Chris Lin, Ian Covert, Su-In Lee

On the Robustness of Removal-Based Feature Attributions

Mateusz Krzyziński

MI^2 Research Seminar

October 23rd, 2023

About authors

Authors: Chris Lin[1], Ian Covert[1], Su-In Lee

[1] equal contribution

About authors

Authors: Chris Lin[1], Ian Covert[1], Su-In Lee

- From AI for bioMedical Sciences (AIMS) lab led by Prof. Su-In Lee.
- Paper accepted at NeurIPS 2023. Unfortunately reviews are not public yet.

[1] equal contribution

Feature attributions

Feature attribution explanation methods assign importance scores to input features of the model of interest.

[2] Sidequest: which methods do not?

[3] We discussed their stability a week ago.

Feature attributions

Feature attribution explanation methods assign importance scores to input features of the model of interest.

Most of them[2] can be categorized into two groups:

- gradient-based[3] – compute gradients of model predictions with respect to input features,
- removal-based – *remove* input features to quantify their impact.

[2] Sidequest: which methods do not?

[3] We discussed their stability a week ago.

Feature attributions

Feature attribution explanation methods assign importance scores to input features of the model of interest.

Most of them[2] can be categorized into two groups:

- gradient-based[3] – compute gradients of model predictions with respect to input features,
- removal-based – *remove* input features to quantify their impact.

The paper under discussion focuses on the robustness of removal-based methods, for which there are fewer results.

[2] Sidequest: which methods do not?
[3] We discussed their stability a week ago.

Motivation

Feature attributions are vulnerable to adversarial attacks, and even without an adversary appear unstable under small changes to the input or model. (...)

Motivation

Feature attributions are vulnerable to adversarial attacks, and even without an adversary appear unstable under small changes to the input or model. (...)

When can we guarantee that feature attributions are robust to small changes in the input or small changes in the model?

Contributions

1. Theoretical analysis (and guarantees) for the robustness of:
   - model predictions with the removal of arbitrary feature subsets (under both input and model perturbations),
   - techniques of summarizing model predictions into feature attributions,

Contributions

1. Theoretical analysis (and guarantees) for the robustness of:
    - model predictions with the removal of arbitrary feature subsets (under both input and model perturbations),
    - techniques of summarizing model predictions into feature attributions,
    - unified attribution explanations (by combining results above).

Contributions

1. Theoretical analysis (and guarantees) for the robustness of:
   - model predictions with the removal of arbitrary feature subsets (under both input and model perturbations),
   - techniques of summarizing model predictions into feature attributions,
   - unified attribution explanations (by combining results above).
2. Empirical validation of theoretical results and demonstration of their practical implications.

Notation

- $f : \mathbb{R}^d \mapsto \mathbb{R}$ – the model whose predictions we want to explain
- $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ – features

Notation

- $f : \mathbb{R}^d \mapsto \mathbb{R}$ – the model whose predictions we want to explain
- $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ – features
- for $S \subseteq [d] \quad \boldsymbol{x}_S = \{\boldsymbol{x}_i : i \in S\}$ – the feature subset
- $\bar{S} = [d] \setminus S$ – the set complement

Notation

- $f : \mathbb{R}^d \mapsto \mathbb{R}$ – the model whose predictions we want to explain
- $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ – features
- for $S \subseteq [d] \quad \boldsymbol{x}_S = \{\boldsymbol{x}_i : i \in S\}$ – the feature subset
- $\bar{S} = [d] \setminus S$ – the set complement
- $p(\boldsymbol{x})$ – the data distribution with support on $\mathcal{X} \subseteq \mathbb{R}^d$

Notation

- $f : \mathbb{R}^d \mapsto \mathbb{R}$ – the model whose predictions we want to explain
- $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_d)$ – features
- for $S \subseteq [d] \quad \boldsymbol{x}_S = \{\boldsymbol{x}_i : i \in S\}$ – the feature subset
- $\bar{S} = [d] \setminus S$ – the set complement
- $p(\boldsymbol{x})$ – the data distribution with support on $\mathcal{X} \subseteq \mathbb{R}^d$
- $||g||_p \equiv (\int |g(x)|^p dx)^{1/p}$ – the L^p norm
  for a function $g : \mathbb{R}^d \mapsto \mathbb{R}$ (the integral is taken over $\mathbb{R}^d$)
- $||g||_{p,\mathcal{X}'}$ denotes the same integral taken over the domain $\mathcal{X}' \subseteq \mathbb{R}^d$.

