Uncertainty-aware Accumulated Local Effects (UALE) for quantifying the heterogeneous effects

Anonymous Author
Anonymous Institution

Abstract

Accumulated Local Effects (ALE) is a popular approach to explainable AI that describes how a feature influences the decisions of a model on average, taking into account potential correlations between the features. However, ALE does not provide an explicit component for modeling the uncertainty of the effect, i.e., the level of heterogeneous effects behind the average explanation, which is crucial for a complete interpretation when the black-box function contains interactions between the correlated features. In this work, we propose Uncertainty-aware ALE (UALE), an extension of ALE for quantifying the heterogeneous effects. We show the importance of bin-splitting for a robust approximation of the uncertainty and we formally prove the conditions for an unbiased estimation. We finally propose a computationally-grounded algorithm for finding the optimal sequence of bins given the limited instances of the training set. We demonstrate through synthetic and real datasets, that UALE quantifies the average and heterogeneous effects correctly and approximates them robustly by optimizing the bin-splitting procedure.

1 INTRODUCTION

Recently, Machine Learning (ML) has flourished in critical domains, such as healthcare and finance. In these areas, we need a combination of accurate predictions along with meaningful explanations to support them. For this reason there is an increased interest in Explainable AI (XAI), the field that provides interpretations about the behavior of complex black-box models. XAI literature distinguishes between local and global explainability techniques (Molnar et al., 2020a). Local methods explain a specific prediction, whereas global methods explain the entire model behavior. Global methods provide a universal explanation, summarizing the various local explanations into a single interpretable outcome, usually a number or a plot. If a user wants to get a rough overview about which features are significant (feature importance) or whether a particular feature has a positive or negative effect on the output (feature effect), they should opt for a global explainability technique. On the other hand, aggregating the individual explanations for producing a concise global one is vulnerable to misinterpretations. Under strong interactions and correlations between features, the global explanation may obfuscate heterogeneous effects that exist under the hood (Herbinger et al., 2022); a phenomenon called aggregation bias (Mehrabi et al., 2021).

Feature effect (FE) (Grömping, 2020) is a fundamental category of global explainability methods. The objective of FE is to the isolate and visualize the impact of a single feature on the output. [1] FE methods suffer from aggregation bias because, often, the rationale behind the average effect is unclear. For example, a feature with zero average effect may indicate no effect on the output or, in contrast, highly positive in some cases and highly negative in others. There are three widely-used FE methods; Partial Dependence Plots (PDP)(Friedman, 2001), Marignal Plots (MP)(Apley and Zhu, 2020) and Aggregated Local Effects (ALE)(Apley and Zhu, 2020). PDP and MP have been criticized for computing erroneous effects when the input features are (highly) correlated, which is a frequent scenario in many ML cases. Therefore, ALE has been established as the state-of-the-art FE method.

However, ALE faces two crucial limitations, the first concerns ALE definition and the second one ALE approximation. Regarding the definition ALE does not inform the user about the level of uncertainty, i.e., the heterogeneous effects hidden behing the average effect due to implicit feature interactions. In contrast, in PDPs the heterogeneous

Preliminary work. Under review by AISTATS 2023. Do not distribute.

[1] Often, FE methods also isolate the combined effect of a pair of features to the output. Combinations of more than two features are not usual, because they encounter, among others, visualization difficulties.

effects can be (partially) spotted by exploring the Individual Conditional Expectations (ICE)(Goldstein et al., 2015). Regarding the approximation, i.e. the estimation usign the (limited) samples of the training set, ALE requires the additional *bin-splitting* step. Bin-splitting consists of partitioning the axis of the feature of interest in a sequence of non-overlapping intervals (regions) and estimating a constant effect in each one from the population of samples that fall inside. Specifying an approriate sequence of regions is of particular importance, since ALE's interpretation is meaningful only inside each region (Molnar, 2022). However, this crucial step has not raised the appropriate attention, making the approximation vulnerable to potential misinterpretations.

In this paper, we present Accumulated Regional Effects (ARE) an extension of ALE for measuring the uncertainty of the explanation, i.e. how certain we are that the average explanation is valid if applied to an instance drawn at random. We also provide an alternative definition of ARE based on a sequence of constant-effect regions and we define a metric of the dissimilarity of the two definitions. We formally prove that under some conditions the dissimilarity is zero and both definitions are equivalent. We, therefore, set up an optimization problem where we search for the sequence of region that minimizes the disimilarity, given enough training points iniside each region for a robust estimation. Finally, we propose a computationally-grounded algorithm for finding the optimal solution.

Contributions. The contribution of this paper are:

- ARE, a FE method that extends ALE for quantifying the heterogeneous effects (uncertainty).
- An interval-based definition of ARE and a formal proof that under some conditions it is equivalent to the initial definition.
- A principled framework for automatically extracting regions with similar effects, improving the estimation and the interpretability of ALE plots.

We provide empirical evaluation of the method in artificial and real datasets. The implementation of our method and the code for reproducing all the experiments is provided in the submission and will become publicly available upon acceptance.

2 BACKGROUND AND RELATED WORK

In this section, we describe the basic methods for FE and for uncertainty quantification, focusing on ALE definition. We, then, review the ALE approximation(Apley and Zhu, 2020; Gkolemis et al.), describing some of its vulnerabilities.

