RHALE: Robust and Heterogeneity-aware Accumulated Local Effects
(Supplementary Material)

September 20, 2023

A Theoretical Evidence

In this Section, we provide proofs for the equations used in the main paper.

A.1 Proof that $\hat{\mu}(z_1, z_2)$ is an unbiased estimator of $\mu(z_1, z_2)$

This proof is required for Theorem 1 (Section A.2). We want to show that

$$\hat{\mu}(z_1, z_2) = \frac{1}{|\mathcal{S}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}} f^s(\mathbf{x}^i)$$

is an unbiased estimator of:

$$\mu(z_1, z_2) = \frac{\int_{z_1}^{z_2} \mathbb{E}_{X_c|z} \left[f^s(z, X_c) \right] \partial z}{z_2 - z_1}$$

under the assumptions that (a) z follows a uniform distribution in $[z_1, z_2]$, i.e., $z \sim \mathcal{U}(z_1, z_2)$, (b) $\tilde{X}$ is a random variable with PDF $p(\tilde{\mathbf{x}}) = p(\mathbf{x_c}|z)p(z) = \frac{1}{z_2 - z_1} p(\mathbf{x_c}|z)$ and (c) the points $\mathbf{x}^i$ are i.i.d. samples from $p(\tilde{\mathbf{x}})$. We want to show that $\mathbb{E}_{\tilde{X}}[\hat{\mu}(z_1, z_2)] = \mu(z_1, z_1)$.

Proof Description We show that (a) $\mu(z_1, z_2) = \mathbb{E}_{\tilde{X}}[f^s(\tilde{X})]$ and we use the fact that (b) the population mean is an unbiased estimator of the expected value.

Proof

$$\mu(z_1, z_2) = \frac{\int_{z_1}^{z_2} \mathbb{E}_{X_c|z}[f^s(z, X_c)] \partial z}{z_2 - z_1} = \mathbb{E}_{z \sim \mathcal{U}(z_1, z_2)}[\mathbb{E}_{X_c|z}[f^s(z, X_c)]] = \mathbb{E}_{\tilde{X}}[f^s(\tilde{X})] = \mathbb{E}_{\tilde{X}}[\hat{\mu}(z_1, z_2)] \tag{1}$$

A.2 Proof that $\hat{\sigma}^2(z_1, z_2)$ is an unbiased estimator of $\sigma_*^2(z_1, z_2)$

This equation is used in Section 3.1 of the main paper. We want to show that

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

is an unbiased estimator of

$$\sigma_*^2(z_1, z_2) = \frac{\int_{z_1}^{z_2} \mathbb{E}_{X_c|X_s=z} \left[(f^s(z, X_c) - \mu(z_1, z_2))^2 \right] \partial z}{z_2 - z_1}$$

under the assumptions that (a) z follows a uniform distribution in $[z_1, z_2]$, i.e., $z \sim \mathcal{U}(z_1, z_2)$, (b) $\tilde{X}$ is a random variable with PDF $p(\tilde{\mathbf{x}}) = p(\mathbf{x_c}|z)p(z) = \frac{1}{z_2 - z_1} p(\mathbf{x_c}|z)$ and (c) the points $\mathbf{x}$ are i.i.d. samples from $p(\tilde{\mathbf{x}})$. We want to show that $\mathbb{E}_{\tilde{X}}[\hat{\sigma}^2(z_1, z_2)] = \sigma_*^2(z_1, z_1)$.

Proof Description We show (a) that $\sigma_*^2(z_1, z_2) = \mathbb{E}_{\tilde{X}} \left[(f^s(\tilde{X}) - \mathbb{E}_{\tilde{X}}[\hat{\mu}(z_1, z_2)])^2 \right]$ and then (b) we use the fact that the sample variance is an unbiased estimator of the distribution variance.