Removal-based feature attributions as the framework

Intuition: Collect predictions after removing subsets of features and summarize how each feature affects the model's prediction.

Removal-based feature attributions as the framework

Intuition: Collect predictions after removing subsets of features and summarize how each feature affects the model's prediction.

Examples of methods: Shapley values, LIME, leave-one-out (occlusion analysis), Banzhaf values, ...

Removal-based feature attributions as the framework

Intuition: Collect predictions after removing subsets of features and summarize how each feature affects the model's prediction.

Examples of methods: Shapley values, LIME, leave-one-out (occlusion analysis), Banzhaf values, ...

There are two choices to be made:

- how feature information is removed,
- how the summarizing is done.

Choice 1: Feature removal

The prediction after feature removal

The prediction given partial input (the set of observed values x_S) is

$$f(x_S) := \mathbb{E}_{q(\boldsymbol{x}_{\bar{S}})}\left[f(x_S, \boldsymbol{x}_{\bar{S}})\right] = \int f(x_S, x_{\bar{S}}) q(x_{\bar{S}}) dx_{\bar{S}}.$$

This is averaging the prediction over a distribution $q(\boldsymbol{x}_{\bar{S}})$ for the removed feature values.

Choice 1: Feature removal

The prediction after feature removal

The prediction given partial input (the set of observed values x_S) is

$$f(x_S) := \mathbb{E}_{q(\boldsymbol{x}_{\bar{S}})}\left[f(x_S, \boldsymbol{x}_{\bar{S}})\right] = \int f(x_S, x_{\bar{S}})q(x_{\bar{S}})dx_{\bar{S}}.$$

This is averaging the prediction over a distribution $q(\boldsymbol{x}_{\bar{S}})$ for the removed feature values.

Three options for $q(\boldsymbol{x}_{\bar{S}})$ are considered:

- $q(\boldsymbol{x}_{\bar{S}}) \equiv \delta_{b_{\bar{S}}}$ (baseline values) setting held-out features to their values from some baseline input $b \in \mathbb{R}^d$,
- $q(\boldsymbol{x}_{\bar{S}}) = p(\boldsymbol{x}_{\bar{S}})$ (marginal distribution),
- $q(\boldsymbol{x}_{\bar{S}}) = p(\boldsymbol{x}_{\bar{S}} \mid x_S)$ (conditional distribution).

Choice 2: Summary technique

Problem: How to use the obtained predictions with partial information to calculate feature attributions $\phi(f, x) = [\phi_1(f, x), \ldots, \phi_d(f, x)] \in \mathbb{R}^d$?

We have $f(x_S)$ for many different $S \subseteq [d]$ (even up to 2^d results). Which ones to use and how to aggregate them?

Choice 2: Summary technique

Problem: How to use the obtained predictions with partial information to calculate feature attributions $\phi(f, x) = [\phi_1(f, x), \dots, \phi_d(f, x)] \in \mathbb{R}^d$?

We have $f(x_S)$ for many different $S \subseteq [d]$ (even up to 2^d results). Which ones to use and how to aggregate them?

Option 1: leave-one-out

$$\phi_i(f, x) = f(x) - f(x_{[d] \setminus \{i\}})$$

Choice 2: Summary technique

Problem: How to use the obtained predictions with partial information to calculate feature attributions $\phi(f, x) = [\phi_1(f, x), \ldots, \phi_d(f, x)] \in \mathbb{R}^d$?

We have $f(x_S)$ for many different $S \subseteq [d]$ (even up to 2^d results). Which ones to use and how to aggregate them?

Option 1: leave-one-out

$$\phi_i(f, x) = f(x) - f(x_{[d] \setminus \{i\}})$$

Option 2: Shapley values

$$\phi_i(f, x) = \frac{1}{d} \sum_{S \subseteq [d] \setminus \{i\}} \binom{d-1}{|S|}^{-1} \left(f(x_{S \cup \{i\}}) - f(x_S)\right)$$

Choice 2: Summary technique cont'd

Calculating attributions based on predictions for feature subsets

The attributions for each method can be calculated by applying a linear operator $A \in \mathbb{R}^{d \times 2^d}$ to a vector $v \in \mathbb{R}^{2^d}$ representing the predictions with each feature set, i.e.,

$$\phi(f, x) = Av,$$

where v is defined as $v_S = f(x_S)$ for each $S \subseteq [d]$.