Notation. We refer to random variables (rv) using uppercase X, to simple variables with plain lowercase x and to vectors with bold $\mathbf{x}$. Often, we partition the input vector $\mathbf{x} \in \mathbb{R}^D$ to the feature of interest $x_s \in \mathbb{R}$ and the rest of the features $\mathbf{x}_c \in \mathbb{R}^{D-1}$. For convenience we denote it as $(x_s, \mathbf{x}_c)$, but we clarify that it implies the vector $(x_1, \cdots, x_s, \cdots, x_D)$. Equivalently, we denote the corresponding rv as $\mathbf{X} = (X_s, \mathbf{X}_c)$. When we refer to an instance of the training set, we use $\mathbf{x}^i = (\mathbf{x}_c^i, x_s^i)$. The black-box function is $f : \mathbb{R}^D \rightarrow \mathbb{R}$ and the FE of the s-th feature is $f^{\texttt{<method>}}(x_s)$, with $< \texttt{method} >$ indicating the particular method in use.[2]

2.1 Feature Effect Methods And ALE Definition

The three well-known feature effect methods are: PDP, MP and ALE. PDP formulates the FE of the s-th attribute as an expectation over the marginal distribution $\mathbf{X}_c$, i.e., $f^{\texttt{PDP}}(x_s) = \mathbb{E}_{\mathbf{X}_c}[f(x_s, \mathbf{X}_c)]$, whereas MP formulates it as an expectation over the conditional $\mathbf{X}_c|X_s$, i.e., $f^{\texttt{MP}}(x_s) = \mathbb{E}_{\mathbf{X}_c|x_s}[f(x_s, \mathbf{X}_c)]$. ALE defines the local effect of the s-th feature at point $x_s = z$ as $f^s(z, \mathbf{x}_c) = \frac{\partial f}{\partial x_s}(z, \mathbf{x}_c)$. All the local explanations at z are, then, weighted by the conditional distribution $p(\mathbf{x}_c|x_s = z)$ and are averaged, to produce the averaged effect $\mu(z)$. ALE is the accumulation of the averaged local effects:

$$f^{\texttt{ALE}}(x_s) = \int_{x_{s,min}}^{x_s} \underbrace{\mathbb{E}_{\mathbf{X}_c|X_s=z}\left[f^s(z, \mathbf{X}_c)\right]}_{\mu(z)} \partial z \tag{1}$$

Eq. (1) has specific advantages which gain particular value in cases of correlated input features. In these cases, PDP integrates over unrealistic instances due to the use of the marginal distribution $\mathbf{X}_c$, and MP computes aggregated effects, i.e., imputes the combined effect of sets of features to a single feature. ALE manages to resolve both issues, and is therefore the most trustable method in cases of correlated features.

2.2 Quantification Of Heterogeneous Effects.

FE methods reply to the question *what is the average (global) effect on the output, if the value of a specific feature is increased/decreased.* It comes naturally to also ask *to what extent the local effects deviate from the global explanation.* Quantifying the uncertainty of the global explanation has attracted a lot of interest. ICE and d-ICE(Goldstein et al., 2015) provide a set of curves that are plot on top-of the PDP. Both methods produce one curve for each instance of the training set; $f_i^{\texttt{ICE}}(x_s) = f(x_s, \mathbf{x}_c^i)$ for ICE and $f_i^{\texttt{d-ICE}}(x_s) = \frac{\partial f}{\partial x_s}(x_s, \mathbf{x}_c^i)$ for d-ICE. The user then visually observes if the curves are homogeneous and

[2] An extensive list of all symbols used in the paper is provided at the Appendix.

to what extent they deviate from the PDP. Some methods try to automate the aforementioned visual exploration, by grouping (d-)ICE plots into clusters (Molnar et al., 2020b; Herbinger et al., 2022; Britton, 2019). Some other approaches, like H-Statistic(Friedman and Popescu, 2008), Greenwell's interaction index(Greenwell et al., 2018) or SHAP interaction values(Lundberg et al., 2018), quantify the level of interaction between the input features, with an interaction value. A strong interaction index is an indicator for the existence of heterogeneous effects.

The aforementioned approaches are under two limitations; They either do not quantify the uncertainty of the FE directly or they are based on PDPs, and, therefore, they are subject to the failure modes of PDPs in cases of correlated features(Baniecki et al., 2021), as we will show in-depth in Section 4.1. To the best of our knowledge, no work exist so far that quantifies the heterogeneous effects based on the formulation of ALE.

2.3 ALE Approximation.

In real ML scenarios, the FE is estimated from the training set examples. Therefore, (Apley and Zhu, 2020) proposed dividing the s-th axis in K bins and estimating the local effects in each bin by evaluating the black box-function at the bin limits:

$$\hat{f}^{\texttt{ALE}}(x_s) = \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[f(z_k, \mathbf{x}^i_c) - f(z_{k-1}, \mathbf{x}^i_c))\right] \tag{2}$$

We denote as k_x the index of the bin that x_s belongs to, i.e. $k_x : z_{k_x - 1} \leq x_s < z_{k_x}$ and $\mathcal{S}_k$ is the set of training instance that lie in the k-th bin, i.e. $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x^i_s < z_k\}$. Afterwards, (Gkolemis et al.) proposed the Differential ALE (DALE) that computes the local effects on the training instances using auto-differentiation:

$$\hat{f}^{\texttt{DALE}}(x_s) = \Delta x \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} f^s(\mathbf{x}^i) \tag{3}$$

Their method has the advantages of remaining on-distribution even when bins become wider and, most importantly, allows the recomputation of the accumulated effect with different bin-splitting with near-zero computational cost.

