How to Address Monotonicity for Model Risk Management?

Dangxing Chen [1] **Weicheng Ye** [1]

Abstract

In this paper, we study the problem of establishing the accountability and fairness of transparent machine learning models through monotonicity. Although there have been numerous studies on individual monotonicity, pairwise monotonicity is often overlooked in the existing literature. This paper studies transparent neural networks in the presence of three types of monotonicity: individual monotonicity, weak pairwise monotonicity, and strong pairwise monotonicity. As a means of achieving monotonicity while maintaining transparency, we propose the monotonic groves of neural additive models. As a result of empirical examples, we demonstrate that monotonicity is often violated in practice and that monotonic groves of neural additive models are transparent, accountable, and fair.

1. Introduction

There has been growing public concern over the misuse of artificial intelligence models in the absence of regulations, despite the success of artificial intelligence (AI) and machine learning (ML) in many fields (Radford et al., 2019; He et al., 2016; Chen & Guestrin, 2016). The European Commission (EC) has proposed the Artificial Intelligence Act (AIA) (EU2, 2021), which represents a significant first step toward filling the regulatory void. Regulations regarding artificial intelligence should consider transparency, accountability, and fairness (Carlo et al., 2021; OCC, 2021).

Many efforts have been made to develop transparent ML models (Rudin, 2019; Agarwal et al., 2021; Yang et al., 2021; Tsang et al., 2020; Hastie, 2017; Caruana et al., 2015; Lou et al., 2013). A transparent model facilitates the explanation of how it makes decisions, therefore allowing us to easily verify conceptual soundness and fairness.

Nevertheless, conceptual soundness and fairness are not necessarily guaranteed for ML models, even if they are transparent. Our focus in this paper is on monotonicity, one of the most important indicators. In recent years, monotonic machine learning models have received extensive research attention (Yanagisawa et al., 2022; Liu et al., 2020; Milani Fard et al., 2016; You et al., 2017; Potharst & Feelders, 2002; Duivesteijn & Feelders, 2008). These studies have led to a more reasonable and fair approach to ML. The majority of papers, however, focus on individual monotonicity, that is, on the fact that a model is monotonic with a particular feature. It was only recently pointed out that individual monotonicity is insufficient to summarize all relevant information (Chen & Ye, 2022; Gupta et al., 2020). It is also important to consider pairwise monotonicity, a monotonicity that considers monotonicity between different features. Furthermore, most of these models are not necessarily transparent.

In this paper, pairwise monotonicity is explored in more detail, particularly in the context of transparent machine learning models. We divide pairwise monotonicity into two types: the pairwise monotonicity introduced in (Chen & Ye, 2022) is classified as weak pairwise monotonicity, and monotonic dominance discussed in (Gupta et al., 2020) is classified as strong pairwise monotonicity. Time and severity are the two most common causes of pairwise monotonicity. In terms of time, recent information should often be considered more important than older information. For example, in credit scoring, if there is one past due, the credit score should be lower if the past due occurred recently. It is important to take into account such pairwise monotonicity in order to give people the opportunity to improve. It is important that all individuals have the opportunity to succeed without being solely based on their past behaviors. In terms of severity, some events are intrinsically more severe than others due to the nature of justice. A felony, for example, is more serious than a misdemeanor in criminal justice. It is important to maintain pairwise monotonicity as justice is an important component of fairness and a good society should have a system of reward and punishment that is fair. Furthermore, weak and strong pairwise monotonicity are distinguished based on whether two features can only be compared at the same magnitude. Strong pairwise monotonicity occurs

[1]Zu Chongzhi Center for Mathematics and Computational Sciences, Duke Kunshan University, Kunshan, Jiangsu, China. Correspondence to: Dangxing Chen <dangxing.chen@dukekunshan.edu.cn>.

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

when two features can be compared at any level. Justice usually dictates the making of such comparisons.

Pairwise monotonicity is analyzed and its impact on statistical interactions is discussed. The traditional way to check additive separability should incorporate monotonicity constraints. Features with strong pairwise monotonicity and diminishing marginal effects should not be separated, even if data indicate otherwise. A new class of monotonic groves of neural additive models (MGNAMs) is presented to incorporate three types of monotonicity into transparent neural networks. We demonstrate empirically that pairwise monotonicities frequently occur in a wide range of fields, including finance, criminology, and healthcare. Overall, MGNAMs provide a transparent, accountable, and fair framework.

2. Monotonicity

For problem setup, assume we have $\mathcal{D} \times \mathcal{Y}$, where $\mathcal{D}$ is the dataset with n samples and m features and $\mathcal{Y}$ is the corresponding numerical values in regression and labels in classification. We assume the data-generating process

$$y = f(\mathbf{x}) + \epsilon \tag{1}$$

for regression problems and

$$y|\mathbf{x} = \text{Bernoulli}(f(\mathbf{x})) \tag{2}$$

for binary classification problems. For simplicity, we assume $\mathbf{x} \in \mathbb{R}^m$. Then ML methods are applied to approximate f.

2.1. Individual Monotonicity

Throughout the paper, without loss of generality, we focus on the monotonic increasing functions. Suppose $\boldsymbol{\alpha}$ is the list of all individual monotonic features and $\neg\boldsymbol{\alpha}$ its complement, then the input $\mathbf{x}$ can be partitioned into $\mathbf{x} = (\mathbf{x}_{\boldsymbol{\alpha}}, \mathbf{x}_{\neg\boldsymbol{\alpha}})$. Then we have the following definition.

Definition 2.1. We say f is **individually monotonic** with respect to $\mathbf{x}_{\boldsymbol{\alpha}}$ if

$$\begin{aligned} &f(\mathbf{x}_{\boldsymbol{\alpha}}, \mathbf{x}_{\neg\boldsymbol{\alpha}}) \leq f(\mathbf{x}'_{\boldsymbol{\alpha}}, \mathbf{x}_{\neg\boldsymbol{\alpha}}), \\ &\mathbf{x}_{\boldsymbol{\alpha}} \leq \mathbf{x}'_{\boldsymbol{\alpha}}, \forall \mathbf{x}_{\boldsymbol{\alpha}}, \mathbf{x}'_{\boldsymbol{\alpha}}, \mathbf{x}_{\neg\boldsymbol{\alpha}}, \end{aligned} \tag{3}$$

where $\mathbf{x}_{\boldsymbol{\alpha}} \leq \mathbf{x}'_{\boldsymbol{\alpha}}$ denotes the inequality for all entries, i.e., $x_{\alpha_i} \leq x'_{\alpha_i}, \forall i$.

Here is an example of individual monotonicity.

Example 2.2. In credit scoring, the probability of default should increase as the number of past due increases.