Choice 2: Summary technique cont'd

Calculating attributions based on predictions for feature subsets

The attributions for each method can be calculated by applying a linear operator $A \in \mathbb{R}^{d \times 2^d}$ to a vector $v \in \mathbb{R}^{2^d}$ representing the predictions with each feature set, i.e.,

$$\phi(f, x) = Av,$$

where v is defined as $v_S = f(x_S)$ for each $S \subseteq [d]$.

Summary	**Method**	A_{iS} $(i \in S)$	A_{iS} $(i \notin S)$
Leave-one-out	Occlusion	$\mathbb{I}\{\lvert S\rvert = d\}$	$-\mathbb{I}\{\lvert S\rvert = d-1\}$
Shapley value	SHAP	$\frac{(\lvert S\rvert-1)!(d-\lvert S\rvert)!}{d!}$	$-\frac{\lvert S\rvert!(d-\lvert S\rvert-1)!}{d!}$
Banzhaf value	Banzhaf	$1/2^{d-1}$	$-1/2^{d-1}$
Mean when included	RISE	$1/2^{d-1}$	0
Weighted least squares	LIME	depends on implementation choices	

Goals

Question 1

Whether $\|\phi(f, x) - \phi(f, x')\|$
is controlled by $\|x - x'\|$ (**input perturbation**)?

Goals

Question 1

Whether $\|\phi(f, x) - \phi(f, x')\|$
is controlled by $\|x - x'\|$ (**input perturbation**)?

Question 2

Whether $\|\phi(f, x) - \phi(f', x)\|$
is controlled by $\|f - f'\|$ (**model perturbation**)?

Assumptions

Assumption 1

The model f is globally L-Lipschitz continuous, i.e,

$$\exists L \geq 0 \;\; \forall x, x' \in \mathbb{R}^d \quad |f(x) - f(x')| \leq L \cdot \|x - x'\|_2.$$

Assumptions

Assumption 1

The model f is globally L-Lipschitz continuous, i.e,

$$\exists L \geq 0 \;\; \forall x, x' \in \mathbb{R}^d \quad |f(x) - f(x')| \leq L \cdot \|x - x'\|_2.$$

Assumption 2

The model f has bounded predictions, i.e.,

$$\exists B \in \mathbb{R} \;\; \forall x \in \mathbb{R}^d \quad |f(x)| \leq B.$$

Theoretical contents

- Prediction robustness to input perturbations.
- Prediction robustness to model perturbations.
- Summary technique robustness to predictions.

Theoretical contents

- Prediction robustness to input perturbations.
- Prediction robustness to model perturbations.
- Summary technique robustness to predictions.
- **Feature attribution robustness.**

Prediction robustness to input perturbations

Lemma A1

When removing features using either the **baseline** or **marginal** approaches, the prediction function $f(\boldsymbol{x}_S)$ for any feature set $\boldsymbol{x}_S$ is L-Lipschitz continuous:

$$|f(x_S) - f(x'_S)| \leq L \cdot \|x_S - x'_S\|_2 \quad \forall x_S, x'_S.$$

Prediction robustness to input perturbations

Lemma A1

When removing features using either the **baseline** or **marginal** approaches, the prediction function $f(\boldsymbol{x}_S)$ for any feature set $\boldsymbol{x}_S$ is L-Lipschitz continuous:

$$|f(x_S) - f(x'_S)| \leq L \cdot \|x_S - x'_S\|_2 \quad \forall x_S, x'_S.$$

Lemma A2

When removing features using the **conditional** approach, the prediction function $f(\boldsymbol{x}_S)$ for a feature set $\boldsymbol{x}_S$ satisfies:

$$|f(x_S) - f(x'_S)| \leq L \cdot \|x_S - x'_S\|_2 + 2B \cdot d_{TV}\Big(p(\boldsymbol{x}_{\bar{S}} \mid x_S), p(\boldsymbol{x}_{\bar{S}} \mid x'_S)\Big).$$

$$d_{TV}\Big(p(\boldsymbol{x}_{\bar{S}} \mid x_S), p(\boldsymbol{x}_{\bar{S}} \mid x'_S)\Big) = \frac{1}{2}\Big\|p(\boldsymbol{x}_{\bar{S}} \mid x_S) - p(\boldsymbol{x}_{\bar{S}} \mid x'_S)\Big\|_1$$