Both approximations ask from the user to blindly decide the number of bins, denoted with K, for splitting the axis into equally-sized bins, an approach with crucial limitations. Setting K to a small value may hide fine-grain effects due to large bins and setting K to a high value leads to poor bin-effect estimations from limited samples. In general, the user may face contradictory explanations for different K without a clue to decide which one to trust.

3 Accumulated Regional Effects (ARE)

ARE extends ALE for quantifying the level of heterogeneous effects (uncertainty) and automates the bin-splitting step for robust estimations. At Section 3.1, we define ARE and at Section 3.2 we provide an alternative definition of ARE based on variable-size bins, providing important remarks and proofs for connecting it with the initial definition. In Section 3.4, we define the problem of optimal bin-splitting and we propose an algorithmic solution to the problem.

3.1 ARE: ALE With Uncertainty Quantification

We extend ALE definition of Eq. (1) with a component for quantifying the uncertainty. We denote as $\mathcal{H}(z)$ the uncertainty of the local effects at a specific point $x_s = z$ and we quantify it as the standard deviation of the local explanations:

$$\mathcal{H}(z) := \sigma(z) = \sqrt{\mathbb{E}_{\mathbf{X}_c|z}\left[(f^s(z, X_c) - \mu(z))^2\right]} \tag{4}$$

The uncertainty emerges from the natural characteristics of the experiment, i.e., the feature correlations existent in the data generating distribution and the implicit interactions of the black-box function. We also define the accumulated uncertainty at x_s, as the accumulation of the standard deviation of the local effects along the axis:

$$f^{\texttt{ALE}}_{\sigma}(x_s) = \int_{x_{s,min}}^{x_s} \sigma(z) \partial z \tag{5}$$

ARE method formulates the effect at a specific point x_s with a compact tuple that consists of the average effect and the uncertainty $(\mu(z), \sigma(z))$ and visualizes them as a continuous curve with a confidence region, i.e. $f^{\texttt{ARE}}(x_s) := f^{\texttt{ALE}}_{\mu}(x_s) \pm f^{\texttt{ALE}}_{\sigma}(x_s)$.

3.2 Interval-Based Formulation and Approximation

In real scenarios, the estimations are based on the limited instances of the training set. Estimating $\mu(z), \sigma(z)$ at the granularity of a point is impossible, because the probability of observing a sample inside the region $[x_s - h, x_s + h]$ tends to zero, when $h \to 0$. Therefore we define the regional-effect $\mu(z_1, z_2)$ and the regional-uncertainty $\sigma(z_1, z_2)$, as the summary statistics that characterize the effect and the uncertainty inside the interval $[z_1, z_2)$:

$$\mu(z_1, z_2) = \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x}_c|z}\left[f^s(z, \mathbf{x}_c)\right] \partial z \tag{6}$$

$$\sigma(z_1, z_2) = \sqrt{\frac{\int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x}_c|x_s=z}\left[(f^s(z, \mathbf{x}_c) - \mu(z))^2\right] \partial z}{z_2 - z_1}} \tag{7}$$

The regional-effect and the regional-uncertainty are the mean values of the local effects and local uncertainty inside an interval. In other words, they quantify what is the expected effect and the expected uncertainty of a sample drawn at random from $[z_1, z_2]$ given uniform distribution $\mathcal{U}(z_1, z_2)$.

If we also define as $\mathcal{Z}_s$ the sequence of $K+1$ points that partition the s-th feature axis into K variable-size intervals, i.e. $\mathcal{Z}_s = \{z_0, \dots, z_K\}$, we can redefine ARE using an interval-based formulation $\tilde{f}^{\texttt{ARE}}(x_s) := \tilde{f}^{\texttt{ALE}}_{\mu}(x_s) \pm \tilde{f}^{\texttt{ALE}}_{\sigma}(x_s)$, where:

$$\tilde{f}^{\texttt{ALE}}_{\mu}(x_s) = \sum_{k=1}^{k_x} \mu(z_{k-1}, z_k)(z_k - z_{k-1}) \tag{8}$$

$$\tilde{f}^{\texttt{ALE}}_{\sigma}(x_s) = \sum_{k=1}^{k_x} \sigma(z_{k-1}, z_k)(z_k - z_{k-1}) \tag{9}$$

3.3 Interval-Based Approximation

The mean effect (Eq. (6)) and the uncertainty (Eq. (7)) can be directly estimated from the set $\mathcal{S}_k$ of dataset instances with the s-th feature lying inside the interval $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x_s^i < z_k\}$:

$$\hat{\mu}(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[f^s(\mathbf{x}^i) \right] \tag{10}$$

$$\hat{\sigma}^2(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left(f^s(\mathbf{x}^i) - \hat{\mu}(z_1, z_2) \right)^2 \tag{11}$$

Under the assumption that the points are uniformly distributed inside the intervals Eq.(10) is an unbiased estimatior of Eq. (6). The same does not hold in the general case for Eq.(11). In the general case, $\hat{\sigma}^2(z_{k-1}, z_k)$ is an unbiased estimator of the observable standard deviation $\sigma_{obs}(z_1, z_2) = \sqrt{\frac{\int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x}_c|x_s=z}[(f^s(z,\mathbf{x}_c) - \mu(z_1,z_2))^2] \partial z}{z_2 - z_1}}$, which as we prove in Theorem 1 is an overestimation of the real regional variance.

