Uncertainty-aware Accumulated Local Effects (UALE) for quantifying the heterogeneity of instance-level feature effects

Anonymous Author
Anonymous Institution

Abstract

Accumulated Local Effects (ALE) is a popular explainable AI method that quantifies how a feature influences the decisions of a model, handling well feature correlations. In case of complex interactions between features, instance-level feature effects may deviate from the ALE curve. It is therefore crucial to quantify this deviation, namely, the uncertainty of the effect. In this work, we define Uncertainty-aware ALE (UALE) to quantify and visualize, on a single plot, both the average effect and its uncertainty. We show that UALE quantifies uncertainty effectively, even in case of correlated features. We also note that as in ALE, UALE's approximation requires partitioning the feature domain into non-overlapping intervals (bin-splitting). The average effect and the uncertainty are computed from the instances that lie in each bin. We formally prove that to achieve an unbiased approximation of the uncertainty in each bin, bin-splitting must follow specific constraints. Based on this, we propose a method to determine the optimal intervals, [balancing the estimation bias and variance.] We demonstrate, through synthetic and real datasets, (a) the advantages of modeling the uncertainty with UALE compared to alternative methods and (b) the effectiveness of UALE's appropriate bin splitting for a [good] approximating of the average effect and its uncertainty.

1 INTRODUCTION

Recently, Machine Learning (ML) has been adopted in mission critical domains, such as healthcare and finance. In these areas, we need a combination of accurate predictions along with meaningful model explanations. For this reason, there is an increased interest in Explainable AI (XAI) to understand the decision mechanism of complex black-box models. XAI literature distinguishes between local and global techniques (Molnar et al., 2020a). Local methods provide instance-level explanations, i.e., explain the prediction for a specific instance, whereas global methods summarize the entire model behavior. Global methods create a global explanation by aggregating the instance-level explanations into a single interpretable outcome, usually a number or a plot.

A popular class of global XAI are feature effect (FE) methods (Grömping, 2020) that quantify the average (over all instances) effect of a single feature on the output[1]. There are three widely-used FE methods: *Partial Dependence Plots* (PDP)(Friedman, 2001), *Marginal Plots* (MP)(Apley and Zhu, 2020) and *Accumulated Local Effects* (ALE)(Apley and Zhu, 2020). ALE is established as the state-of-the-art in quantifying the average effect, since PDP and MP have been criticized (Grömping, 2020) of being innacurate when the input features are correlated.

When complex interactions between features exist, the instance-level (local) feature effects may significantly deviate from the aggregated outcome, a phenomenon called *aggregation bias* (Mehrabi et al., 2021). Aggregation bias leads to ambiguities; for example, a feature with zero average effect may indicate (a) no effect on the output or (b) an effect that is highly positive for some instances and highly negative for some others. As a result, FE methods should visualize the heterogeneity of instance-level explanations, in addition to the average effect. In the case of PDP, this heterogeneity is visualized via the Individual Conditional Explanations (ICE) (Goldstein et al., 2015). ICE plots, however, have the same limitations as PDP in case of correlated features. To the best of our knowledge, no such heterogeneity visualization method has been proposed for ALE.

In this work, we propose UALE, a global feature-effect

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

[1] FE methods also isolate the combined effect of a pair of features to the output. Combinations of more than two features are uncommon, since they are difficult to estimate and visualize.

method that extends ALE for accurately modelling the *uncertainty of the feature effect*, in case of correlated features. The uncertainty of the feature effect expresses the heterogeneity of the instance-level (local) effects, which in our case, is quantified by their standard deviation. As with ALE, UALE's approximation requires partitioning the feature domain into non-overlapping intervals (bin-splitting). As we formally prove, a fixed equi-width partitioning that does not consider the characteristics of the instance-level effects may lead to an erroneous (biased) estimation of the uncertainty. Therefore, we propose an uncertainty-driven variable-width partitioning that lead to an unbiased approximation.

The contributions of this paper are:

- A global feature effect method (UALE) that quantifies both the average effect and the uncertainty, i.e., the heterogeneity of instance-level effects.
- An unbasied approximation of the uncertainty through a dynamic uncertainty-driven bin partitioning of the feature domain.
- The evaluation of UALE on both synthetic and real datasets.

The implementation of our method and the code for reproducing all the experiments is provided along with the manuscript and will become publicly available upon acceptance.

2 BACKGROUND AND RELATED WORK

Here we review existing methods for modeling the average effect and the heterogeneity. We also provide the necessary background for ALE definition and approximation.