For a differentiable function f, individual monotonicity with respect to $\mathbf{x}_{\boldsymbol{\alpha}}$ can be verified if

$$\min_{\mathbf{x},i} \frac{\partial f(\mathbf{x})}{\partial x_{\alpha_i}} \geq 0. \tag{4}$$

2.2. Pairwise Monotonicity

There are some features that are intrinsically more important than others in practice. Analog to (3), we partition $\mathbf{x} = (x_\beta, x_\gamma, \mathbf{x}_\neg)$. Without loss of generality, we assume x_β is more important than x_γ. As a result of multiple features encountering pairwise monotonicity, we record them in two lists $\mathbf{u}$ and $\mathbf{v}$ such that u_i is more important than v_i. Lastly, we require all features with pairwise monotonicity also satisfy individual monotonicity.

2.2.1. Weak Pairwise Monotonicity

We classify the pairwise monotonicity introduced in (Chen & Ye, 2022) as the weak pairwise monotonicity. The definition is given as follows.

Definition 2.3. We say f is **weakly monotonic** with respect to x_β over x_γ if

$$\begin{aligned} &f(x_\beta, x_\gamma + c, \mathbf{x}_\neg) \leq f(x_\beta + c, x_\gamma, \mathbf{x}_\neg), \\ &\forall x_\beta, x_\gamma \text{ s.t. } x_\beta = x_\gamma, \forall \mathbf{x}_\neg, \forall c \in \mathbb{R}^+. \end{aligned} \tag{5}$$

We give an example of weak pairwise monotonicity below.

Example 2.4. Functions should be weakly monotonic with respect to features containing current information over features containing past information. Following Example 2.2, let x_β and x_γ count the number of past dues within two years and two years ago, then the probability of default is weakly monotonic with respect to x_β over x_γ.

Such monotonicity is considered weak due to the condition of $x_\beta = x_\gamma$. Using this condition ensures that the effects of features on the function are compared at the same magnitude, and can therefore be viewed as a more general definition.

Suppose f is differentiable and is weakly monotonic with respect to u_i over v_i for all i in lists $\mathbf{u}$ and $\mathbf{v}$, then the weak pairwise monotonicity can be verifed as

$$\min_{\widetilde{\mathbf{x}},i} \left(\frac{\partial f}{\partial x_{u_i}}(\widetilde{\mathbf{x}}) - \frac{\partial f}{\partial x_{v_i}}(\widetilde{\mathbf{x}}) \right) \geq 0. \tag{6}$$

where $\widetilde{x}_{u_i} = \widetilde{x}_{v_i}$ in $\widetilde{\mathbf{x}}$.

2.2.2. Strong Pairwise Monotonicity

In addition to the weak pairwise monotonicity, there exists a stronger condition of pairwise monotonicity. We classify the monotonic dominance introduced in (Gupta et al., 2020) as the strong pairwise monotonicity.

Definition 2.5. We say f is **strongly monotonic** with respect to x_β over x_γ if

$$\begin{aligned} &f(x_\beta, x_\gamma + c, \mathbf{x}_\neg) \leq f(x_\beta + c, x_\gamma, \mathbf{x}_\neg), \\ &\forall x_\beta, x_\gamma, \mathbf{x}_\neg, \forall c \in \mathbb{R}^+. \end{aligned} \tag{7}$$

The difference between strong/weak monotonicity is whether the condition $x_\beta = x_\gamma$ is needed. Strong monotonicity implies the impacts of increments of some features are more important than others at any point. Note that the features in Example 2.4 are only weakly pairwise monotonic, not strongly pairwise monotonic. Adding more past dues to the credit score will have a different impact based on the number of past dues. Thus, current and past features cannot be directly compared, unless they are of equal magnitude. We provide an example of strong pairwise monotonicity below.

Example 2.6. In criminology, an additional felony is always considered more serious than an additional misdemeanor. Therefore, the probability of recidivism should be strongly monotonic with respect to felonies over misdemeanors.

Clearly, we have the following Lemma.

Lemma 2.7. *If f is strongly monotonic with respect to x_β over x_γ, then f is also weakly monotonic with respect to x_β over x_γ.*

For a differentiable function f, suppose f is strongly monotonic with respect to u_i over v_i for all i in lists $\mathbf{u}$ and $\mathbf{v}$, then the strong pairwise monotonicity can be verifed as

$$\min_{\mathbf{x},i} \left(\frac{\partial f}{\partial x_{u_i}}(\mathbf{x}) - \frac{\partial f}{\partial x_{v_i}}(\mathbf{x}) \right) \geq 0. \tag{8}$$

Strong pairwise monotonicity is transitive and we provide the following Lemma, where proof is provided in Appendix A.1.

Lemma 2.8. *If f is strongly monotonic with respect to x_β over x_γ and x_γ over x_δ, then f is strongly monotonic with respect to x_β over x_δ.*

3. Statistical Interactions

The study of transparent machine learning models has become increasingly popular in order to improve explanation and compliance with regulatory requirements. As a general rule, we should avoid interactions between features if they do not exist in order to maintain transparency in our models. One popular class of transparent models is generalized additive models (GAMs) (Hastie, 2017) of the form

$$f(\mathbf{x}) = \alpha + \sum_{i=1}^{m} f_i(x_i). \tag{9}$$

GAMs are transparent in that statistical interactions are not included. (Agarwal et al., 2021; Caruana et al., 2015) have shown that combination of GAMs with ML models achieved high accuracy for many datasets. In this section, we discuss whether we could incorporate three types of monotonicity into GAMs.

3.1. Individual and Weak Pairwise Monotonicity for GAMs

In GAMs, individual and weak pairwise monotonicity can be easily enforced. Assume that f follows the GAM (9) of the form and is differentiable. If f is individually monotonic with respect to x_α, then we need

$$f'_\alpha(x) \geq 0, \ \forall x \in \mathbb{R}. \tag{10}$$

Similarly, if f is weakly monotonic with respect to x_β over x_γ, then the weak pairwise monotonicity requires that

$$f'_\beta(x) \geq f'_\gamma(x), \forall x \in \mathbb{R}. \tag{11}$$

Constraints such as these can be easily implemented (Chen & Ye, 2022). Furthermore, without statistical interactions, weak pairwise monotonicity is also transitive, as illustrated in the following Lemma with proof in Appendix A.1.

Lemma 3.1. *If f follows the GAM (9), f is weakly monotonic with respect to x_β over x_γ and x_γ over x_δ, then f is weakly monotonic with respect to x_β over x_δ.*

3.2. Additive Separability

Statistical interactions can be determined by checking additive separability. For simplicity, suppose there are two groups: $\mathbf{x}$ can be split into two components $\mathbf{x}_U$ and $\mathbf{x}_V$, with $U \cup V = D$ and $U \cap V = \emptyset$, where $D = \{1, \ldots, m\}$. Extending it to multiple groups is straightforward.

Definition 3.2. We say a function f with D is strictly additive separable for U and V if

$$f(\mathbf{x}) = g(\mathbf{x}_U) + h(\mathbf{x}_V) \tag{12}$$

for some functions g and h, $U \cup V = D$, and $U \cap V = \emptyset$.