Proof

$$\begin{aligned}
\sigma_*^2(z_1, z_2) &= \frac{\int_{z_1}^{z_2} \mathbb{E}_{X_c|z} \left[(f^s(z, X_c) - \mu(z_1, z_2))^2 \right] \partial z}{z_2 - z_1} && (2) \\
&= \mathbb{E}_{z \sim \mathcal{U}(z_1, z_2)} \mathbb{E}_{X_c|z} \left[(f^s(z, X_c) - \mu(z_1, z_2))^2 \right] && (3) \\
&= \mathbb{E}_{\tilde{X}} \left[(f^s(\tilde{X}) - \mu(z_1, z_2))^2 \right] && (4) \\
&= \mathbb{E}_{\tilde{X}} \left[(f^s(\tilde{X}) - \mathbb{E}_{\tilde{X}}[\hat{\mu}(z_1, z_2)])^2 \right] && (5) \\
&= \mathbb{E}_{\tilde{X}} \left[\hat{\sigma}^2(z_1, z_2) \right] && (6)
\end{aligned}$$

A.3 Proof Of Theorem 1

Theorem 3.1 *If we define (a) the residual $\rho(z)$ as the difference between the expected effect at z and the bin effect, i.e, $\rho(z) = \mu(z) - \mu(z_1, z_2)$, and (b) $\mathcal{E}(z_1, z_2)$ as the mean squared residual of the bin, i.e., $\mathcal{E}(z_1, z_2) = \frac{\int_{z_1}^{z_2} \rho^2(z) \partial z}{z_2 - z_1}$, then it holds*

$$\sigma_*^2(z_1, z_2) = \sigma^2(z_1, z_2) + \mathcal{E}^2(z_1, z_2) \tag{7}$$

We want to show that $\sigma_*^2(z_1, z_2) = \sigma^2(z_1, z_2) + \mathcal{E}^2(z_1, z_2)$, where (a) the bin-error $\mathcal{E}^2(z_1, z_2)$ is the mean squared residual of the bin, i.e. $\mathcal{E}^2(z_1, z_2) = \frac{\int_{z_1}^{z_2} \rho^2(z) \partial z}{z_2 - z_1}$ and (b) the residual $\rho(z)$ is the difference between the expected effect at z and the bin effect, i.e $\rho(z) = \mu(z) - \mu(z_1, z_2)$.

Proof Description We use that $\forall z \in [z_1, z_2]$, it holds that $\mu(z_1, z_2) = \mu(z) - \rho(z)$ and then we split the terms appropriately to complete the proof.

Proof

$$
\begin{aligned}
\sigma_*^2(z_1, z_2) &= \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{X_c|z} \left[\left(f^s(z, X_c) - \mu(z_1, z_2) \right)^2 \right] \partial z && (8) \\
&= \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{X_c|z} \left[\left(f^s(z, X_c) - \mu(z) + \rho(z) \right)^2 \right] \partial z && (9) \\
&= \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{X_c|z} \left[(f^s(z, X_c) - \mu(z))^2 + \rho(z)^2 + 2(f^s(z, X_c) - \mu(z))\rho(z) \right] \partial z && (10) \\
&= \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \left(\underbrace{\mathbb{E}_{X_c|z} \left[(f^s(z, X_c) - \mu(z))^2 \right]}_{\sigma^2(z)} + \underbrace{\mathbb{E}_{X_c|z} \left[\rho^2(z) \right]}_{\rho^2(z)} + 2(\underbrace{\mathbb{E}_{X_c|z} \left[(f^s(z, X_c) \right]}_{\mu(z)} - \mu(z))\rho(z)) \right) \partial z && (11) \\
&= \underbrace{\frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \sigma^2(z) \partial z}_{\sigma^2(z_1, z_2)} + \underbrace{\frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \rho^2(z) \partial z}_{\mathcal{E}^2(z_1, z_2)} && (12) \\
&= \sigma^2(z_1, z_2) + \mathcal{E}^2(z_1, z_2) && (13)
\end{aligned}
$$

A.4 Proof Of Corollary 2

We want to show that, *if a bin-splitting $\mathcal{Z}$ minimizes the accumulated error, then it also minimizes* $\sum_{k=1}^{K} \sigma_*^2(z_1, z_2)\Delta z_k$. In mathematical terms, we want to show that:

$$
\mathcal{Z}^* = arg\, min_{\mathcal{Z}} \sum_{k=1}^{K} \sigma_*^2(z_{k-1}, z_k)\Delta z_k \Leftrightarrow \mathcal{Z}^* = arg\, min_{\mathcal{Z}} \sum_{k=1}^{K} \mathcal{E}^2(z_{k-1}, z_k)\Delta z_k
$$

Proof Description The key-point for the proof is that the term $\sum_{k=1}^{K} \sigma^2(z_{k-1}, z_k)\Delta z_k$ is independent of the bin partitioning $\mathcal{Z}$. In Eq.(16) we use Eq.8 of the main paper.