Notation: Let $\mathcal{X} \in \mathbb{R}^d$ be the d-dimensional feature space, $\mathcal{Y} \in \mathbb{R}$ the target space and $f(\cdot) : \mathcal{X} \to \mathcal{Y}$ the black-box function. We use index $s \in \{1, \dots, d\}$ for the feature of interest and $c = \{1, \dots, d\} - s$ for the set with all other indexes. For convenience, we denote the feature vector $\mathbf{x} = (x_1, \cdots, x_s, \cdots, x_D)$ with $(x_s, \mathbf{x}_c)$ and the corresponding random variables $X = (X_1, \cdots, X_s, \cdots, X_D)$ with (X_s, X_c). The training set is $\mathcal{D} = \{(\mathbf{x}^i, y^i)\}_{i=1}^N$ sampled i.i.d. from the distribution $\mathbb{P}_{X,Y}$. Finally, we use $f^{<\texttt{method}>}(x_s)$ for denoting the s-th FE, where $<\texttt{method}>$ indicates the particular method in use, for example ALE. An extensive list with all symbols used in this paper is provided at Appendix A.

2.1 Feature Effect Methods and ALE

The three well-known feature effect methods are: *Partial Dependence Plots* (PDP), *Marginal Plots* (MP) and *Accumulated Local Effects* (ALE). PDP formulates the global FE as an expectation over the distribution of X_c, i.e., $f^{\texttt{PDP}}(x_s) = \mathbb{E}_{X_c}[f(x_s, X_c)]$, whereas MP as an expectation over the distribution of $X_c|X_s$, i.e., $f^{\texttt{MP}}(x_s) = \mathbb{E}_{X_c|x_s}[f(x_s, X_c)]$. Both methods suffer from misestimations in case of correlated features; PDP integrates over unrealistic instances and MP computes aggregated effects, i.e., imputes the combined effect of sets of features to a single feature. [cite]

ALE overcomes these problems. Specifically, ALE defines the local effect of the s-th feature at a specific point $(z, \mathbf{x}_c)$ of the input space $\mathcal{X}$ with the partial derivative $f^s(z, \mathbf{x}_c) = \frac{\partial f}{\partial x_s}(z, \mathbf{x}_c)$. All the local explanations at z are then weighted by the conditional distribution $p(\mathbf{x}_c|z)$ and are averaged, to produce the averaged effect $\mu(z)$. ALE plot is the accumulation of the averaged local effects:

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

where $x_{s,min}$ is the minimum value of the s-th feature.

2.2 Heterogeneity Of Instance-Level Effects

The global effect is computed as an expectation over local (instance-level) effects. In addition to this global effect, it is important to know to what extent the local effects deviate from the global explanation, i.e. to quantify the uncertainty of the global effect. ICE plots and similar methods (e.g., d-ICE plots (Goldstein et al., 2015)) provide a set of curves illustrated on top-of PDP. Each curve corresponds to one instance of the dataset, $f_i^{\texttt{ICE}}(x_s) = f(x_s, \mathbf{x}_c^i)$. The user then visually observes if the curves are homogeneous, i.e., all instances have similar effect plots, and to what extent they deviate from the PDP. There are 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). Unfortunatelly, these methods are subject to the failure modes of PDPs in cases of correlated features (Baniecki et al., 2021), as we also confirm in our experimental analysis (c.f., Section 4.1).

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 interactions between the input features. [A strong interaction is an indication of the existence of heterogeneous effects. These methods, however, do not quantify the level of heterogeneity.]

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

ALE is estimated from the available dataset instances. (Apley and Zhu, 2020) proposed to divide the feature domain in K bins of equal size and to estimate the local effects in each bin by evaluating the model f at the bin limits:

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

where 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 instances of the k-th bin, i.e. $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x_s^i < z_k\}$. (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 significant computational advantages and allows the recomputation of the accumulated effect with different bin-splitting with near-zero additional computational cost. Moreover, by avoiding the use of artificial samples at the bin limits DALE allows the use of wider bins without out-of-distribution sampling. In both cases, the approximations partition the feature domain in K equally-sized bins, without considering the underlying local effects.

3 Uncertainty-Aware ALE (UALE)

UALE extends ALE by quantifying the level of heterogeneous effects (uncertainty) and by optimally partitioning the feature domain into bins. In the following, we first define UALE, then we reformulate it using an interval-based formulation, provide UALE approximation and, finally, define and solve the problem of optimal bin-splitting.