Prediction robustness to input perturbations cont'd

Additional assumption

We assume that there exists a constant M
such that for all $S \subseteq [d]$

$$d_{TV}\Big(p(\boldsymbol{x}_{\bar{S}} \mid x_S), p(\boldsymbol{x}_{\bar{S}} \mid x'_S)\Big) \leq M \cdot \|x_S - x'_S\|_2 \quad \forall x_S, x'_S.$$

For example, $M = 0$ for independent features.

Prediction robustness to input perturbations cont'd

Additional assumption

We assume that there exists a constant M
such that for all $S \subseteq [d]$

$$d_{TV}\Big(p(\boldsymbol{x}_{\bar{S}} \mid x_S), p(\boldsymbol{x}_{\bar{S}} \mid x'_S)\Big) \leq M \cdot \|x_S - x'_S\|_2 \quad \forall x_S, x'_S.$$

For example, $M = 0$ for independent features.

Lemma A3

Under this additional assumption, when removing features using **conditional** approach the prediction function $f(\boldsymbol{x}_S)$ for any feature set $\boldsymbol{x}_S$ is Lipschitz continuous with constant $L + 2BM$.

Prediction robustness to model perturbations

Lemma B1

For two models $f, f' : \mathbb{R}^d \mapsto \mathbb{R}$ and a subdomain $\mathcal{X}' \subseteq \mathbb{R}^d$, the prediction functions $f(\boldsymbol{x}_S), f'(\boldsymbol{x}_S)$ for any feature set x_S satisfy

$$|f(x_S) - f'(x_S)| \leq \|f - f'\|_{\infty, \mathcal{X}'} \cdot Q_{x_S}(\mathcal{X}') + 2B \cdot \big(1 - Q_{x_S}(\mathcal{X}')\big),$$

where $Q_{x_S}(\mathcal{X}')$ is the probability of imputed samples lying in $\mathcal{X}'$ based on the distribution $q(\boldsymbol{x}_{\bar{S}})$:

$$Q_{x_S}(\mathcal{X}') = \mathbb{E}_{q(\boldsymbol{x}_{\bar{S}})}\left[\mathbb{I}\left\{(x_S, \boldsymbol{x}_{\bar{S}}) \in \mathcal{X}'\right\}\right].$$

Prediction robustness to model perturbations

Lemma B1

For two models $f, f' : \mathbb{R}^d \mapsto \mathbb{R}$ and a subdomain $\mathcal{X}' \subseteq \mathbb{R}^d$, the prediction functions $f(\boldsymbol{x}_S), f'(\boldsymbol{x}_S)$ for any feature set x_S satisfy

$$|f(x_S) - f'(x_S)| \leq \|f - f'\|_{\infty,\mathcal{X}'} \cdot Q_{x_S}(\mathcal{X}') + 2B \cdot \big(1 - Q_{x_S}(\mathcal{X}')\big),$$

where $Q_{x_S}(\mathcal{X}')$ is the probability of imputed samples lying in $\mathcal{X}'$ based on the distribution $q(\boldsymbol{x}_{\bar{S}})$:

$$Q_{x_S}(\mathcal{X}') = \mathbb{E}_{q(\boldsymbol{x}_{\bar{S}})}\left[\mathbb{I}\left\{(x_S, \boldsymbol{x}_{\bar{S}}) \in \mathcal{X}'\right\}\right].$$

Hence we have:

- $|f(x_S) - f'(x_S)| \leq \|f - f'\|_\infty$ for **baseline** or **marginal** approaches,
- $|f(x_S) - f'(x_S)| \leq \|f - f'\|_{\infty,\mathcal{X}}$ for **conditional** approach.

Summary technique robustness

Remark: We can represent attributions as $\phi(f, x) = Av$.