Recently, statistical interactions have been studied extensively in the existing literature (Sorokina et al., 2008; Tsang et al., 2018b; 2020). Roughly speaking, we wish to know whether there are interactions between groups U and V. As implied from the name, the conclusion is often drawn according to whether such interactions are statistically significant. There are many different rules to check statistical significance, as a simple example, we might consider a threshold ϵ and check whether the accuracy deteriorates if no interactions are assumed.

Verify additive separability:

$$|\text{Acc}(f(\mathbf{x})) - \text{Acc}(g(\mathbf{x}_U) + h(\mathbf{x}_V))| < \epsilon. \tag{13}$$

If the criteria are satisfied, it seems reasonable to conclude that there are no interactions between U and V. When it comes to GAMs, if a GAM achieves similar accuracy as the black-box ML model, we may conclude that no interaction is necessary.

3.3. Additive Separability in the Presence of Monotonicity

In the context that monotonicity is required, we should add the monotonicity into the requirement of additive separability. That motivates us to modify the rule of Equation 13.

Verify additive separability with monotonicity:

$$|\text{Acc}(f(\mathbf{x})) - \text{Acc}(g(\mathbf{x}_U) + h(\mathbf{x}_V))| < \epsilon, \quad (14)$$

f and $g + h$ have required monotonicity.

For statistical interactions with monotonicity, monotonicity constraints for $g + h$ are essential, since we may not have sufficient data for statistically significant results. In spite of this, neglecting such a statistical interaction may have catastrophic consequences. To illustrate our idea, consider the following example of credit scoring.

Example 3.3. Suppose $\mathbf{x} = (x_\beta, x_\gamma)$ where x_β counts the number of past dues of more than 60 days and x_γ counts the number of past dues between 30 and 59 days. Assume f calculates the probability of defaults. Clearly, f should be strongly monotonic with respect to x_β over x_γ. For simplicity, consider the values of f in the region where $x_\beta + x_\gamma \leq 2$. Suppose the true function f and an additive approximation $\widetilde{f} = f_1 + f_2$ are given in Table 1. If there are no data for $\mathbf{x} = (1, 1)$, then $\widetilde{f}$ exactly fits f in all training data. According to the criteria (13), x_β and x_γ can be well separated. However, $\widetilde{f}$ doesn't have strong pairwise monotonicity and $\widetilde{f}(1, 1) > \widetilde{f}(2, 0)$ causes algorithmic unfairness. Furthermore, such rules could encourage people with $\mathbf{x} = (1, 1)$ to wait for an additional month to pay back to change to $\mathbf{x} = (2, 0)$ in order to obtain a lower probability of default, and therefore higher credit score. Even worse, ML models might not recognize this from data in the long run, as people would intentionally avoid the state $\mathbf{x} = (1, 1)$. Data does not reveal such a problem, thus it must be considered in advance.

3.4. In the Presence of Strong Pairwise Monotonicity

We argue that there exists a common situation in which features with strong pairwise monotonicity cannot be separated, except in the trivial case. Let us consider the following proposition, whereas the proof is in Appendix A.1.

Proposition 3.4. *Suppose f takes the GAM form (9), f is differentiable, individually monotonic with respect to x_β and x_γ, and strongly monotonic with respect to x_β over x_γ. If there exists x^* such that $f'_\beta(x^*) = 0$, then $f_\gamma(x)$ is a constant function.*

According to the Proposition, under such additive forms, f_γ is a constant function, which can be inconsistent with reality. Sadly, such phenomena are common in practice, and one of the most common causes is diminishing marginal effects.

Table 1. Comparison between f with strong pairwise monotonic features and an additive approximation $\widetilde{f} = f_1 + f_2$. f is strongly monotonic with respect to x_β over x_γ. $\widetilde{f}$ violates strong pairwise monotonicity at $\mathbf{x} = (1, 1)$.

TRUE f			
2	0.4		
1	0.3	**0.35**	
0	0	0.2	0.3
$x_\beta \backslash x_\gamma$	0	1	2
$\widetilde{f} = f_1 + f_2$			
2	0.4		
1	0.3	**0.5**	
0	0	0.2	0.3
$x_\beta \backslash x_\gamma$	0	1	2

We provide the definition below.

Definition 3.5. Suppose $\mathbf{x} = (x_\alpha, \mathbf{x}_{\neg\alpha})$. We say a differentiable function f has the **diminishing marginal effect (DME)** with respect to x_α if followings hold

1. $\frac{\partial}{\partial x_\alpha} f(x_\alpha, \mathbf{x}_{\neg\alpha}) > 0$
2. $\frac{\partial^2}{\partial x_\alpha^2} f(x_\alpha, \mathbf{x}_{\neg\alpha}) < 0$
3. $\lim_{x_\alpha \to \infty} \frac{\partial}{\partial x_\alpha} f(x_\alpha, \mathbf{x}_{\neg\alpha}) = 0.$

As a matter of fact, DMEs are quite common in practice. For example, the Cobb-Douglas utility function, $u(x, y) = x^a y^{1-a}$ with $0 < a < 1$, is commonly used to illustrate diminishing marginal utility in economics.

Proposition 3.4 suggests that DMEs may prevent us from separating features with strong pairwise monotonicity. Features with strong pairwise monotonicity that exhibits DME patterns must be assumed to be non-separable at the time of its emergence. Therefore, GAMs are insufficient to incorporate strong pairwise monotonicity in this case.

3.5. Implications on Binary Features

There is an exception to the previous analysis, which is when features are binary since DMEs do not apply. In this case, we have the following Lemma, whereas the proof is left in Appendix A.1.

Lemma 3.6. *For binary features, weak pairwise monotonicity coincides with strong pairwise monotonicity.*

In this case, features can still be additive separable in the linear form. Consider the linear regression of the following form for simplicity

$$f(\mathbf{x}) = \alpha + \sum_{i=1}^{m} \beta_i x_i.$$

Suppose f is monotonic with respect to x_γ over x_δ, then we require $\beta_\gamma > \beta_\delta$, and the additive separability can be achieved.

4. Monotonic Groves of Neural Additive Models

There has been an increasing demand for transparent models recently. In this direction, Neural additive models (NAMs) (Agarwal et al., 2021) and its monotonic version (Chen & Ye, 2022) provide the most transparent neural networks by avoiding statistical interactions, and have been very successful. NAMs have assumed that each f_i in Equation (9) is parametrized by neural networks (NNs). Despite their success, they cannot handle strong pairwise monotonicity, as discussed above. We aim to develop a new model that will maintain transparency to the greatest extent possible, in the manner of NAMs, as well as incorporate strong pairwise monotonicity. Thus, we consider a more general form, namely the groves of neural additive models (GNAMs), similar to (Sorokina et al., 2008),

$$f(\mathbf{x}) = \alpha + \sum_{p:p\in P} f_p(x_p) + \sum_{q:q\in Q} f_q(\mathbf{x}_q), \qquad (15)$$

f_p and f_q are parametrized by NNs. There exists five types of features:

- Nonmonotonic features
- Features with only individual monotonicity
- Features with only weak pairwise monotonicity
- Features with only strong pairwise monotonicity
- Features with both strong and weak pairwise monotonicity

The first three types of features are trained by 1-dimensional functions f_p, just like monotonic NAMs (MNAMs) (Chen & Ye, 2022). Different from MNAMs, $\mathbf{x}_q$ can be higher-dimensional. For the last two types, when there is strong pairwise monotonicity involved, features with pairwise monotonicity should be grouped together in q. Note we group features with both strong and weak pairwise monotonicity to avoid unfair comparisons. Detailed explanations can be found in Appendix A.2.