3.1 UALE Definition

UALE extends ALE with a component for quantifying the uncertainty. Let us denote as $\mathcal{H}(z)$ the uncertainty of the local effects at a specific point $x_s = z$ and quantify it using the standard deviation of the local explanations $\mathcal{H}(z) = \sigma(z)$, where:

$$\sigma^2(z) = \mathbb{E}_{X_c|X_s=z} \left[\left(f^s(z, X_c) - \mu(z) \right)^2 \right] \tag{4}$$

The uncertainty emerges from the natural characteristics of the experiment. These include the feature correlations present in the data generating distribution and the implicit feature 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:

Figure 1: UALE-concept-figure

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

UALE defines the effect at a specific point x_s with a pair that consists of the average effect and the uncertainty $(\mu(z), \sigma(z))$ and visualizes it with a continuous curve $f_{\mu}^{\texttt{ALE}}(x_s)$ as defined in Eq. (1) and a confidence interval $f_{\mu}^{\texttt{ALE}}(x_s) \pm f_{\sigma}^{\texttt{ALE}}(x_s)$. For notational convenience, we use $f^{\texttt{UALE}}(x_s) := f_{\mu}^{\texttt{ALE}}(x_s) \pm f_{\sigma}^{\texttt{ALE}}(x_s)$ to represent this plot.

3.2 UALE Interval-Based Formulation

In this Section, we adapt the definition the average feature effect Eq. (1) and the uncertainty Eq. (5) to intervals. We define the bin-effect $\mu(z_1, z_2)$ and the bin-uncertainty $\mathcal{H}(z_1, z_2) = \sigma(z_1, z_2)$ as:

$$\mu(z_1, z_2) = \mathbb{E}_{z \sim \mathcal{U}(z_1, z_2)}[\mu(z)] = \frac{\int_{z_1}^{z_2} \mu(z) \partial z}{z_2 - z_1} \tag{6}$$

$$\sigma^2(z_1, z_2) = \mathbb{E}_{z \sim \mathcal{U}(z_1, z_2)}[\sigma^2(z)] = \frac{\int_{z_1}^{z_2} \sigma^2(z) \partial z}{z_2 - z_1} \tag{7}$$

Intuitively, the regional-effect and the regional-uncertainty quantify the expected average effect and the expected uncertainty if we randomly draw a point z^* given a uniform distribution $z^* \sim \mathcal{U}(z_1, z_2)$. If we also define as $\mathcal{Z}$ the sequence of $K+1$ points that partition the domain of the s-th feature into K variable-size intervals, i.e. $\mathcal{Z} = \{z_0, \ldots, z_K\}$, we can redefine UALE using an interval-based formulation $\tilde{f}_{\mathcal{Z}}^{\texttt{UALE}}(x_s) := \tilde{f}_{\mathcal{Z},\mu}^{\texttt{ALE}}(x_s) \pm \tilde{f}_{\mathcal{Z},\sigma}^{\texttt{ALE}}(x_s)$, where:

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

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

3.3 UALE Interval-Based Approximation

For approximating the mean effect (Eq. (6)) and the uncertainty (Eq. (7)) we use the set $\mathcal{S}_k$ of dataset instances with the s-th feature lying inside the k-th bin, i.e.,$\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x_s^i < z_k\}$. The mean effect is approximated with

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

which, as shown by(Gkolemis et al.), is an unbiased estimator of Eq. (6) under the assumption that the points are uniformly distributed inside the intervals. The corresponding approximation for the uncertainty (Eq. (7)) is:

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

which, in the general case, is a biased estimator of Eq. (7). In Theorem 3.1, we prove that Eq. (11) is an approximator of $\sigma^2_*(z_1, z_2) = \frac{\int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x}_c | x_s = z} \left[(f^s(z, \mathbf{x}_c) - \mu(z_1, z_2))^2 \right] \partial z}{z_2 - z_1}$ which is an overestimation of $\sigma^2(z_1, z_2)$.

Theorem 3.1. *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 it holds*

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

Proof. The proof is ... □

Theorem 3.1 reveals an important connection between the ground-truth and the uncertainty inside a region. For convenience, we denote as $\mathcal{E}^2(z_1, z_2) = \frac{\int_{z_1}^{z_2} \rho^2(z) \partial z}{z_2 - z_1}$ the error term that models the expected square residual inside the interval and as $\mathcal{H}_{obs}(z_1, z_2) := \sigma_{obs}(z_1, z_2)$ the observable uncertainty. It holds that the observable uncertainty is an overestimation of the correct uncertainty

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

and, therefore, the estimation is unbiased only in case $\mathcal{E}^2(z_1, z_2) = 0$.