Theorem 1. *If we define the residual $\rho(z)$ as the difference between the expected effect at x_s and the regional effect, i.e $\rho(z) = \mu(z) - \mu(z_1, z_2)$, then, the regional variance $\sigma^2(z_1, z_2)$ equals to:*

$$\sigma^2_{obs}(z_1, z_2) = \sigma^2(z) + \frac{\int_{z_1}^{z_2} \rho^2(z) \partial z}{z_2 - z_1} \tag{12}$$

Theorem 1 reveals an important connection between the ground-truth and the observable uncertainty inside a region. Setting as ground-truth the regional uncertainty $\mathcal{H}_{true}(z_1, z_2) = \sigma(z_1, z_2)$ and as observable uncertainty the one we estimate with Eq. (11), i.e. $\mathcal{H}_{obs}(z_1, z_2) := \sigma_{obs}(z_1, z_2)$, then there is an error term quantified by the residual inside the interval, i.e. $\mathcal{E}(z_1, z_2) = \sqrt{\frac{\int_{z_1}^{z_2} \rho^2(z) \partial z}{z_2 - z_1}}$, such that:

$$\mathcal{H}^2_{obs}(z_1, z_2) = \mathcal{H}^2_{true}(z_1, z_2) + \mathcal{E}^2(z_1, z_2) \tag{13}$$

Eq. (13) shows that the error term quantifies how much the observable uncertainty of the region $[z_1, z_2]$ deviates for the one we would like to measure and provides a quantified metric to evaluate a particular splitting. As we prove in the Appendix, UALE uncertainty approximation is unbiased if and only if the partitioning is such that ther error term is zero for all bins, i.e. $\sum_{k=1}^{K} \mathcal{E}^2(z_{k-1}, z_k) = 0$

3.4 Bin-Splitting: Finding Regions With Homogeneous Effects

In this section we formulate bin-splitting as an unsupervised clustering problem. The main idea of the clustering problem is to find the best bin-splitting solution, handling appropriately two conflicting objectives. On the one hand, we want to minimize the error term $\mathcal{E}$ defined above for minimizing the error in the uncertainty estimation and on the other hand, we want wide-enough bins for including a fair population of samples for a robust estimation of $\hat{\mu}(z_1, z_2), \hat{\sigma}(z_1, z_2)$. We set-up the following optimization problem:

$$\begin{aligned}
\min_{\mathcal{Z}=\{z_0,\dots,z_K\}} \quad & \mathcal{L} = \sum_{k=1}^{K} \tau_k \hat{\sigma}^2(z_{k-1}, z_k) \Delta z_k \\
\text{where} \quad & \Delta z_k = z_k - z_{k-1} \\
& \tau_k = 1 - \alpha \frac{|S_k|}{N} \\
\text{s.t.} \quad & |\mathcal{S}_k| \geq N_{\texttt{NPB}} \\
& z_0 = x_{s,min} \\
& z_K = x_{s,max}
\end{aligned} \tag{14}$$

We search for the sequence of intervals $z_0, \dots, z_K$ that minimizes the sum of bin costs. The cost of each bin is the approximated variance $\hat{\sigma}_k^2$ scaled by the bin length Δz_k and discounted by the term τ_k. The term τ_K acts as a tie-braker that favors the selection of a bigger bin in case it has similar variance with the aggregate variance of many smaller ones. The constraint of at least $N_{\texttt{PPB}}$ points per bin sets the lowest-limit for a *robust* estimation. The user can choose to what extent they favor the creation of wide bins through the discount parameter by setting the parameter α and where they set the minimum for robust approximation with the parameter $N_{\texttt{PPB}}$. For providing a rough idea of some sensible value, in our experiments we set $\alpha = 0.2$ which means that the discount ranges in $[0\%, 20\%]$ and $N_{\texttt{PPB}} = \frac{N}{20}$. It is also important to clarify that by minimizing $\hat{\sigma}_k^2 \approx \mathcal{H}^2_{obs}$, we actually minimize the squared error term $\mathcal{E}^2$, since the term $\mathcal{H}^2_{true}$ is independent of the bin splitting procedure.