Regularized algorithms are used to enforce monotonicity. In GNAMs' architecture, motivated by conditions (4), (6), and (8), we consider the optimization problem:

$$\min_{\mathbf{\Theta}} \ell(\mathbf{\Theta}) + \lambda_1 h_1(\mathbf{\Theta}) + \lambda_2 h_2(\mathbf{\Theta}) + \lambda_3 h_3(\mathbf{\Theta}), \qquad (16)$$

where $\ell(\mathbf{\Theta})$ is the mean-squared error for regressions and log-likelihood function for classifications, and

- Individual monotonicity: suppose $\boldsymbol{\alpha}$ is the list of individual monotonic features, then

$$h_1(\mathbf{\Theta}) = \sum_{\alpha\in\boldsymbol{\alpha}} \int_{\mathbb{R}^m} \max\left(0, -\frac{\partial f(\mathbf{x};\mathbf{\Theta})}{\partial x_\alpha}\right)^2 d\mathbf{x}.$$

- Weak pairwise monotonicity: suppose $\mathbf{u}$ and $\mathbf{v}$ are weak pairwise monotonic lists such that f is weakly monotonic with respect to u_i over v_i, then

$$h_2(\mathbf{\Theta}) = \sum_{i=1}^{|\mathbf{u}|} \int_{\mathbb{R}^{m-1}} \max\left(0, \Delta f(\widetilde{\mathbf{x}}_i, \mathbf{\Theta})\right)^2 d\widetilde{\mathbf{x}}_i$$

  where

$$\Delta f(\widetilde{\mathbf{x}}_i, \mathbf{\Theta}) = -\frac{\partial f(\widetilde{\mathbf{x}}_i;\mathbf{\Theta})}{\partial x_{u_i}} + \frac{\partial f(\widetilde{\mathbf{x}}_i;\mathbf{\Theta})}{\partial x_{v_i}}$$

  and $x_{u_i} = x_{v_i}$ in $\widetilde{\mathbf{x}}_i$.

- Strong pairwise monotonicity: suppose $\mathbf{y}$ and $\mathbf{z}$ are strong pairwise monotonic lists such that f is strongly monotonic with respect to y_i over z_i, then

$$h_3(\mathbf{\Theta}) = \sum_{i=1}^{|\mathbf{y}|} \int_{\mathbb{R}^m} \max\left(0, \Delta f_i(\mathbf{x}, \mathbf{\Theta})\right)^2 d\mathbf{x}$$

  where

$$\Delta f_i(\mathbf{x}, \mathbf{\Theta}) = -\frac{\partial f(\mathbf{x};\mathbf{\Theta})}{\partial x_{y_i}} + \frac{\partial f(\mathbf{x};\mathbf{\Theta})}{\partial x_{z_i}}.$$

In the GNAM's architecture (15), computational dimensions can be reduced. For example, when calculating partial derivatives for features in the group q, it is sufficient to evaluate ∂f_q instead of ∂f. In practice, we replace the integral with the equispaced discrete approximations. In the optimization procedure, we also replace all $\max(0, \cdot)$ with $\max(\epsilon, \cdot)$.

We gradually increase λ_1, λ_2, and λ_3 until penalty terms vanish. The two-step procedure is summarized in Algorithm 1. We refer to the GNAM that satisfies all required monotonic constraints (3), (5), and (7) as the monotonic groves of neural additive model (MGNAM).

5. Empirical Examples

This section evaluates the performance of models for a variety of datasets in different fields, including finance, criminology, and health care. We compare fully-connected neural networks (FCNNs), neural additive models (NAMs), monotonic neural additive models (MNAMs), and monotonic groves of neural additive models (MGNAMs). For MNAMs, strong pairwise monotonicity is replaced by weak pairwise

Algorithm 1 Monotonic Groves of Neural Additive Model

Initialization: $\lambda_1 = \lambda_2 = \lambda_3 = 0$, the architecture of the GNAM (P and Q)
Train a GNAM by (16)
while $\min(h_1, h_2, h_3) > 0$ **do**
  Increase λ_i if $h_i > 0$
  Retrain the GNAM by (16).
end while

Table 2. Model performance of the GMSC dataset. All ML models perform similarly.

MODEL/METRICS	CLASSIFICATION ERROR	AUC
FCNN	6.6%	79.5%
NAM	6.6%	79.8%
MNAM	6.6%	80.0%
MGNAM	6.5%	80.2%

monotonicity. We use FCNNs to check the accuracy of black-box ML models and NAMs/MNAMs for visualizations. We do not consider other models here as the general comparison of accuracy is not our focus, but the conceptual soundness and fairness. More details of the dataset, models, and experiments setup are provided in Appendix A.3.

5.1. Finance - Credit Scoring

In credit scoring, statistical models are used to assess an individual's creditworthiness. A popularly used dataset is the Kaggle credit score dataset [1]. In this dataset, we have included three delinquency features that quantify the number of past dues and their duration: 30-59 days, 60-89 days, and 90+ days. To demonstrate the strong pairwise monotonicity of this dataset, we focus on these three features. Without loss of generality, we denote them as x_1, x_2, and x_3. When an additional past due exceeds 90 days, the system should take it much more seriously than when it exceeds 60-89 days, which should take it much more seriously than when it exceeds 30-59 days. We, therefore, impose strong pairwise monotonicity on this order. In the event that such strong pairwise monotonicity is violated, customers with longer past dues could have a higher credit score, thereby causing algorithmic unfairness. In addition, customers with shorter past dues may wish to delay their payments in order to increase their credit score.

A summary of the model performance is provided in Table 2. There is no significant difference in accuracy between the different methods, indicating that transparent neural networks are sufficient for this dataset.

Next, we evaluate conceptual soundness and fairness. For simplicity, we focus on the number of past dues in each period that are less than or equal to two, that is, $0 \leq x_1, x_2, x_3 \leq 2$. We will begin by examining the result of the NAM since it is straightforward to visualize. A comparison of the associated functions is provided in Figure 1. The pairwise monotonicity is clearly violated when there is more than one past due. For example, the feature with 30-59 days past due becomes more important than the feature with 60-89 days past due. Then, we evaluate the MNAM with function values in Table 3. Both individual monotonicity and weak pairwise monotonicity are satisfied. But when statistical interactions are involved for large x, the strong pairwise monotonicity is violated. As an example, consider an applicant who has three past dues, with $x_3 = 1$. If $(x_1, x_2, x_3) = (0, 2, 1)$, then it should be punished more severely than $(x_1, x_2, x_3) = (1, 1, 1)$; however, according to the MNAM, $f_1(0) + f_2(2) + f_3(1) = 3.4$, which is less than $f_1(1) + f_2(1) + f_3(1) = 3.9$. Therefore, based on the MNAM, for the person with (0,1,1), if the applicant did not pay for one month and received (1,1,1), then he or she should wait and pay one payment in an additional month to achieve (0,2,1) for a higher credit score (lower probability of default). Clearly, the fairness of this situation has been violated.