3.4 Bin-Splitting: Finding Regions With Homogeneous Effects

In this section, we propose optimally select the bin-splitting $\mathcal{Z}$ with the aim to minimize the error term $\mathcal{E}$. At the same time we want to maintain a sufficient population of samples within each bin to allow 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 the bin costs. The cost of each bin is the approximated variance $\hat{\sigma}^2_k$ scaled by the bin length Δz_k and discounted by the term τ_k. The term τ_K 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, in our experiments we set $\alpha = 0.2$ which means that the discount ranges between $[0\%, 20\%]$ and $N_{\texttt{PPB}} = \frac{N}{20}$. It is also important to clarify that by minimizing $\hat{\sigma}^2_k \approx \mathcal{H}^2_{obs}$, we actually minimize the squared error term $\mathcal{E}^2$, since the term $\mathcal{H}^2_{true}$ is independent of the bin splitting.

3.4.1 Solve Bin-Splitting with Dynamic Programming

[Review all the description] 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

In XAI it is common (Aas et al., 2021; Herbinger et al., 2022) to use simulation examples for compering methods, because the data-generating distribution $p(\mathbf{X})$ and the predictive function $f(\cdot)$ are known, and, therefore, the ground-truth is accesible. We follow this common XAI practice and we split the evaluation in two groups of examples. The first group (Section 4.1) compares UALE definition against PDP-ICE in modeling the average effect and the uncertainty. The second group (Section 4.2) compares UALE approximation using the automatic bin-splitting against the fixed-size alternative, in estimating the average effect and the uncertainty.

4.1 Case 1: UALE vs PDP-ICE

In this simulation, we compare UALE against PDP-ICE. The aim of this example is to highlight that when features are correlated, PDP and ICE fail in quantifying the average effect and the heterogeneity, even when the black-box function is relatatively simple.

We use 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 $x_3 \sim \mathcal{N}(0, \sigma_3^2)$ and the following piecewise linear function:

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

where $f_{\texttt{lin}}(\mathbf{x}) = a_1 x_1 + a_2 x_2$ and $f_{\texttt{int}}(\mathbf{x}) = x_1 x_3$. The linear term $f_{\texttt{lin}}$ includes the two correlated features and the interaction term $f_{\texttt{int}}$ the two non-correlated. We evaluate the effect computed by UALE and PDP-ICE in three cases; (a) without interaction ($\alpha = 0$) and equal weights ($a_1 = a_2$), (b) without 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 provide the ground-truth average effect and uncertainty computed with analytical derivations (proofs at the Appendix) and we compare it against the approximations provided by each method, after generating $N = 500$ samples.

No Interaction, Equal weights. In this set-up, $\alpha = 0$ (no interaction) and the $a_1 = a_2 = 1$ (equal weights). The ground truth effect $f_\mu^{\texttt{GT}}(x_1)$ is: x_1 in $[0, 0.25]$, $-x_1$ in $[0.25, 0.5]$ and 0 in $[0.5, 1]$. The uncertainty is $f_{\sigma^2}^{\mathcal{GT}}(x_1) = 0$ in $[0, 1]$ because x_1 does not interact with ohter variables. In Figure 2, we observe that PDP effect is wrong and ICE plots imply the existence of heterogeneous effects. In contrast, UALE models correctly both the average effect and the zero uncertainty and approximates it accurately finding a zero aggregate residual error partitioning of x_1 domain.

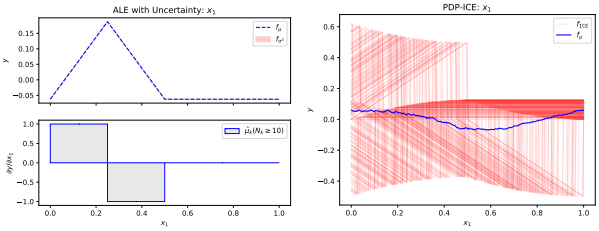

Figure 2: No interaction, Equal weights: Feature effect for x_1 using UALE (Left) and PDP-ICE (Right).