Proof

$$\mathcal{Z}^* = arg\,min_{\mathcal{Z}} \sum_{k=1}^{K} \sigma_*^2(z_{k-1}, z_k)\Delta z_k \tag{14}$$

$$= arg\,min_{\mathcal{Z}} \left[\sum_{k=1}^{K} (\sigma^2(z_{k-1}, z_k) + \mathcal{E}^2(z_{k-1}, z_k))\Delta z_k \right] \tag{15}$$

$$= arg\,min_{\mathcal{Z}} \left[\sum_{k=1}^{K} \left(\frac{\Delta z_k}{\Delta z_k} \int_{z_{k-1}}^{z_k} \sigma^2(z)\partial z + \mathcal{E}^2(z_{k-1}, z_k)\Delta z_k \right) \right] \tag{16}$$

$$= arg\,min_{\mathcal{Z}} \left[\underbrace{\int_{z_0}^{z_K} \sigma^2(z)\partial z}_{\text{independent of } \mathcal{Z}} + \sum_{k=1}^{K} \mathcal{E}^2(z_{k-1}, z_k)\Delta z_k) \right] \tag{17}$$

$$= arg\,min_{\mathcal{Z}} \sum_{k=1}^{K} \mathcal{E}^2(z_{k-1}, z_k)\Delta z_k \tag{18}$$

A.5 Dynamic Programming

We denote with $i \in \{0, \dots, K_{max}\}$ the index of point x_i, as defined at Section 3.2 of the main paper, and with z_j and z_{j+1} the chosen limits (out of the values x_i) for bin j. The states of the problem are then represented by matrices $\mathcal{C}(i, j)$ and $\mathcal{I}(i, j)$. $\mathcal{C}(i, j)$ is the cost of setting $z_{j+1} = x_i$, i.e., the cost of setting the right limit of the j-th bin to x_i, and is computed by the recursive function:

$$\mathcal{C}(i, j) = \begin{cases} \min_{i \in \{0,\dots,K_{max}\}} \left[C(i, j-1) + \mathcal{B}(i, j) \right], & \text{if } j > 0 \\ \mathcal{B}(i, j) & \text{if } j = 0 \end{cases} \tag{19}$$

$\mathcal{I}(i, j)$ is an index matrix indicating the selected values z_j, i.e., the values indicating the right limit of $j-1$ bins. In other words, $z_j = x_{\mathcal{I}(i,j)}$. The value of $\mathcal{I}(i, j)$ is given by $\mathcal{I}(i, j) = \text{argmin}_{i \in \{0,\dots,K_{max}\}} \left[C(i, j-1) + \mathcal{B}(i, j) \right]$. Note that although this procedures always selects $K_{max} + 1$ values for z_j, some of them may be the same point corresponding to zero-width bins. These are dropped when choosing the optimal bin limits $\mathcal{Z}$. Algorithm 1 presents the use of dynamic programming to solve the optimization problem of Eq.13.

B Empirical Evaluation

B.1 Running Example

In the running example, the data generating distribution is $p(\mathbf{x}) = p(x_1)p(x_2)p(x_3|x_1)$, where $p(x_1) = \frac{5}{6}\mathcal{U}(x_1; -0.5, 0) + \frac{1}{6}\mathcal{U}(x_1; 0, 0.5)$, $p(x_2) = \mathcal{N}(x_2; \mu_2 = 0, \sigma_2 = 2)$ and $p(x_3) = \mathcal{N}(x_3; \mu_3 = x_1, \sigma_3 = 0.01)$. So, x_1 is highly correlated with x_3, while x_2 is independent from both x_1 and x_3. The black-box function is:

$$f(\mathbf{x}) = \underbrace{\sin(2\pi x_1)(\mathbb{1}_{x_1<0} - 2\mathbb{1}_{x_3<0})}_{g_1(\mathbf{x})} + \underbrace{x_1 x_2}_{g_2(\mathbf{x})} + \underbrace{x_2}_{g_3(\mathbf{x})} \tag{20}$$

Algorithm 1 Algorithm for solving the optimization problem with dynamic programming