Table 3. Function values for x_1, x_2, x_3 by the MNAM in the GMSC dataset. f is weakly monotonic with respect to x_3 over x_2 and x_2 over x_1. Individual and weak pairwise monotonicity are preserved.

$f \backslash x$	0	1	2
f_1	0	0.8	1.0
f_2	0	1.4	1.7
f_3	0	1.7	2.2

We then examine the result of the MGNAM. We are interested in knowing if delinquency features can be separated additively. We plot the marginal probability of default as a function of $x_1 - x_3$ in Figure 2. The presence of DMEs is evident. By Proposition 3.4, we cannot separate these three features additively and therefore group them together. The values of $f_q(x_1, x_2, x_3)$ calculated by MGNAM are shown in Table 4. The table provides confidence to model users by verifying all monotonicity is achieved. It should be emphasized that without satisfying monotonicity, even the most accurate ML model will not be accepted. Furthermore, the transparent nature of the MGNAM makes it easier to verify conceptual soundness and fairness, which are difficult to achieve with black-box machine learning models.

5.2. Criminal Justice - COMPAS

The COMPAS scoring system was developed to predict recidivism risk and has been scrutinized for its racial bias

[1] https://www.kaggle.com/c/GiveMeSomeCredit/overview

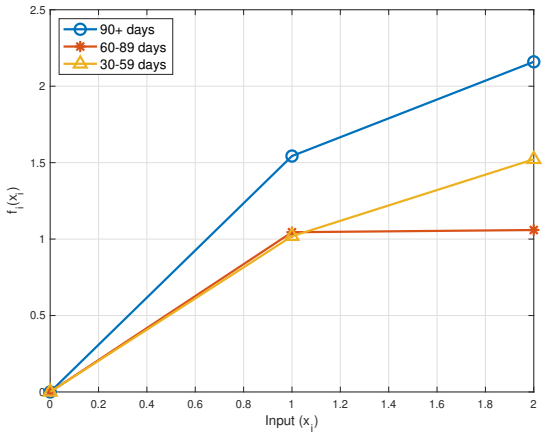

Figure 1. Comparison of functions associated with the number of past dues in different periods by the NAM for the GMSC dataset. Monotonicity is violated between 30-59 days and 60-89 days.

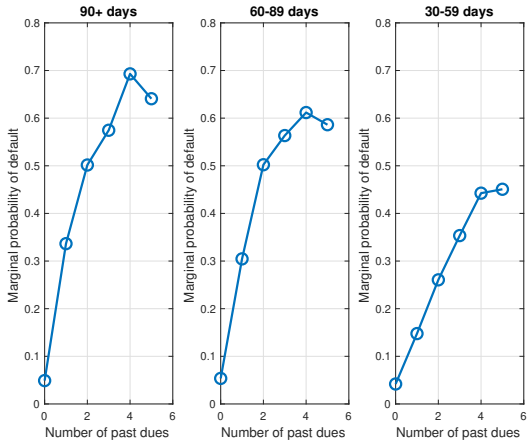

Figure 2. Marginal probability of defaults with respect to $x_1 - x_3$ in the GMSC dataset. Diminishing marginal effects are observed.

(Angwin et al., 2016; Dressel & Farid, 2018; Tan et al., 2018). In 2016, ProPublica published recidivism data for defendants in Broward County, Florida (Pro, 2016). We focus on the simplified cleaned dataset provided in (Dressel & Farid, 2018). Race and gender unfairness have been extensively studied in the past (Foulds et al., 2020; Kearns et al., 2019; 2018; Hardt et al., 2016). Our focus is on the potential unfairness associated with types of offenses. Specifically, a felony is considered more serious than a misdemeanor. Without loss of generality, assume x_1 counts the number of past misdemeanors and x_2 counts the number of past felonies. Due to this, we ask that the probability of recidivism be strongly monotonic with respect to x_2 over x_1. Criminals may consider turning a misdemeanor into a felony in the future if this strong pairwise monotonicity is violated.

Model performance is summarized in Table 5. The performance of all methods is similar. In this regard, algorithmic fairness is more important than accuracy when it comes to the dataset.

Next, we evaluate conceptual soundness and fairness. For simplicity's sake, we restrict ourselves to a maximum of three charges per type. Regarding the architecture of the MGNAM, the diminishing marginal effect is clearly observed for the felony in Figure 3, therefore we should group the felony and misdemeanor together, based on Proposition 3.4. Due to the fact that there are only two features in the group, function values are calculated and compared in tables 6. For small values of x_1 and x_2, functions behave reasonably in the NAM. For larger values, it immediately violates pairwise monotonicity. The individual monotonicity of x_2 is violated when the value of x_1 is fixed. Furthermore, the function contribution is only 0.37 when there are three past felonies ($x_2 = 3$), whereas the function value is 0.65 when there is one felony and one misdemeanor ($x_1 = x_2 = 1$). Compared to the first case, the value is almost doubled, which is a serious violation. Then, we evaluate the MNAM. Both individual monotonicity and weak pairwise monotonicity are satisfied. But when statistical interactions are involved for large x, the strong pairwise monotonicity is violated. Consider the example of $(x_1, x_2) = (0, 2)$ which should be punished more severely than $(1, 1)$. However, according to the MNAM, the value of the function at $(0, 2)$ is 0.37, which is less than the value at $(1, 1)$ as 0.50. Consequently, a person who commits one felony and one misdemeanor will be punished more severely than a person who commits two felonies. There is a seri-

Table 4. Function values for x_1, x_2, x_3 by the MGNAM in the GMSC dataset. f is strongly monotonic with respect to x_3 over x_2 and x_2 over x_1. Monotonicity is preserved.

$x_3 = 0$			
$x_1 \backslash x_2$	0	1	2
0	0	1.7	2.3
1	1.7	2.3	2.8
2	2.3	2.8	3.2
$x_3 = 1$			
$x_1 \backslash x_2$	0	1	2
0	2.2	2.7	3.2
1	2.7	3.1	3.5
2	3.1	3.5	3.7
$x_3 = 2$			
$x_1 \backslash x_2$	0	1	2
0	3.1	3.4	3.7
1	3.4	3.6	3.8
2	3.6	3.8	3.9

Table 5. Model performance of the COMPAS dataset. All ML models perform similarly.