3.4.1 Solve Bin-Splitting with Dynamic Programming

For achieving a computationally-grounded solution we set a threshold K_{max} on the maximum number of bins which also discretizes the solution space. The width of the bin can take discrete values that are multiple of the minimum step $u = \frac{x_{s,max} - x_{s,min}}{K_{max}}$. For defining the solution, we use two indexes. The index $i \in \{0, \dots, K_{max}\}$ denotes the point (z_i) and the index $j \in \{0, \dots, K_{max}\}$ denotes the position of the j-th multiple of the minimum step, i.e., $x_j = x_{s,min} + j \cdot u$. The recursive cost function $T(i,j)$ is the cost of setting $z_i = x_j$:

$$\mathcal{T}(i,j) = \min_{l \in \{0,\dots,K_{max}\}} \left[\mathcal{T}(i-1,l) + \mathcal{B}(x_l, x_j)\right] \tag{15}$$

where $\mathcal{T}(0,j)$ equals zero if $j = 0$ and ∞ in any other case. $\mathcal{B}(x_l, x_j)$ denotes the cost of creating a bin with limits $[x_l, x_j)]$:

$$\mathcal{B}(x_l, x_j) = \begin{cases} \infty, & \text{if } x_j > x_l \text{ or } |\mathcal{S}_{(x_j,x_l)}| < N \\ 0, & \text{if } x_j = x_l \\ \hat{\sigma}^2(x_j, x_l), & \text{if } x_j \leq x_l \end{cases} \tag{16}$$

The optimal solution is given by solving $\mathcal{L} = \mathcal{T}(K_{max}, K_{max})$ and keeping track of the sequence of steps.

- Discuss more aspects (e.g. Computational complexity)

4 SIMULATION EXAMPLES

The simulation examples, where the data-generating distribution $p(\mathbf{X})$ and the predictive function $f(\cdot)$ are defined by us, enable the evaluation of competitive approaches against a solid ground-truth. We follow this common XAI practice (Aas et al., 2021; Herbinger et al., 2022) for providing secure empirical evidence about the superiority of ARE method.

We split the simulation examples in two groups. The first group, Section 4.1, aims at showcasing that ARE method is more accurate than PDP-ICE in quantifying the average effect and the level of heterogeneous effects (uncertainty), when input features are correlated. The second group, Section 4.2, illustrates that ARE method achieves a better approximation (average effect and uncertainty) in cases of limited samples, due to automatic bin-splitting. In both groups, we choose to illustrate the most indicative examples; a more extensive evaluation is provided in the Appendix.

4.1 Case 1: Uncertainty Quantification

In this simulation, we will compare ARE method against PDP-ICE in quantifying the main effect and the uncertainty. Since there is ambiguity about the ground-truth first-order effect in cases of correlated features, e.g. (Apley and Zhu, 2020; Grömping, 2020), we limit ourselves to the following piecewise linear function,

$$f(\mathbf{x}) = \begin{cases} f_{\texttt{lin}} + \alpha f_{\texttt{int}} & \text{if } f_{\texttt{lin}} < 0.5 \\ 0.5 - f_{\texttt{lin}} + \alpha f_{\texttt{int}} & \text{if } 0.5 \leq f_{\texttt{lin}} < 1 \\ \alpha f_{\texttt{int}} & \text{otherwise} \end{cases} \tag{17}$$

where $f_{\texttt{lin}}(\mathbf{x}) = a_1 x_1 + a_2 x_2$ and $f_{int}(\mathbf{x}) = x_1 x_3$. As we observe, the linear term $f_{\texttt{lin}}$ includes the two correlated features and the term $f_{\texttt{int}}$ interacts the two non-correlated variables. The samples that we use in all examples are coming from the following distribution: $p(\mathbf{x}) = p(x_3)p(x_2|x_1)p(x_1)$ where $x_1 \sim \mathcal{U}(0,1)$, $x_2 = x_1$ and $x_3 \sim \mathcal{N}(0, \sigma_3^2)$. We will test the effect computed by ARE and PDP-ICE in three cases; (a) no interaction ($\alpha = 0$) and equal weights ($a_1 = a_2$), (b) no interaction ($\alpha = 0$) and different weigths ($a_1 \neq a_2$) and (c) with interaction ($\alpha \neq 0$) with equal weights ($a_1 = a_2$).

In all cases, we firstly compute the ground-truth average effect and the uncertainty analytically (proofs in the Appendix) and then we compare it against the approximation provided by each method. The approximation is computed after generating $N = 300$ samples. As we will see, despite the model's simplicity, PDP-ICE fail in quantifying the effect correctly, due to correlation between features X_1 and X_2. Finally, besides it is not the focus of the current examples, we will also observe that ARE manages to correctly extract the regions with constant effect. Details for obtaining the ground truth effects and the experimental set-up are in the Appendix.

No Interaction, Equal weights. We test the feature effect when no interaction term is apparent, i.e. $\alpha = 0$, and the weights are equal $a_1 = a_2 = 1$. In this case, the ground truth effect $f^{\texttt{GT}}_{\mu}(x_1)$ is: x_1 when $0 \leq x_1 < 0.25$, $-x_1$ when $0.25 \leq x_1 < 0.5$ and zero otherwise. In each position, there is total absence of heterogeneous effects; therefore, the uncertainty of the global explanation is zero $f^{\mathcal{GT}}_{\sigma^2}(x_1) = 0$. In Figure 1, we observe that PDP main effect is wrong and ICE plots imply the existence of heterogeneous effects. In contrast, ALE captures correctly the average effect and the zero uncertainty and it also groups perfectly the regions with constant-effect.