¡¡¡¡¡¡¡ HEAD

No Interaction, Non-Equal Weights. In this set-up, $\alpha = 0$ (no interaction), $a_1 = 2$ and $a_2 = 0.5$ (non-equal weights). The non-equal weights have implications at both the gradient and the interval of the piece-wise linear regions, i.e., $f_\mu^{\texttt{GT}}(x_1)$ is: $2x_1$ in $[0, 0.2]$, $-2x_1$ in $[0.2, 0.4]$ and 0 in $[0.4, 1]$. As before, the ground-truth uncertainty is $f_{\sigma^2}^{\mathcal{GT}}(x_1) = 0$ because x_1 does not interact with other features. In Figure 3, we observe that PDP estimation is completely opposite to the ground-truth effect, i.e. negative in the region $[0, 0.2)$, positive in $[0.2, 0.4)$, and the ICE erroneously implies the existence of heterogeneous effects. As before, ALE quantifies correctly the ground truth effect, the zero-uncertainty and partioning correctly the x_1 domain. =======

No Interaction, Different weights. As before, no interaction exits $\alpha = 0$, therefore, the ground-truth uncertainty is zero, i.e. $f_{\sigma^2}^{\mathcal{GT}}(x_1) = 0$. The weights are $a_1 = 2, a_2 = 0.5$, therefore, the ground-truth effect is $f_\mu^{\texttt{GT}}(x_1)$ is, $2x_1$ when $0 \leq x_1 < 0.2$, $-2x_1$ when $0.2 \leq x_1 < 0.4$ and zero otherwise. In Figure 3, we observe that PDP estimation is completely opposite to the ground-truth effect, i.e. negative in the region $[0, 0.2)$ and positive in $[0.2, 0.4)$, and the ICE erroneously implies the existence of heterogeneous effects. As before, ALE quantifies correctly the ground truth effect, the zero-uncertainty and extracts the constant effect regions. ¿¿¿¿¿¿¿ origin/overleaf-2022-10-11-1029

Uncertainty, Equal weights. In this set-up, we activate the interaction term, i.e. $a = 1$, keeping the weights equal $a_1 = a_2 = 1$. The interaction term provokes heterogeneous effects for features x_1 and x_3, because the local effects at x_1 depend on the unknown value of X_3 and vice-versa. The effect of x_2 has zero-uncertainty since it does not appear in

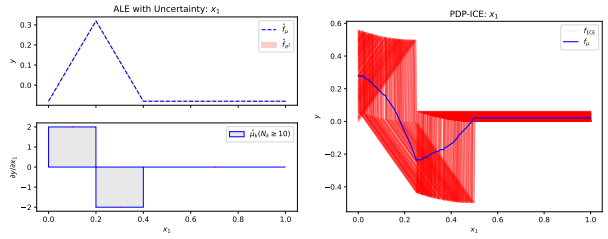

Figure 3: No interaction, Different weights: Feature effect for x_1 using UALE (Left) and PDP-ICE (Right).

any interaction term. As we observe in Figure 4, UALE correcly models the average effect and the uncertainty in all cases. On the other hand, PDP-ICE quantifies correctly the average effect and the uncertainty only in the case of feature x_3, because X_3 is independent from other features. For the correlated features (x_1, x_2) both the average effect (PDP) and the uncertainty (ICE) are wrong.

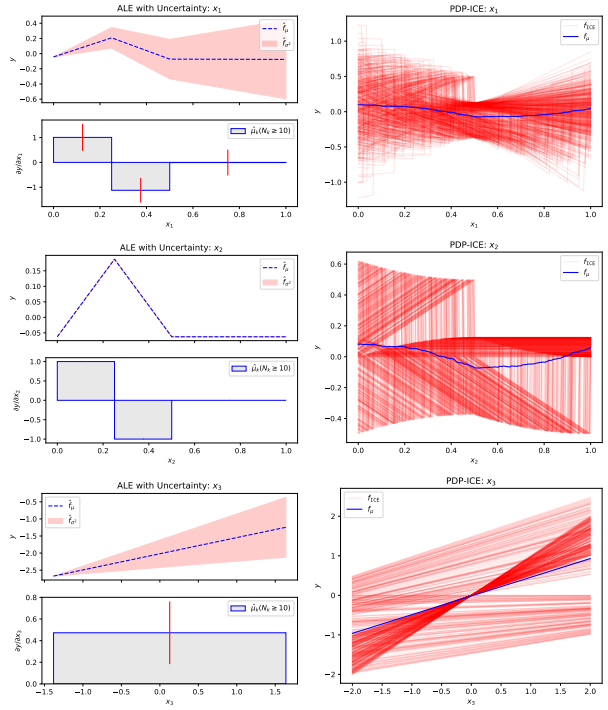

Figure 4: With interaction, equal weights: From top to bottom, feature effect for features $\{x_1, x_2, x_3\}$ using UALE (left column) and PDP-ICE (right column).