MODEL/METRICS	CLASSIFICATION ERROR	AUC
FCNN	33.8%	71.9%
NAM	34.1%	71.8%
MNAM	33.5%	71.7%
MGNAM	34.3%	71.9%

ous violation of the principle of fairness in this situation. Additionally, if someone with one felony commits another crime, he or she may consider it to be a felony rather than a misdemeanor, leading to difficulties in society. In our model, this issue has been avoided, since the value of the function at $(0, 2)$ is 0.54, which is larger than 0.53 at $(1, 1)$. There are many other similar examples of violations. In the absence of such strong pairwise monotonicity, the algorithm should not be used.

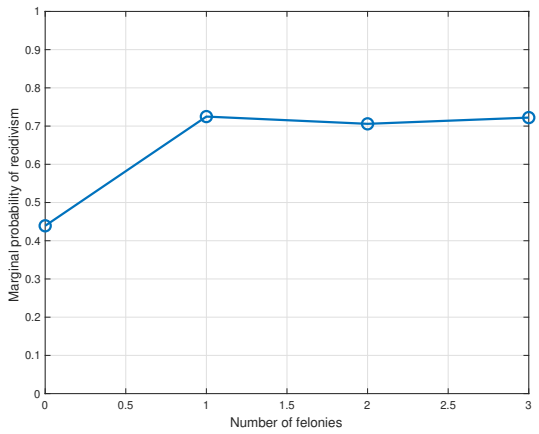

Figure 3. Marginal probability of recidivism with respect to the number of felonies in the COMPAS dataset. The diminishing marginal effect is observed.

5.3. Healthcare - Heart Failure Clinical Records

This dataset (Ahmad et al., 2017; Chicco & Jurman, 2020) contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features. This study aims to predict the survival of patients suffering from heart failure. Conceptual soundness is a very important aspect of health datasets. With limited dataset, machine learning models are very easy to overfit, which can be mitigated by imposing constraints. In the case that one needs to determine the priority of patients, then fairness is also a very important factor. For this dataset, we focus on four features: smoking, anemia, high blood pressure, and diabetes. Without loss of generality, we denote them as x_1, x_2, x_3, and x_4. Anemia, high blood pressure, and diabetes are considered to be more serious health risks than smoking. Thus, f should be monotonic with respect to $x_2 - x_4$ over x_1. Due to the fact that they are all binary features, strong monotonicity is the same as weak monotonicity, by Lemma 3.6.

Table 6. Function values for x_1, x_2 by the MNAM and MGNAM of the COMPAS dataset. There are multiple violations of monotonicity for the NAM, for example, between $(2, 2)$ and $(2, 3)$, and between $(0, 3)$ and $(1, 1)$. Violations are also observed for the MNAM, for example, between $(0, 2)$ and $(1, 1)$. The MGNAM preserves monotonicity.

MGNAM				
$x_1 \backslash x_2$	0	1	2	3
0	0	0.35	0.54	0.56
1	0.21	0.53	0.56	0.56
2	0.49	0.55	0.56	0.56
3	0.55	0.56	0.56	0.56
NAM				
$x_1 \backslash x_2$	0	1	2	3
0	0	0.41	0.40	**0.37**
1	0.24	**0.65**	0.65	0.62
2	0.32	0.72	**0.72**	**0.69**
3	0.33	0.74	0.73	0.70
MNAM				
$x_1 \backslash x_2$	0	1	2	3
0	0	0.33	**0.37**	0.37
1	0.17	**0.50**	0.54	0.54
2	0.19	0.53	0.57	0.57
3	0.20	0.53	0.57	0.57

Table 7. Model performance of the heart dataset. All ML models perform similarly.

MODEL/METRICS	CLASSIFICATION ERROR	AUC
FCNN	20.3%	87.0%
NAM	18.9%	89.8%
MGNAM	17.6%	90.6%

A summary of the results is provided in Table 7. Since the NAM performs similarly to the FCNN and features associated with pairwise monotonicity are only binary, we do not consider interactions, and the MGNAM coincides with the MNAM. The MGNAM also has a similar level of accuracy.

Next, we evaluate conceptual soundness and fairness. For blood and diabetes in the NAM, both individual and pairwise monotonicity are violated, as shown in Figure 4 and Figure 5. This problem has been avoided by MGNAM. According to the NAM, high blood pressure and diabetes are actually beneficial for survival. Furthermore, smoking is more dangerous than both of them.

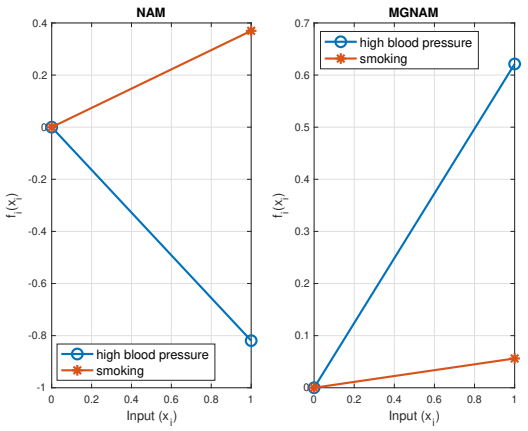

Figure 4. Comparison of blood and smoking functions by the NAM and the MGNAM for the heart dataset. The individual and pairwise monotonicity are both violated by the NAM.

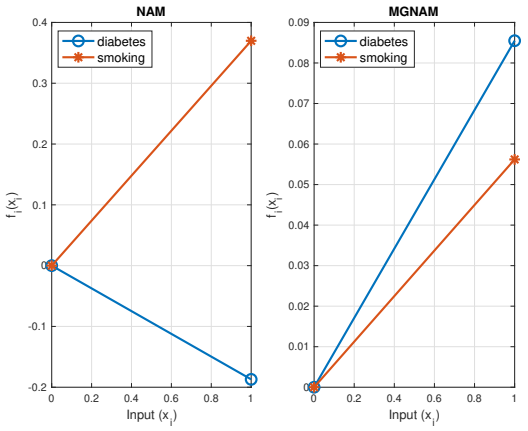

Figure 5. Comparison of diabetes and smoking functions by the NAM and the MGNAM for the diabetes dataset. The individual and pairwise monotonicity are both violated by the NAM.

6. Related work

Monotonic Models: Most of previous work (Yanagisawa et al., 2022; Liu et al., 2020; Milani Fard et al., 2016; You et al., 2017) focus on individual monotonicity. Weak pairwise monotonicity is considered in (Chen & Ye, 2022) and strong pairwise monotonicity is considered in (Gupta et al., 2020). Our paper has considered three types of monotonicity.

Transparent Models: There has been enormous literature on designing transparent machine learning models. (Agarwal et al., 2021; Chen & Ye, 2022; Yang et al., 2021; Lou et al., 2012) starts with transparent generalized additive models. Another direction specifies neural network models based on statistical interactions (Janizek et al., 2021; Tsang et al., 2018b; 2020; 2018a). However, these approaches haven't yet included the discussion between monotonicity and transparency.