No Interaction, Different weights. As before, there is no interaction term and, therefore, the ground-truth is zero, i.e. $f^{\mathcal{GT}}_{\sigma^2}(x_1) = 0$. The weigts are different $a_1 = 2, a_2 = 0.5$, therefore, the ground-truth effect is $f^{\texttt{GT}}_{\mu}(x_1)$ is: $2x_1$

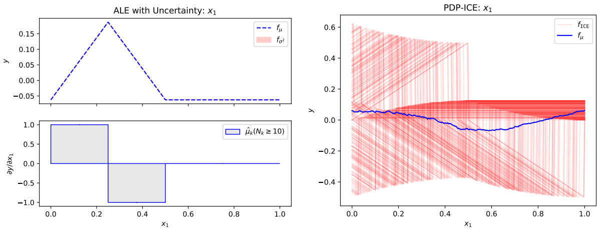

Figure 1: No interaction, Equal weights: Feature effect for x_1; Ground-truth vs (a) ARE method at the left and (b) PDP-ICE at the right.

when $0 \leq x_1 < 0.2$, $-2x_1$ when $0.2 \leq x_1 < 0.4$ and zero otherwise. In Figure 2, we observe that PDP estimation is completely opposite to the ground-truth effect, i.e. negative in $[0, 0.2)$ and positive in $[0.2, 0.4)$, and the ICE erroneously implies heterogeneous effects. As before, ALE quantifies perfectly the ground truth effect and the zero-uncertainty, extracting correcltly the constant effect regions.

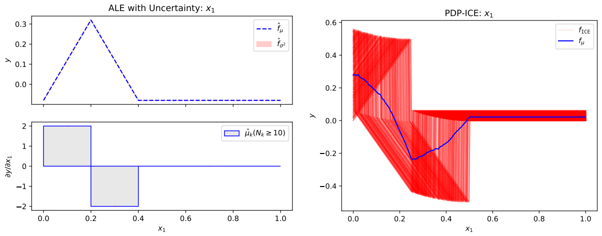

Figure 2: No interaction, Different weights: Feature effect for x_1; Ground-truth vs (a) ARE method at the left and (b) PDP-ICE at the right.

Uncertainty, Equal weights. In this case we activate the interaction term, i.e. $a = 1$, and we set equal weights $a_1 = a_2 = 1$. The interaction term provokes heterogeneous effects in features x_1, x_3, i.e., at any position x_1, the local effects dependend on the unknown value of X_3 and vice-versa. The effect of x_2 continues to have zero-uncertainty, since it does not appear in any interaction term. As we observe in Figure 3, ARE captures perfectly the effect of all features, and the uncertainty in all cases. As expected, PDP-ICE quantify the effect and the uncertainty correctly only in the case of x_3, since X_3 is independent from other features. For the correlated features, i.e. x_1, x_2, the average effect computed by PDP and the uncertainty implied by ICE plots are wrong.

4.2 Case 2: Bin-Splitting

In this simulation, we aim to to quantify the advantages of automatic bin-splitting based on the objective of Eq.(14), through the following validation framework. We generate a big dataset with dense sampling ($N = 10^6$) and we treat the DALE estimation with dense fixed-size bins ($K = 10^3$) as ground-truth. Afterwards, we generate less samples ($N = 500$) and we compare (a) the fixed-size DALE estimation for many different K versus (b) the auto-bin splitting algorithm. In all cases, we use a bivariate black-box function $f(\cdot)$ where the samples are instances of the distribution $p(\mathbf{x}) = p(x_2|x_1)p(x_1)$ where $x_1 \sim \mathcal{U}(0,1)$ and $x_2 \sim \mathcal{N}(x_1, \sigma_2^2 = 0.5)$.

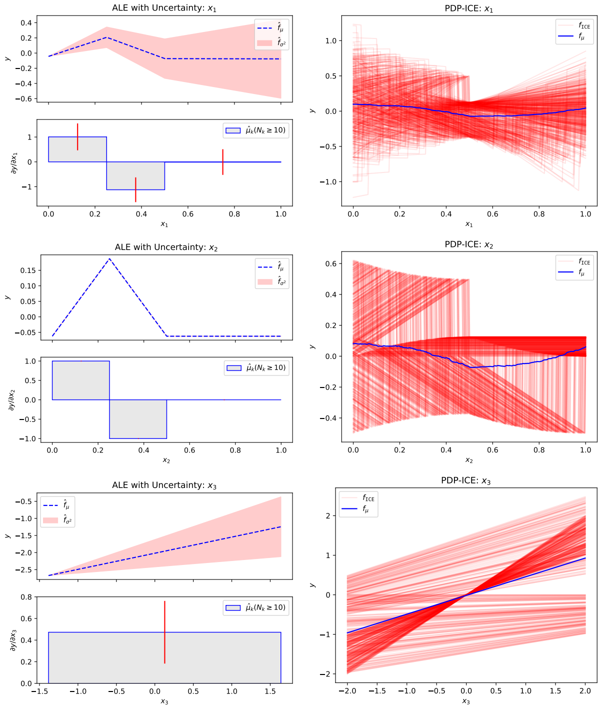

Figure 3: With interaction, equal weights: Feature effect for all features, x_1 to x_3 from top to bottom; Ground-truth vs (a) ARE method at the left columns and (b) PDP-ICE at the right column.