Discussion. Despite the model's simplicity, PDP-ICE fail in modeling both the average effect and the uncertainty. The error is due to the method's formulation, i.e. the use of the marginal distribution that ignores correlations between features, and not due to poor approximation because of limited samples. In contrast, UALE models both correctly and estimates them accurately extracting the regions with constant effect. The examples above do not cover the case of an interaction term between correlated features, for example a term $x_1 x_2$, because there is an open debate about the ground-truth effect in this case (Grömping, 2020).

4.2 Case 2: Bin-Splitting

In this simulation, we illustrate the advantages of automatic bin-splitting against the fixed-size alternative. For this reason, we generate a very big dataset applying dense sampling ($N = 10^6$) and we treat the estimation with dense fixed-size bins ($K = 10^3$) as the ground-truth UALE. Afterwards, we generate less samples ($N = 500$) and we compare the fixed-size estimation (for many different K) against UALE automatic bin-splitting. In all set-ups, we sample from $p(\mathbf{x}) = p(x_2|x_1)p(x_1)$ where $x_1 \sim \mathcal{U}(0, 1)$ and $x_2 \sim \mathcal{N}(x_1, \sigma_2^2 = 0.5)$. We denote as $\mathcal{Z}^{\texttt{DP}} = \{z_{k-1}^{\texttt{DP}}, \cdots, z_K^{\texttt{DP}}\}$ the sequence obtained by automatic bin-splitting and with $\mathcal{Z}^{\texttt{K}}$ the fixed-size splitting with K bins. The evaluation metrics we report are the average number of $t = 30$ independent runs, using each time $N = 500$ different samples.

Metrics The evaluation is done with regard to two metrics counting the mean error per bin, where the bin error is definde as the absolute difference of the approximation from the ground-truth. The first metric, Eq. (18), counts the mean bin error wrt average effect and the second, Eq. (19) the mean bin error wrt uncertainty:

$$\mathcal{L}^{\mu}_{\texttt{DP}|\texttt{K}} = \frac{1}{|\mathcal{Z}^{\texttt{DP}|\texttt{K}}|} \sum_{k \in \mathcal{Z}^{\texttt{DP}|\texttt{K}}} |\mu(z_k - z_{k-1}) - \hat{\mu}(z_k - z_{k-1})| \tag{18}$$

$$\mathcal{L}^{\sigma}_{\texttt{DP}|\texttt{K}} = \frac{1}{|\mathcal{Z}^{\texttt{DP}|\texttt{K}}|} \sum_{k \in \mathcal{Z}^{\texttt{DP}|\texttt{K}}} |\sigma(z_k - z_{k-1}) - \hat{\sigma}(z_k - z_{k-1})| \tag{19}$$

The ground truth μ_k (σ_k) is the mean value of the bin-effects (bin-uncertainties) of the bins inside the interval defined by the wide bin. For better interpretation of the uncertainty error, we also provide the mean (per bin) error term $\mathcal{L}^{\rho}_{\texttt{DP}|\texttt{K}} = \frac{1}{|\mathcal{Z}^{\texttt{DP}|\texttt{K}}|} \sum_{k \in \mathcal{Z}^{\texttt{DP}|\texttt{K}}} \mathcal{E}(z_{k-1}, z_k)$.

Piecewise-Linear Function. In this set-up, $f(\mathbf{x}) = a_1 x_1 + x_1 x_2$ is a piecewise-linear function with 5 different-width regions, 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 5, UALE's approximation is better than all fixed-size alternatives, in both mean effect ($\mathcal{L}^{\mu}$) and uncertainty ($\mathcal{L}^{\sigma}$). For understanding the importance of variable-size splitting, we analyze the fixed-bin error. In the top-right of Figure 5, we observe the Law of Large Numbers (LLN);

$\mathcal{L}^{\mu}_{\mathrm{K}}$ and K have positive correlation because as Δx becomes smaller less points contributing to each estimation. The interpretation of $\mathcal{L}^{\sigma}$ is more complex. Given that the smallest piecewise-linear region is $\Delta x = 0.05$ any K that is not a multiple of $\frac{1}{\Delta x} = 20$, adds nuisance uncertainty $\mathcal{L}^{\rho}_{\mathrm{K}} > 0$ leading to a biased uncertainty approximation. For these K, uncertainy error decreases for bigger K because the length of the bins that lie in the limit between two piecewise linear regions becomes smaller. For K that are multiple of 20, i.e. $K = \{20, 40, 60, 80, 100\}$, the uncertainty approximation is unbiased $\mathcal{L}^{\rho}_{\mathrm{K}} \approx 0$. Here, as in $\mathcal{L}^{\mu}$, uncertainty estimation is worse as K grows larger (LLN). As we observe in the top left of Figure 5, automatic bin-splitting separates the fine-grain bins, e.g. regions $[0.4, 0.45]$, $[0.45, 0.5]$, and unites constant-effect regions into a single bin, e.g. region $[0.5, 1]$.