7. Conclusion

In this paper, we analyze three types of monotonicity and propose monotonic groves of neural additive models (MGNAMs) for transparency and monotonicity.

There are many avenues for future directions. First, the regularized algorithm with discretized integrals in the penalty functions in order to enforce monotonicity. It is possible to achieve high accuracy with continuous features by using a large number of points, however, certification is not yet available for three types of monotonicity. Second, there are many applications in which these integrals are appropriate when the dimensions of pairwise monotonic features are small. Nevertheless, there is the possibility of having a large collection of pairwise monotonic features in some contexts. In the future, we plan to investigate possible fast algorithms for implementing pairwise monotonicity. Third, in the spirit of neural additive models, we keep MGNAM architectures as simple as possible to preserve the transparency of models. There is, however, a possibility that some datasets will exhibit other interactions. The detection of statistical interactions will be studied in the future in the presence of three types of monotonicity. Further, conjoint measurement and multiple criteria decision analysis (Bouyssou & Pirlot, 2016; Grabisch & Labreuche, 2018) are also concerned with transparency when dealing with complex statistical interactions. The inclusion of some analysis will be of interest.

Acknowledgements

We are grateful to the reviewers for their valuable and constructive comments.

References

Propublica. compas data and analysis for "machine bias"., 2016. URL https://github.com/propublica/compas-analysis.

European union's artificial intelligence act, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Model risk management, 2021. URL https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich,

B., Caruana, R., and Hinton, G. E. Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, 34, 2021.

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., and Raza, M. A. Survival analysis of heart failure patients: A case study. *PloS one*, 12(7):e0181001, 2017.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. In *Ethics of Data and Analytics*, pp. 254–264. Auerbach Publications, 2016.

Bouyssou, D. and Pirlot, M. Conjoint measurement tools for mcdm: A brief introduction. *Multiple criteria decision analysis: State of the art surveys*, pp. 97–151, 2016.

Carlo, C. d., Bondt, M. D., and Evgeniou, T. Ai regulation is coming. *Harvard Business Review*, 2021. URL https://hbr.org/2021/09/ai-regulation-is-coming.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

Chen, D. and Ye, W. Monotonic neural additive models: Pursuing regulated machine learning models for credit scoring. In *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 70–78, 2022.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Chicco, D. and Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1):1–16, 2020.

Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Duivesteijn, W. and Feelders, A. Nearest neighbour classification with monotonicity constraints. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I 19*, pp. 301–316. Springer, 2008.

Foulds, J. R., Islam, R., Keya, K. N., and Pan, S. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921. IEEE, 2020.

Grabisch, M. and Labreuche, C. Monotone decomposition of 2-additive generalized additive independence models. *Mathematical Social Sciences*, 92:64–73, 2018.

Gupta, M., Louidor, E., Mangylov, O., Morioka, N., Narayan, T., and Zhao, S. Multidimensional shape constraints. In *International Conference on Machine Learning*, pp. 3918–3928. PMLR, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Hastie, T. J. Generalized additive models. In *Statistical models in S*, pp. 249–307. Routledge, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Janizek, J. D., Sturmfels, P., and Lee, S.-I. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22 (104):1–54, 2021.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572. PMLR, 2018.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 100–109, 2019.

Liu, X., Han, X., Zhang, N., and Liu, Q. Certified monotonic neural networks. *Advances in Neural Information Processing Systems*, 33:15427–15438, 2020.

Lou, Y., Caruana, R., and Gehrke, J. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 150–158, 2012.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623–631, 2013.

Milani Fard, M., Canini, K., Cotter, A., Pfeifer, J., and Gupta, M. Fast and flexible monotonic functions with ensembles of lattices. *Advances in neural information processing systems*, 29, 2016.

Potharst, R. and Feelders, A. J. Classification trees for problems with monotonicity constraints. *ACM SIGKDD Explorations Newsletter*, 4(1):1–10, 2002.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Sorokina, D., Caruana, R., Riedewald, M., and Fink, D. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, pp. 1000–1007, 2008.

Tan, S., Caruana, R., Hooker, G., and Lou, Y. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 303–310, 2018.

Tsang, M., Cheng, D., and Liu, Y. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=ByOfBggRZ.

Tsang, M., Liu, H., Purushotham, S., Murali, P., and Liu, Y. Neural interaction transparency (nit): Disentangling learned interactions for improved interpretability. *Advances in Neural Information Processing Systems*, 31, 2018b.

Tsang, M., Rambhatla, S., and Liu, Y. How does this interaction affect me? interpretable attribution for feature interactions. *Advances in neural information processing systems*, 33:6147–6159, 2020.

Tshitoyan, V. Simple neural network, 2023. URL https://github.com/vtshitoyan/simpleNN.

Yanagisawa, H., Miyaguchi, K., and Katsuki, T. Hierarchical lattice layer for partially monotone neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Yang, Z., Zhang, A., and Sudjianto, A. Gami-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192, 2021.

You, S., Ding, D., Canini, K., Pfeifer, J., and Gupta, M. Deep lattice networks and partial monotonic functions. *Advances in neural information processing systems*, 30, 2017.

A. Appendix

A.1. Proof

Proof of Lemma 2.8. We partition $\mathbf{x} = (x_\beta, x_\gamma, x_\delta, \mathbf{x}_\neg)$, based on Definition 2.5, we have

$$f(x_\beta, x_\gamma + c, x_\delta, \mathbf{x}_\neg) \leq f(x_\beta + c, x_\gamma, x_\delta, \mathbf{x}_\neg), \tag{17}$$
$$f(x_\beta, x_\gamma, x_\delta + c, \mathbf{x}_\neg) \leq f(x_\beta, x_\gamma + c, x_\delta, \mathbf{x}_\neg), \tag{18}$$
$$\forall x_\beta, x_\gamma, x_\delta, \mathbf{x}_\neg, \forall c \in \mathbb{R}^+. \tag{19}$$

These imply that

$$f(x_\beta, x_\gamma, x_\delta + c, \mathbf{x}_\neg) \leq f(x_\beta + c, x_\gamma, x_\delta, \mathbf{x}_\neg). \tag{20}$$

Thus, we conclude. □

Proof of Lemma 3.1. If f is weakly monotonic with respect to x_β over x_γ and x_γ over x_δ, we have

$$f'_\beta(x) \geq f'_\gamma(x), f'_\gamma(x) \geq f'_\delta(x), \ \forall x \in \mathbb{R}. \tag{21}$$

Therefore, we have $f'_\beta(x) \geq f'_\delta(x), \forall x \in \mathbb{R}$. □

Proof of Proposition 3.4. f has the form (9), f is differentiable, and is strongly monotonic with respect to x_β over x_γ, therefore

$$\min_x f'_\beta(x) \geq \max_x f'_\gamma(x) \geq 0. \tag{22}$$

where we have assumed monotonically increasing without loss of generality in the content. Now at x^*, $f'_\beta(x^*) = 0$. Therefore, $f'_\gamma(x) = 0, \forall x$, and $f_\gamma(x)$ is a constant function. □

Proof of Lemma 3.6. We partition $\mathbf{x} = (x_\beta, x_\gamma, \mathbf{x}_\neg)$, where x_β and x_γ are binary. Suppose f is weakly monotonic with respect to x_β over x_γ, then we have

$$f(0, 1, \mathbf{x}_\neg) \leq f(1, 0, \mathbf{x}_\neg). \tag{23}$$

Note this coincides with the inequality required for strong pairwise monotonicity for the binary feature. Strong pairwise monotonicity implies weak pairwise monotonicity based on Lemma 2.7.