Summary technique robustness

Remark: We can represent attributions as $\phi(f, x) = Av$.

Summary technique robustness can be analyzed with respect to changes in v – there is no need to study robustness to input/model perturbations separately.

Summary technique robustness

Remark: We can represent attributions as $\phi(f, x) = Av$.

Summary technique robustness can be analyzed with respect to changes in v – there is no need to study robustness to input/model perturbations separately.

Lemma C1

The difference in attributions given the same summary technique A and different model outputs v, v' can be bounded as

$$\|Av - Av'\|_2 \leq \|A\|_2 \cdot \|v - v'\|_2 \qquad \text{or} \qquad \|Av - Av'\|_2 \leq \|A\|_{1,\infty} \cdot \|v - v'\|_\infty,$$

where $\|A\|_2$ is the spectral norm, and the operator norm $\|A\|_{1,\infty}$ is the square root of the sum of squared row 1-norms.

Summary technique robustness cont'd

We restrict attention to reasonable choices for an operator A. Some good properties are:

Summary technique robustness cont'd

We restrict attention to reasonable choices for an operator A. Some good properties are:

- boundedness – each feature's attribution is bounded by its smallest and largest contributions,

Summary technique robustness cont'd

We restrict attention to reasonable choices for an operator A. Some good properties are:

- boundedness – each feature's attribution is bounded by its smallest and largest contributions,
- symmetry – two features that make equal marginal contributions to all feature sets must have equal attributions,

Summary technique robustness cont'd

We restrict attention to reasonable choices for an operator A. Some good properties are:

- boundedness – each feature's attribution is bounded by its smallest and largest contributions,
- symmetry – two features that make equal marginal contributions to all feature sets must have equal attributions,
- efficiency – the attributions sum to the difference in predictions for the complete and empty sets.

Summary technique robustness cont'd

We restrict attention to reasonable choices for an operator A. Some good properties are:

- boundedness – each feature's attribution is bounded by its smallest and largest contributions,
- symmetry – two features that make equal marginal contributions to all feature sets must have equal attributions,
- efficiency – the attributions sum to the difference in predictions for the complete and empty sets.

Lemma C2

When the linear operator A satisfies the boundedness property, we have

$$\|A\|_{1,\infty} = 2\sqrt{d}.$$

Note: it applies to to leave-one-out, Shapley values and Banzhaf values.

Summary technique robustness cont'd

We restrict attention to reasonable choices for an operator A. Some good properties are:

- boundedness – each feature's attribution is bounded by its smallest and largest contributions,
- symmetry – two features that make equal marginal contributions to all feature sets must have equal attributions,
- efficiency – the attributions sum to the difference in predictions for the complete and empty sets.

Summary technique robustness cont'd

We restrict attention to reasonable choices for an operator A. Some good properties are:

- boundedness – each feature's attribution is bounded by its smallest and largest contributions,
- symmetry – two features that make equal marginal contributions to all feature sets must have equal attributions,
- efficiency – the attributions sum to the difference in predictions for the complete and empty sets.

Observation C1

Banzhaf value is the most robust among summary methods satisfying both boundedness and symmetry properties.

Summary technique robustness cont'd

We restrict attention to reasonable choices for an operator A. Some good properties are:

- boundedness – each feature's attribution is bounded by its smallest and largest contributions,
- symmetry – two features that make equal marginal contributions to all feature sets must have equal attributions,
- efficiency – the attributions sum to the difference in predictions for the complete and empty sets.

Observation C1

Banzhaf value is the most robust among summary methods satisfying both boundedness and symmetry properties.

Observation C2

Shapley value is the most robust among summary methods satisfying both boundedness and efficiency properties.