We denote as $\mathcal{Z}^{\text{DP}} = \{z^{\text{DP}}_{k-1}, \cdots, z^{\text{DP}}_{K}\}$ the sequence obtained by automatic bin-splitting based on the optimisation problem of Eq. (14) and with $\mathcal{Z}^{\text{K}}$ the fixed-size splitting with K bins. The evaluation is done with regard to two metrics; $\mathcal{L}^{\mu}_{\text{DP|K}} = \frac{1}{|\mathcal{Z}^{\text{DP|K}}|} \sum_{k \in \mathcal{Z}^{\text{DP|K}}} |\mu_k - \hat{\mu}_k|$ quantifies the average error in estimating the the expected effect and $\mathcal{L}^{\sigma}_{\text{DP|K}} = \frac{1}{|\mathcal{Z}^{\text{DP|K}}|} \sum_{k \in \mathcal{Z}^{\text{DP|K}}} |\sigma_k - \hat{\sigma}_k|$ the average error in estimating the variance of the effect. The metric $\mathcal{L}^{\rho}_{\text{DP|K}} = \frac{1}{|\mathcal{Z}^{\text{DP|K}}|} \sum_{k \in \mathcal{Z}^{\text{DP|K}}} \rho_k$ quantifies the average nuisance uncertainty. The optimal bin-splitting is the one that minimises, at the same time, both $\mathcal{L}^{\mu}$ and $\mathcal{L}^{\sigma^2}$ errors. For consistent results, in all the examples below, we regenerate samples and repeat the computations for $t = 10$ times, providing the mean value for all metrics.

Piecewise-Linear Function. In this example, we define $f(\mathbf{x}) = a_1 x_1 + x_1 x_2$ with 5 piecewise-linear regions of different-size, i.e., a_1 equals to $= \{2, -2, 5, -10, 0.5\}$ in the intervals defined by the sequence $\{0, 0.2, 0.4, 0.45, 0.5, 1\}$. As we observe, in Figure 6, ARE method extracts a sequence of intervals with better $\mathcal{L}^{\mu}_{\mathtt{DP}}$ and $\mathcal{L}^{\sigma^2}_{\mathtt{DP}}$ error compared to any fixed-size splitting. Analyzing the fixed-size errors helps us understand the importance of variable-size splitting. In Figure 6(b), we observe a positive trend between $\mathcal{L}^{\mu}_{\mathtt{K}}$ and K, concluding that bin effect estimation is more incosistent as Δx becomes smaller, due to less points contributing to each bin. The interpretation of variance error is slightly more complex. Given that the smallest interval is $\Delta x = 0.05 \Rightarrow K = 20$ and all intervals are multiples of the smallest interval, any K that is not a multiple of 20 adds nuisance uncertainty $\mathcal{L}^{\rho^2}_{\mathtt{K}}$ leading to a high variance error $\mathcal{L}^{\sigma^2}_{\mathtt{K}}$. In these cases, the variance error reduces as K grows bigger because the length of the bins that lie in the limit between two piecewise linear regions becomes smaller. For $K = \{20, 40, 60, 80, 100\}$ where $\mathcal{L}^{\rho^2}_{\mathtt{K}} \approx 0$, we conclude the same as with with the mean effect error, i.e. the estimation becomes more incosistent as K grows larger.

Variable-size extracts correctly the fine-grain bins, e.g., intervals $[0.4, 0.45]$, $[0.45, 0.5]$, and acts as a merging mechanism to create wider bins when the effect remains constant, e.g. interval $[0.5, 1]$, leading to an optimal solution.

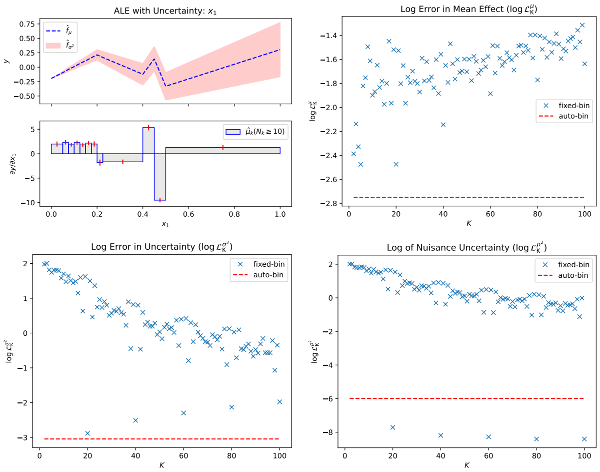

Figure 4: Figure 1

Non-Linear Function. In this example, we define a black-box function $f(\mathbf{x}) = 4x_1^2 + x_2^2 + x_1x_2$, where the effect is non-linear in all the range of x. This case has two specialties. First, there is no obvious advantage of variable-size versus fixed-size splitting, and, second, there is not an apriori optimal bin-size. Widening a bin will increase the resolution error $\mathcal{L}^{\rho^2}$ and narrowing will make less robust. In Figure 5, we observe that automatic bin splitting finds a solution that compromises the conflicting objectives, i.e., it keeps as low as possible both the main effect $\mathcal{L}^{\mu}_{\mathtt{K}}$ and the variance error $\mathcal{L}^{\sigma^2}_{\mathtt{K}}$.