□

A.2. Remarks about Architectures of MGNAMs

Remark A.1. Additional consideration should be given to the case in which there is a mixture of strong and weak pairwise interactions. Suppose f is strongly monotonic with respect to x_δ over x_η, we group them together as $\mathbf{x}_\beta$. Consider the case for $\mathbf{x} = (\mathbf{x}_\beta, x_\gamma)$, where f is weakly monotonic with respect to x_η in $\mathbf{x}_\beta$ over x_γ, then we shouldn't separate $\mathbf{x}_\beta$ and x_γ as there will be some unfair comparisons. More specifically, $\mathbf{x}_\beta$ can take different choices of values, whereas x_γ is only one-dimensional. It follows that if there is strong pairwise monotonicity involved, then all pairwise related features should be grouped together. As a concrete example, let x_δ count the number of past dues with 60+ days within one year, x_η count the number of past dues with 30-59 days within one year, and x_γ counts the number of past dues with 30-59 days one year ago. As there is strong pairwise monotonicity between x_δ and x_η, we group them together. Suppose now we take the additive form that

$$f(\mathbf{x}) = g(x_\delta, x_\eta) + h(x_\gamma), \tag{24}$$

for some differentiable functions g and h. The weak pairwise monotonicity between x_η and x_γ would requires that

$$\frac{\partial}{\partial x_\eta} g(x_\delta, y) \geq \frac{\partial}{\partial x_\gamma} h(y), \ \forall y, x_\delta. \tag{25}$$

Note g is impacted by values of x_δ and h is not. This is inconsistent with our intention and is an unfair comparison.

A.3. Empirical Examples

For all our experiments, the dataset is randomly partitioned into 75% training and 25% test sets. All neural networks contain 1 hidden layer with 2 units, logistic activation, and no regulation. We monitored the model selection cross-validation empirical results and observe no obvious improvement in accuracy based on in/out-of-sample results, except for the healthcare dataset due to insufficient data. We tested models with up to 20 hidden units and two hidden layers. No obvious improvement in accuracy was observed. Additionally, we checked existing literature or public codes online; our accuracy is comparable. For accuracy, we check classification errors and the area under the curve (AUC). The code is built and modified based on (Tshitoyan, 2023).

A.3.1. Finance - Credit Scoring

A popularly used dataset is the Kaggle credit score dataset [2]. For simplicity, data with missing variables are removed. Past dues greater than four times are truncated. Further careful data cleanings could potentially improve model performance but are not the primary concern of this paper. Among the total 120969 observations, 8,357 (6.95$\%$) relate to the cardholders with default payments. This shows that the data are seriously imbalanced. The dataset contains 10 features as explanatory variables:

- x_1: Total balance on credit cards and personal lines of credit except for real estate and no installment debt such as car loans divided by the sum of credit limits
- x_2: Age of borrower in years
- x_3: Number of times borrower has been 30-59 days past due but no worse in the last 2 years
- x_4: Monthly debt payments, alimony, and living costs divided by monthly gross income
- x_5: Monthly income
- x_6: Number of open loans (installments such as car loan or mortgage) and lines of credit (e.g., credit cards)
- x_7: Number of times borrower has been 90 days or more past due
- x_8: Number of mortgage and real estate loans including home equity lines of credit
- x_9: Number of times borrower has been 60-89 days past due but no worse in the last 2 years
- x_{10}: Number of dependents in the family, excluding themselves (spouse, children, etc.)
- y: Client's behavior; 1 = Person experienced 90 days past due delinquency or worse

The feature age is further excluded to avoid potential discrimination.

A.3.2. Criminal Justice - COMPAS

COMPAS is a proprietary score developed to predict recidivism risk, which is used to guide bail, sentencing, and parole decisions. It has been criticized for racial bias(Angwin et al., 2016; Dressel & Farid, 2018; Tan et al., 2018). A report published by ProPublica in 2016 provided recidivism data for defendants in Broward County, Florida (Pro, 2016). We focus on the simplified cleaned dataset provided in (Dressel & Farid, 2018). Three thousand and fifty-one (45%) of the 7,214 observations committed a crime within two years. This study uses a binary response variable, recidivism, as the response variable. The dataset here contains nine features, which were selected after some feature selection was conducted.

- x_1: Races include White (Caucasian), Black (African American), Hispanic, Asian, Native American, and Others
- x_2: Sex, male or female
- x_3: Age

[2] https://www.kaggle.com/c/GiveMeSomeCredit/overview

- x_4: Total number of juvenile felony criminal charges
- x_5: Total number of juvenile misdemeanor criminal charges
- x_6: Total number of non-juvenile criminal charges
- x_7: A numeric value corresponding to the specific criminal charge
- x_8: An indicator of the degree of the charge: misdemeanor or felony
- x_9: An numeric value between 1 and 10 corresponds to the recidivism risk score generated by COMPAS software (a small number corresponds to a low risk, and a larger number corresponds to a high risk)
- y: Whether the defendant recidivated two years after the previous charge

To avoid discrimination, we further exclude races and sexes. The COMPAS score is also excluded as it is not the focus of this study and is correlated with other features, making its interpretation more difficult. As there are too few samples, we truncate the number of juveniles exceeding three. Otherwise, if monotonicity is requested, NN functions will become flat, which is not useful.

A.3.3. Healthcare - Heart Failure Clinical Records

This dataset (Ahmad et al., 2017; Chicco & Jurman, 2020) focuses on the prediction of patients' survival with heart failure in 2015. In total, there are 299 patients. The concept of fairness may be relevant here, for example, if doctors are required to decide which operation should be performed first based on the patient's condition. Death is used as the response variable in this study. This dataset contains a total of 12 features.

- x_1: Age
- x_2: Anaemia, a decrease of red blood cells or hemoglobin
- x_3: High blood pressure, if the patient has hypertension
- x_4: Creatinine phosphokinase
- x_5: If the patient has diabetes
- x_6: Ejection fraction, percentage of blood leaving the heart at each contraction
- x_7: Platelets in the blood (kiloplatelets/mL)
- x_8: Sex
- x_9: Level of serum creatinine in the blood (mg/dL)
- x_{10}: Level of serum sodium in the blood (mEq/L)
- x_{11}: If the patient smokes or not
- x_{12}: Time, follow-up period (days)
- y: Death event, if the patient deceased during the follow-up period