Robustness to input perturbations

Theorem 1

The robustness of removal-based explanations to **input perturbations** is given by the following meta-formula,

$$\|\phi(f,x) - \phi(f,x')\|_2 \leq g(\text{removal}) \cdot h(\text{summary}) \cdot \|x - x'\|_2,$$

where the factors for each method are defined as follows:

$$g(\text{removal}) = \begin{cases} L & \text{if removal} \in \{\textbf{baseline}, \textbf{marginal}\} \\ L + 2BM & \text{if removal} = \textbf{conditional}, \end{cases}$$

$$h(\text{summary}) = \begin{cases} 2\sqrt{d} & \text{if summary} \in \{\textbf{Shapley}, \textbf{Banzhaf}, \textbf{leave-one-out}\} \\ \sqrt{d} & \text{if summary} = \textbf{mean when included}. \end{cases}$$

Robustness to model perturbations

Theorem 2

The robustness of removal-based explanations to model perturbations is given by the following meta-formula,

$$\|\phi(f,x) - \phi(f',x)\|_2 \leq h(\text{summary}) \cdot \|f - f'\|,$$

where the functional distance and factor associated with the summary technique are defined as follows:

$$\|f - f'\| = \begin{cases} \|f - f'\|_\infty & \text{if removal} \in \{\textbf{baseline}, \textbf{marginal}\} \\ \|f - f'\|_{\infty,\mathcal{X}} & \text{if removal} = \textbf{conditional}, \end{cases}$$

$$h(\text{summary}) = \begin{cases} 2\sqrt{d} & \text{if summary} \in \{\textbf{Shapley}, \textbf{Banzhaf}, \textbf{leave-one-out}\} \\ \sqrt{d} & \text{if summary} = \textbf{mean when included}. \end{cases}$$

Validation of results for input perturbation: setup

- logistic regression
- $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$
- $\rho = 1$ (unless stated otherwise)
- $\boldsymbol{y} \sim Bern(1, \sigma(\beta^\top \boldsymbol{x}))$
- $\beta^\top = [5, 0, 3, 1]$

Validation of results for input perturbation: setup

- logistic regression
- $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$
- $\rho = 1$ (unless stated otherwise)
- $\boldsymbol{y} \sim Bern(1, \sigma(\beta^\top \boldsymbol{x}))$
- $\beta^\top = [5, 0, 3, 1]$

Lipschitz constant can be analytically computed.

Validation of results for input perturbation: results

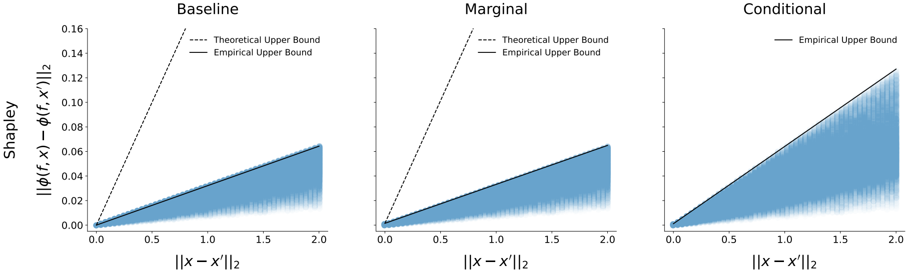

Figure: Theoretical vs. empirical upper bounds for Shapley attribution differences and various feature removal approaches.

Validation of results for input perturbation: results cont'd

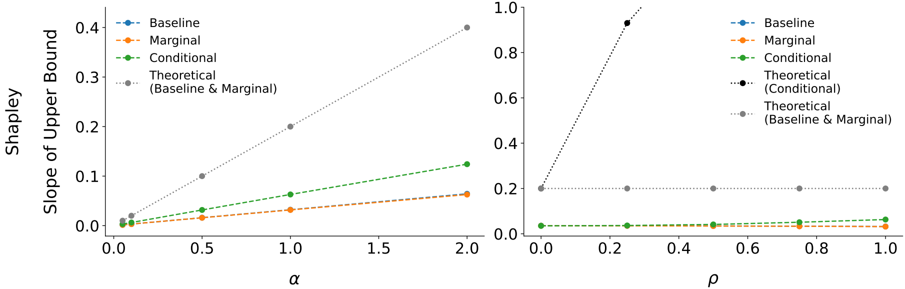

Figure: Slopes of theoretical and empirical upper bounds for Shapley attribution differences w.r.t. parameters of the experiment.

