alpha^m \, DF^m \, DS_\alpha^{m-1} \, DF^{m-1} ... DS_\alpha^2 \, DF^2 \, DS_\alpha^1 \, DF^1$$

Thus,

$$\|\nabla F_\alpha\|_\infty \leq \Pi_{k=1}^m \|DS_\alpha^k(F^k \circ ... \circ F^1)\|_\infty \times \|DF^k(S_\alpha^{k-1} \circ ... \circ F^1)\|_\infty$$

Now $\|DS_\alpha^k(F^k \circ ... \circ F^1)\|_\infty \leq 1$ since S_α^k is either softplus or the identity mapping for each input. Also, F^k is Lipshitz in a bounded domain, and $S_\alpha^{k-1} \circ ... \circ F^1$ converges uniformly on U to a function with a bounded range. Thus $\|DF^k(S_\alpha^{k-1} \circ ... \circ F^1)\|_\infty$ is bounded on U, and $\|\nabla F_\alpha\|_\infty$ is bounded. ■