Input: $\mathcal{B}(i,j)$: function that gives the cost of bin $[x_i, x_j)$, K_{max}: max number of bins
Output: $\mathcal{Z}$: the optimal partitioning

```
  C(i, j) = +∞, ∀i, j                                   ▷ Initiate the cost matrix with +∞
  I(i, j) = 0, ∀i, j                                    ▷ Initiate the index matrix with 0
  C(i, 0) = B(0, i)∀i                                   ▷ Set cost of the first bin
  for j = 0, ..., K_max − 1 do
      for i = 0, ..., K_max do
          for k = 0, ..., K_max do
              L(k) = C(k, j − 1) + B(k, j)
          end for
          C(i, j) = min_k L(k)
          I(i, j) = arg min_k L(k)
      end for
  end for
  Z(j) = 0∀j = {0, ..., K_max}                          ▷ Initialize list with limits
  Z(0) = 0, Z(K_max) = K_max,                           ▷ First and last limit are always the same
  for j = K_max − 1, ..., 1 do
      Z(j) = I(j, Z(j + 1))                             ▷ Follow the inverse indexes
  end for
  Invert Z and drop Z items that show to the same point
  Z ← x_min + Z(j)ΔX_min                                ▷ Convert indexes to points
```

Ground truth effect. For $g_1(\mathbf{x})$, $x_1 \approx x_3$ so $\mathbb{1}_{x_1<0} - 2\mathbb{1}_{x_3<0} = -\mathbb{1}_{x_1<0}$ and therefore $g_1(x_1) = -\sin(2\pi x_1)\mathbb{1}_{x_1<0}$. For $g_2(\mathbf{x})$, x_2 is independent from x_1, so $\mathbb{E}_{x_2|x_1}[x_1 x_2] = \mathbb{E}_{x_2}[x_1 x_2] = x_1 \mathbb{E}_{x_2}[x_2] = 0$ and therefore $g_2(x_1) = 0$. For $g_3(\mathbf{x})$, it does not include x_1, so $g_3(x_1) = 0$. Therefore, the ground truth feature effect is

$$f^{\mathtt{GT}}(x_1) = -\sin(2\pi x_1)\mathbb{1}_{x_1<0} \tag{21}$$

Ground truth heterogeneity. For the heterogeneity, it is not easy to compute the ground truth, because each method defines and visualizes it in a different way. However, we use the fact that the heterogeneity is induced by the variability of the interaction terms. For $g_1(\mathbf{x})$, $x_1 \approx x_3$ so $\mathbb{1}_{x_1<0} - 2\mathbb{1}_{x_3<0} = -\mathbb{1}_{x_1<0}$ and therefore g_1 does not introduce variability. The variability of $g_3(\mathbf{x})$ is also zero. The only term with variability is $g_2(\mathbf{x}) = x_1 x_2$. Since x_1, x_2 are independent the effect of this term varies according to the variation of x_2 that has a standard deviation of σ_2. Therefore, independently of how each method computes the heterogeneity, the user should be able to understand a variation of σ_2 on the local effects.

RHALE. We compute in an analytic form the feature effect $f^{\mathtt{RHALE}}(x_1)$ and the heterogeneity $\sigma(z)$ for the RHALE method.

$$f^{\texttt{RHALE}}(x_1) = \int_{x_{1,min}}^{x_1} \mathbb{E}_{x_2,x_3|z}\left[\frac{\partial f}{\partial x_1}(z, x_2, x_3)\right]\partial z \tag{22}$$

$$= \int_{x_{1,min}}^{x_1} (\mathbb{E}_{x_3|z}[2\pi z\cos(2\pi z)(\mathbb{1}_{z<0} - 2\mathbb{1}_{x_3<0})] + \underbrace{\mathbb{E}_{x_2|z}[x_2]}_{0})\partial z \tag{23}$$

$$= \int_{x_{1,min}}^{x_1} 2\pi z\cos(2\pi z)\mathbb{E}_{x_3|z}[(\mathbb{1}_{z<0} - 2\mathbb{1}_{x_3<0})]\partial z \tag{24}$$

$$\approx \int_{x_{1,min}}^{x_1} \underbrace{2\pi z\cos(2\pi z)(-\mathbb{1}_{z<0})}_{\mu(z)}\partial z \tag{25}$$

$$\approx -\sin(2\pi x_1)\mathbb{1}_{x_1<0} \tag{26}$$

$$\sigma^2(z) = \mathbb{E}_{x_2,x_3|z}\left[\left(\frac{\partial f}{x_1}(z, x_2, x_3) - \mu(z)\right)^2\right] \tag{27}$$

$$= \mathbb{E}_{x_2,x_3|z}\left[(2\pi z\cos(2\pi z)(\mathbb{1}_{z<0} - 2\mathbb{1}_{x_3<0}) + x_2 - 2\pi z\cos(2\pi z)(-\mathbb{1}_{z<0}))^2\right] \tag{28}$$