Validation of results for model perturbation: setup

- setup almost the same as before
- coefficients for the original model f: $\beta^\top = [5, 0, 3, 1]$
- coefficients for the perturbed model f': $\beta^\top = [0, 5, 3, 1]$
- with default value $\rho = 1$, $\|f - f'\|_{\infty,\mathcal{X}} = 0$

Validation of results for model perturbation: results

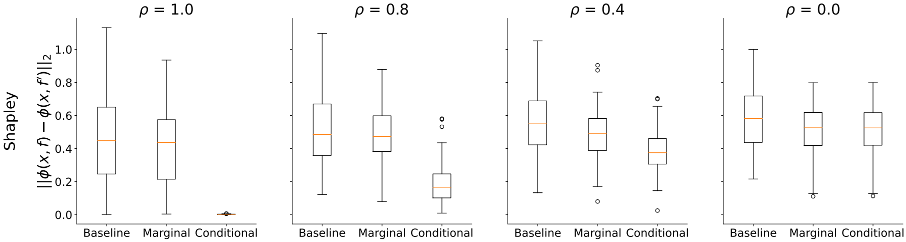

Figure: Shapley attribution differences between two logistic regression models f and f' for various feature removal approaches and rho values.

Robustness of regularized neural networks to input perturbations: setup

- datasets: UCI white wine quality, MNIST
- models: FCNs and CNNs with different weigh decays
- for MNIST only baseline feature removal approach considered
- input perturbations of different norms

Robustness of regularized neural networks to input perturbations: results

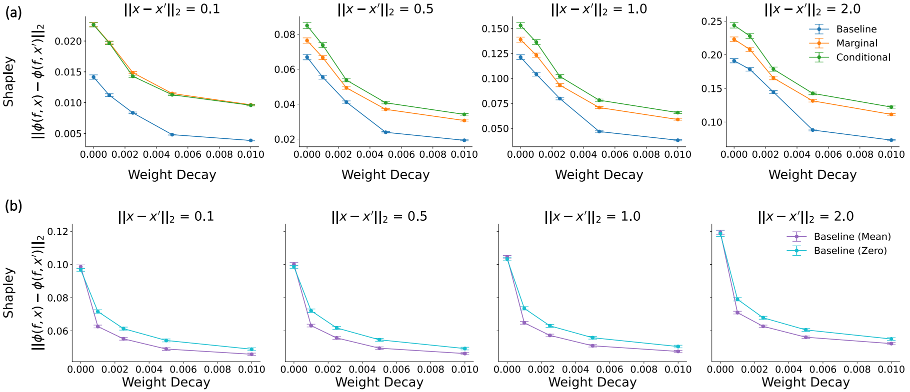

Figure: Shapley attribution differences for networks trained with increasing weight decay, various feature removal approaches and input perturbation norms. Results for: (a) the wine quality dataset, (b) MNIST.

Robustness of regularized neural networks to model perturbations: setup

- setup the same as before
- model perturbations obtained using *cascading randomization* procedure – sequentially randomization of parameters from the output to input layer

Robustness of regularized neural networks to model perturbations: results

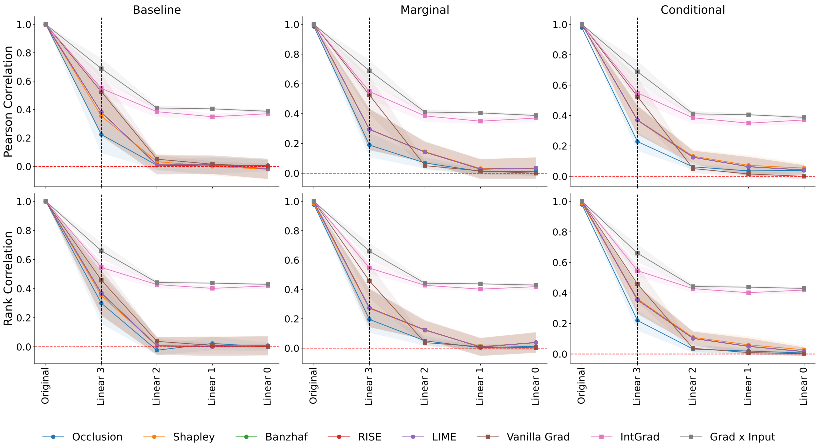

Figure: Sanity checks for attributions for the FCN trained on the wine quality dataset. Note: not only removal-based methods are considered.

Short summary

- Attribution changes are proportional to the input perturbation strength.
- Attribution changes are proportional to the functional distance between the perturbed and original models.
- The main limitation to the results is that they are conservative (worst-case bounds that are in many cases overly conservative).