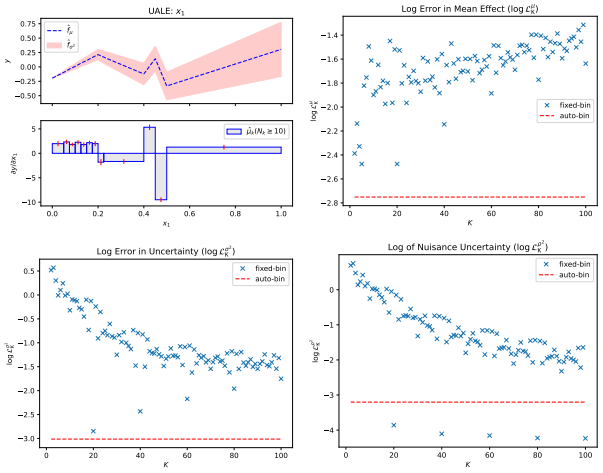

Figure 5: Figure 1

Non-Linear Function. In this set-up, $f(\mathbf{x}) = 4x_1^2 + x_2^2 + x_1x_2$, so the effect is non-linear in $0 \leq x_1 \leq 1$. In this set-up, wide bins increase $\mathcal{L}^{\rho^2}$ making the uncertainty approximation more biased and narrow bins lead to a worse approximation. In Figure 6, we observe that automatic bin splitting manages to compromise the conflicting objectives, [Explain better]

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 $D = 8$ numerical features with characteristics about the building blocks of California, e.g. lat-

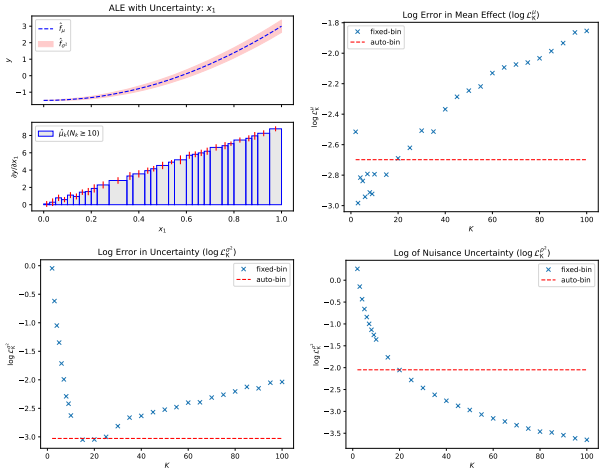

Figure 6: Figure 1

itude, longitude, population of the block or median age of houses in the block. The target variable is the median value of the houses inside the block in dollars that ranges between $[15, 500] \cdots 10^3$, with a mean value of $\mu_Y \approx 201 \cdot 10^3$ and a standard deviation of $\sigma_Y \approx 110 \cdot 10^3$.

We exclude instances with missing and outlier values. As outlier we define the feature values which ar over three standard deviations away from the mean feature value. We also normalize all features to zero-mean and unit standard deviation. 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), which corresponds to a MAE of $37 \cdot 10^3$ dollars.

Below, we illustrate the feature effect for three features: latitude x_2, population x_6 and median income x_8. The particular features cover the main FE cases, e.g. positive/negative trend and linear/non-linear curve, and they are appropriate for illustration purposes. The complete results for all features, along with in-depth information about the preprocessing, training and evaluation parts 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. We selected the particular experiment, because, in broad terms, UALE and PDP-ICE plots aggree in the estimation of the average effect and uncertainty. Therefore, we can focus on juding the quality of the information provided by the two methods.

In Figure7, we observe, from top to bottom, the effects for the latitude, population and the median income. The effect of UALE and PDP-ICE are similar for the population and

the median income. The population has a negative impact that progressively decreases: from 400 to 1500 people the house value decreases with a rate of $-150(\pm 140)$ dollars per added person, a rate that decreases from $-80(\pm 80)$ to $-60(\pm 60)$ dollars per added person as we move from 1500 to 2800 people. The level of uncertainty indicates significant variance in absolute value of the rate, but in the grant majority of instances the rate is negative. With the same inspection, we observe that the median annual income has a positive impact on the value (all numbers are thousands of dollars): 20 ± 15 per 10 of added median income for incomes in $[8, 15]$, 32 ± 20 per 10 added income in $[15, 60]$ and 40 ± 15 per 10 added income in $[60, 70]$. The uncertainty indicates that there are less heterogeneous effects about the median income compared to the number of people. In both cases, we can end-up to the same conclusion by inspecting the PDP-ICE plots. For the latitude, there is a small difference in the explanations for the region $[32, 35]$, where UALE estimates a less negative slope with less uncertainty than PDP, while the explanations are similar for the range $[35, 39]$, where both methods reveal an increase in the uncertainty around the feature value 37.5.