$$= \mathbb{E}_{x_2,x_3|z}\left[(2\pi z\cos(2\pi z)(2\mathbb{1}_{z<0} - 2\mathbb{1}_{x_3<0}) + x_2)^2\right] \tag{29}$$

$$= (4\pi z\cos(2\pi z))^2\mathbb{E}_{x_3|z}[(\mathbb{1}_{z<0} - \mathbb{1}_{x_3<0})^2] + \mathbb{E}_{x_2|z}[x_2^2] + \mathbb{E}_{x_2,x_3|z}[4\pi z\cos(2\pi z)(\mathbb{1}_{z<0} - \mathbb{1}_{x_3<0})x_2] \tag{30}$$

$$= (4\pi z\cos(2\pi z))^2\mathbb{E}_{x_3|z}[(\mathbb{1}_{z<0} - \mathbb{1}_{x_3<0})^2] + \sigma_2^2 + \mathbb{E}_{x_2|z}[x_2]\underbrace{\mathbb{E}_{x_3|z}[4\pi z\cos(2\pi z)(\mathbb{1}_{z<0} - \mathbb{1}_{x_3<0})]}_{0} \tag{31}$$

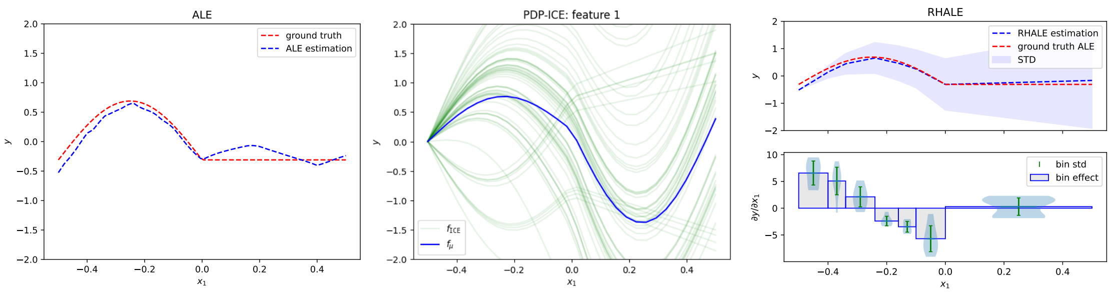
$$= (4\pi z\cos(2\pi z))^2\mathbb{E}_{x_3|z}[(\mathbb{1}_{z<0} + \mathbb{1}_{x_3<0} - 2\mathbb{1}_{z<0}\mathbb{1}_{x_3<0}] + \sigma_2^2 \tag{32}$$

$$= (4\pi x_1\cos(2\pi x_1))^2(2\mathbb{1}_{z<0} - 2\mathbb{1}_{z<0}) + \sigma_2^2 \tag{33}$$

$$= \sigma_2^2 \tag{34}$$

PDP-ICE. We compute in an analytic form the feature effect $f^{\texttt{PDP}}(x_1)$ and the heterogeneity heterogeneity visualized by $f^{\texttt{ICE}}(x_1)$.

The PDP effect uses



(a) ALE plot (50 fixed-size bins) (b) PDP-ICE plot (c) RHALE plot

Figure 1: Feature effect of x_1 on the example defined by Equation 20. ALE does not quantify the heterogeneity and fixed-size splitting leads to a bad estimation. PDP-ICE plots fail in both main effect and heterogeneity, failing to capture feature correlations. RHALE, on the other hand, provides a robust estimation of the main effect and the heterogeneity.

$$
\begin{aligned}
f^{\texttt{PDP}}(x_1) &= \mathbb{E}_{x_2,x_3}[f(\mathbf{x})] && (35)\\
&= \sin(2\pi x_1)\mathbb{E}_{x_3}\left[\mathbb{1}_{x_1<0} - 2\mathbb{1}_{x_3<0}\right] + \mathbb{E}_{x_2}[x_1x_2] + \mathbb{E}_{x_2}[x_2] && (36)\\
&= \sin(2\pi x_1)(\mathbb{1}_{x_1<0} - 2\mathbb{E}_{x_3}[\mathbb{1}_{x_3<0}]) + \underbrace{x_1\mathbb{E}_{x_2}[x_2]}_{0} + \underbrace{\mathbb{E}_{x_2}[x_2]}_{0} && (37)\\
&= \sin(2\pi x_1)\left(\mathbb{1}_{x_1<0} - 2\int_{-0.5}^{0.5}\mathbb{1}_{x_3<0}p(x_3)\partial x_3\right) && (38)\\
&= \sin(2\pi x_1)\left(\mathbb{1}_{x_1<0} - 2\int_{-0.5}^{0} 2\frac{5}{6}\mathbb{1}_{x_3<0}\partial x_3 + \int_{0}^{0.5} 2\frac{1}{6}\mathbb{1}_{x_3<0}\partial x_3\right) && (39)\\
&= \sin(2\pi x_1)\left(\mathbb{1}_{x_1<0} - 2\frac{5}{6}\right) && (40)
\end{aligned}
$$