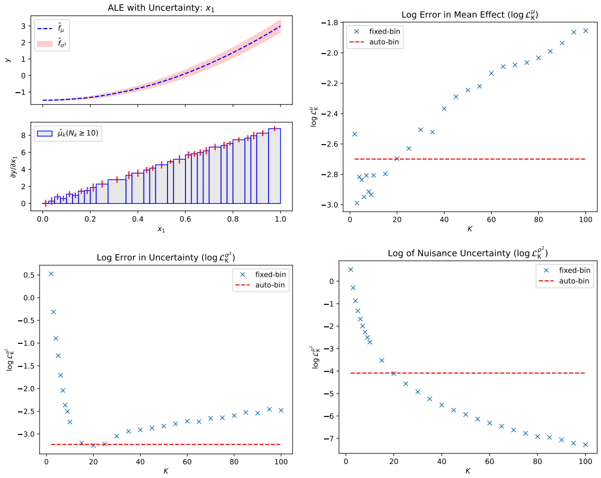

Figure 5: Figure 1

5 REAL-WORLD EXAMPLE

Here, we aim at demonstrating the usefuleness of uncertainty quantification and the advantages of automatic bin-splitting, on the real-world California Housing dataset (Pace and Barry, 1997).

ML setup The California Housing is a largely-studied dataset with approximately 20000 training instances, making it appropriate for robust Monte-Carlo approximations. The dataset contains $D = 8$ features describing building blocks of California, like latitude, longitude, total number of people residing in the block etc., and the target is the median house value inside the block. We exclude instances with missing or outlier values, i.e. $x^i_d > sig\hat{m}a_j$ in any feature and normalize them to have $\hat{\mu}_d = 0, \hat{\sigma_d}$. We split the dataset into $N_{tr} = 15639$ training and $N_test = 3910$ test examples (80/20 split) and we fit a Neural Network with 3 hidden layers of 256, 128 and 36 units respectively. After 15 epochs using the Adam optimizer with learning rate $\eta = 0.02$, the model achieves a normalized mean squared error (R-Squared) of 0.25 (0.75). In depth information about preprocessing, training and evaluation are provided in the Appendix.

Uncertainty Quantification In real-world datasets, it is infeasible to obtain the ground truth FE for seamlessly evaluating the competitive methods. Therefore, the main purpose of this part is to demonstrate the usefulness of uncertainty quantification in ARE, i.e., illustrate that modeling the heterogeneous effects is important for the interpretation of an ALE plot. For this reason, Secondly, we discuss about the FE plots obtaing with ARE and PDP-ICE.

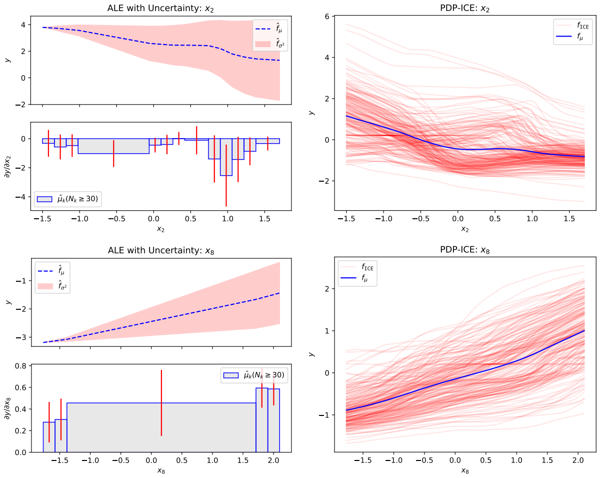

Figure 6: Figure 1

Bin Splitting For evaluating ARE in terms of bin-splitting we set-up the following evaluation framework. We treat as ground-truth the effects computed on the full training-set $N = 20000$ with dense fixed-size bin-splitting ($K = 80$). Given the big number of samples, we make the hypothesis that the approximation with dense binning is close to the ground truth. Afterwards we select less samples $N = 1000$ and we compare ARE approximation with all possible fixed-size alternatives, in terms of accuracy, i.e. we provide the quantities $\mathcal{L}^{\mu}, \mathcal{L}^{\sigma^2}, \mathcal{L}^{\rho^2}$.

6 CONCLUSION

Acknowledgments

All acknowledgments go at the end of the paper, including thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support. To preserve the anonymity, please include acknowledgments *only* in the camera-ready papers.

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020.

Hubert Baniecki, Wojciech Kretowicz, and Przemyslaw Biecek. Fooling partial dependence via data poisoning. *arXiv preprint arXiv:2105.12837*, 2021.

Matthew Britton. Vine: visualizing statistical interactions in black box models. *arXiv preprint arXiv:1904.00561*, 2019.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The annals of applied statistics*, pages 916–954, 2008.

Vasilis Gkolemis, Theodore Dalamagas, and Christos Diou. Dale: Differential accumulated local effects for efficient and accurate global explanations.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

Ulrike Grömping. Model-agnostic effects plots for interpreting machine learning models, 03 2020.

Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. Repid: Regional effect plots with implicit interaction detection. In *International Conference on Artificial Intelligence and Statistics*, pages 10209–10233. PMLR, 2022.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL `https://christophm.github.io/interpretable-ml-book`.

Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer, 2020a.

Christoph Molnar, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio. Model-agnostic feature importance and effects with dependent features–a conditional subgroup approach. *arXiv preprint arXiv:2006.04628*, 2020b.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.