In general, we observe that UALE complements ALE in the quantification of the heterogeneous effects, similarly to as ICE complements PDP. The automatic extraction of constant-effect and constant-uncertainty regions provided by UALE is helpful for an easier interpretation. On the other hand, ICE plots sometimes are more descritive on locating the type of heterogeneity, whereas UALE quantifies only the level (not the type) of heterogeneity.

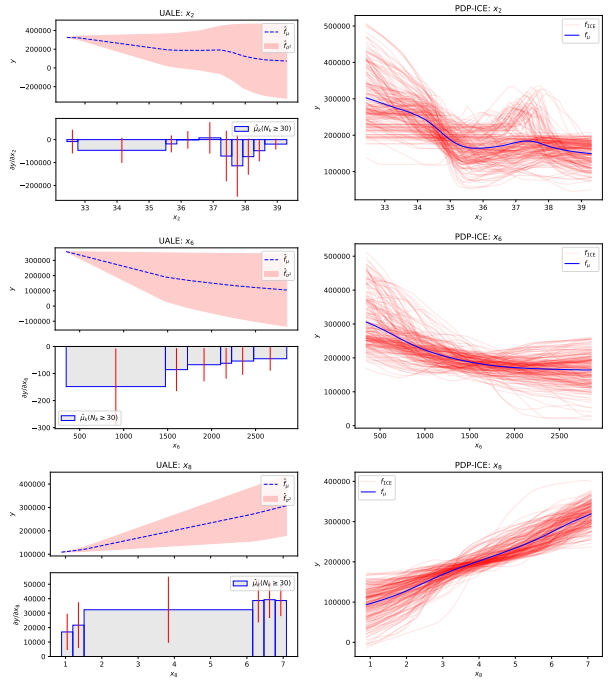

Figure 7: Figure 1

Bin Splitting In this part, we evaluate the robustness of the approximation using automatic bin-splitting. Following the evaluation framework of Section 4.2, 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 randomly select less samples $N = 1000$ and we compare UALE approximation with all possible fixed-size alternatives, repeating this process for 30 times for robust results. In Figure 8, we illustrate the mean values for $\mathcal{L}^{\mu}, \mathcal{L}^{\sigma}$ of the 30 repetitions. We observe that automatic bin-spliting provides (close to the) best approximation in the three features. In the Appendix, we provide the same evaluation for all features.

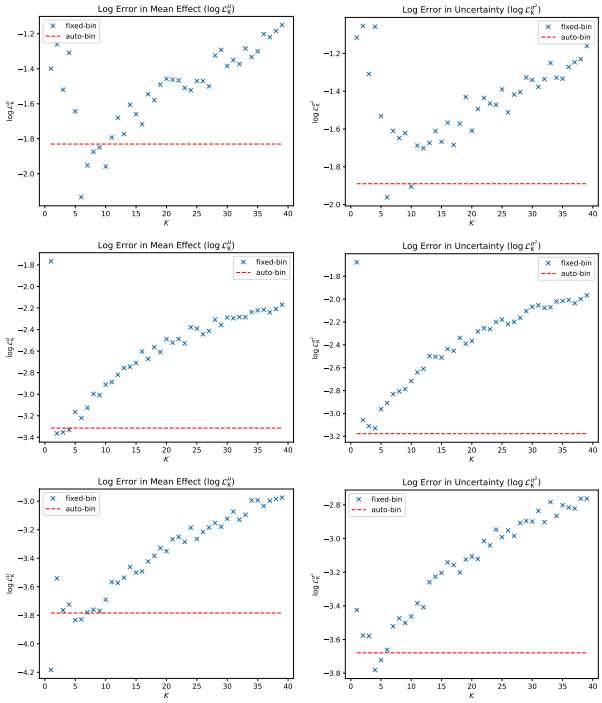

Figure 8: Figure 1

6 DISCUSSION

[Add Conclusion, limitations, and future work]

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.

A List with Symbols