For the ICE plots:

$$
\begin{aligned}
f^{\texttt{ICE}}(x_1^i) &= \sin(2\pi x_1)(\mathbb{1}_{x_1<0} - 2\mathbb{1}_{x_3^i<0}) + x_1x_2^i + x_2^i && (41)\\
&= \sin(2\pi x_1)(\mathbb{1}_{x_1<0} - 2\mathbb{1}_{x_3^i<0}) + x_1x_2^i + c && (42)
\end{aligned}
$$

So if $x_3^i < 0$, which happens in almost $\frac{5}{6}$ of the instances, then $f^{\texttt{ICE}}(x_1^i)(x_1) = -\sin(2\pi x_1) + x_1x_2^i + c$, and in almost $\frac{1}{6}$ of the instances, $f^{\texttt{ICE}}(x_1^i)(x_1) = \sin(2\pi x_1) + x_1x_2^i + c$.

Discussion. The derivations above are reflected in Figure 1. We observe that PDP and ICE provide misleading explanations which are *not* due to some approximation error, e.g., due to limited samples. As shown by Equation 35 and Equation 41 PDP and ICE systematically produce misleading explanations [Apley and Zhu, 2020] for the feature effect and the heterogeneity in cases of correlated features. In contrast, we confirm our previous knowledge that ALE handles well these cases and we observe that the deviation from the ground is only due to approximation issues, which are addressed by RHALE.

B.2 Simulation Study

The data generating distribution is $p(\mathbf{x}) = p(x_3)p(x_2|x_1)p(x_1)$, where $x_1 \sim \mathcal{U}(0,1)$, $x_2 = x_1 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.01)$ is a small additive, noise and $x_3 \sim \mathcal{N}(0, \sigma_3^2 = \frac{1}{4})$. The predictive function is:

$$f(\mathbf{x}) = \underbrace{\alpha f_2(\mathbf{x})}_{g_3(\mathbf{x})} + \underbrace{f_1(\mathbf{x})\mathbb{1}_{f_1(\mathbf{x})\leq \frac{1}{2}}}_{g_1(\mathbf{x})} + \underbrace{(1 - f_1(\mathbf{x}))\mathbb{1}_{\frac{1}{2}<f_1(\mathbf{x})<1}}_{g_2(\mathbf{x})} \tag{43}$$

where $f_1(\mathbf{x}) = a_1x_1 + a_2x_2$ is a linear combination of x_1, x_2, and $f_2(\mathbf{x}) = x_1x_3$ interacts the non-correlated features x_1, x_3. We evaluate the effect computed by RHALE 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 weights ($a_1 \neq a_2$) and (c) with interaction ($\alpha > 0$) and equal weights ($a_1 = a_2$).

Ground truth for case (a) In this case, the weights are $a_1 = a_2 = 1$ and there is no interaction term $\alpha = 0$). Therefore:

$$f(\mathbf{x}) = f_1(\mathbf{x})\mathbb{1}_{f_1(\mathbf{x})\leq \frac{1}{2}} + (1 - f_1(\mathbf{x}))\mathbb{1}_{\frac{1}{2}<f_1(\mathbf{x})<1} \tag{44}$$

where $f_1(\mathbf{x}) = x_1 + x_2$. For the ground truth feature effect, we use the fact that $x_1 \approx x_2$, therefore knowing only the value of x_1 we can automatically infer the value of x_2 and therefore the value of $f_1(\mathbf{x})$. For example, when $0 \leq x_1 \leq \frac{1}{4}$ then $0 \leq f_1(\mathbf{x}) \leq \frac{1}{2}$ and, therefore, $f_1(x_1) = a_1x_1$. In a similar way, we compute the effect of x_2. The effect of x_3 is zero.

$$f^{\mathtt{GT}}(x_1) = x_1\mathbb{1}_{0\leq x_1\leq \frac{1}{4}} + \left(\frac{1}{4} - x_1\right)\mathbb{1}_{\frac{1}{4}<x_1<\frac{1}{2}} \tag{45}$$

$$f^{\mathtt{GT}}(x_2) = x_2\mathbb{1}_{0\leq x_2\leq \frac{1}{4}} + \left(\frac{1}{4} - x_2\right)\mathbb{1}_{\frac{1}{4}<x_2<\frac{1}{2}} \tag{46}$$

$$f^{\mathtt{GT}}(x_3) = 0 \tag{47}$$

The heterogeneity is zero for all features because the heterogeneity is induced by the variability of the interaction terms and, since, $x_1 \approx x_2$, the terms $\mathbb{1}_{f_1(\mathbf{x})\leq\frac{1}{2}}$ and $\mathbb{1}_{\frac{1}{2}<f_1(\mathbf{x})\leq 1}$, do not vary.

Ground truth for case (b) In this case, the weights are $a_1 = 2$ and $a_2 = \frac{1}{2}$ and there is no interaction term $\alpha = 0$. Therefore:

$$f(\mathbf{x}) = f_1(\mathbf{x})\mathbb{1}_{f_1(\mathbf{x})\leq \frac{1}{2}} + (1 - f_1(\mathbf{x}))\mathbb{1}_{\frac{1}{2}<f_1(\mathbf{x})<1} \tag{48}$$

where $f_1(\mathbf{x}) = 2x_1 + \frac{1}{2}x_2$. As in case (a), we use again the fact that $x_1 \approx x_2$, to compute the ground truth feature effect:

$$f^{\mathtt{GT}}(x_1) = 2x_1\mathbb{1}_{0\leq x_1\leq \frac{1}{5}} + (\frac{2}{5} - 2x_1)\mathbb{1}_{\frac{1}{4}<x_1<\frac{2}{5}} \tag{49}$$

$$f^{\mathtt{GT}}(x_2) = 2x_2\mathbb{1}_{0\leq x_2\leq \frac{1}{5}} + (\frac{2}{5} - 2x_2)\mathbb{1}_{\frac{1}{4}<x_2<\frac{2}{5}} \tag{50}$$

$$f^{\mathtt{GT}}(x_3) = 0 \tag{51}$$

The heterogeneity is zero for all features because the heterogeneity is induced by the variability of the interaction terms and, since, $x_1 \approx x_2$, the terms $\mathbb{1}_{f_1(\mathbf{x})\leq\frac{1}{2}}$ and $\mathbb{1}_{\frac{1}{2}<f_1(\mathbf{x})\leq 1}$, do not vary.

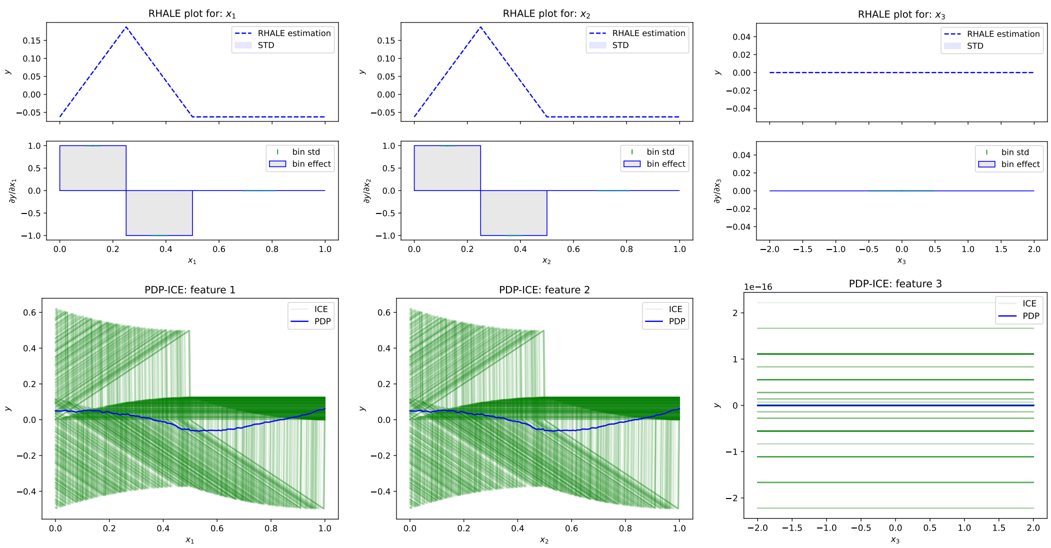


Figure 2: Case (a)

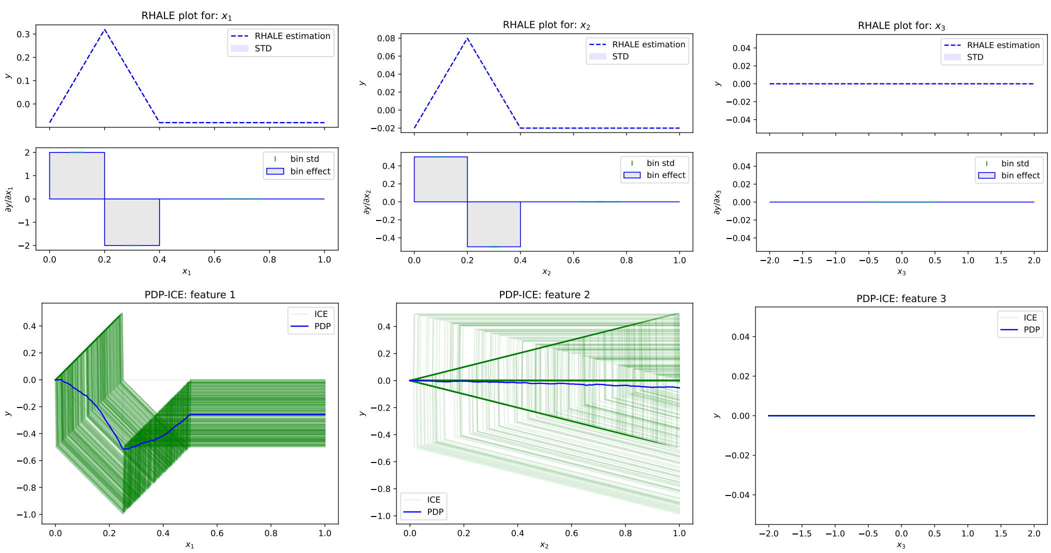


Figure 3: Case (b)

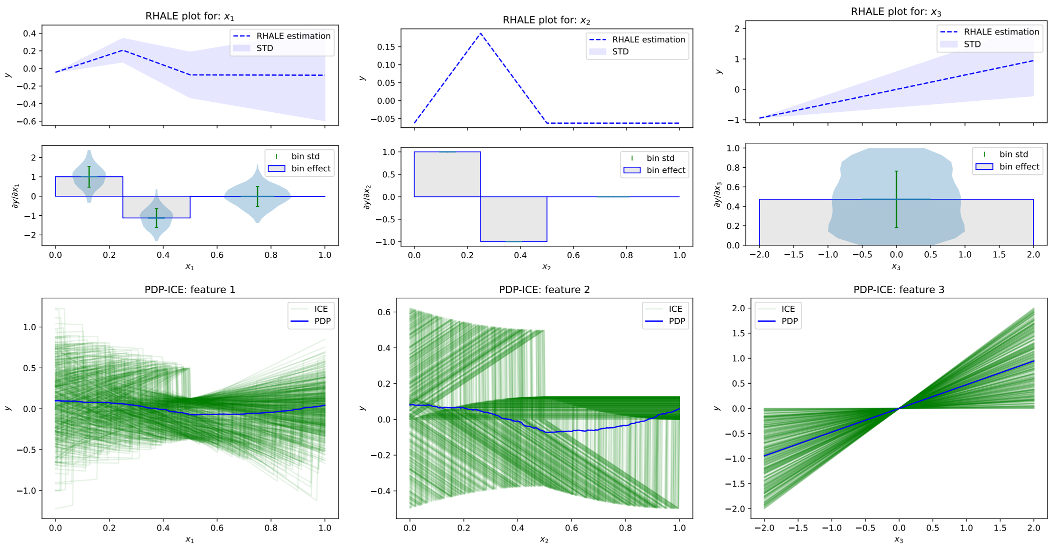


Figure 4: Case (b)

Ground truth for case (c) In this case, the weights are equal $a_1 = a_2 = 1$ and the is interaction term is enabled ($\alpha = 1$). Therefore:

$$f(\mathbf{x}) = f_2(\mathbf{x}) + f_1(\mathbf{x})\mathbb{1}_{f_1(\mathbf{x})\leq\frac{1}{2}} + (1 - f_1(\mathbf{x}))\mathbb{1}_{\frac{1}{2}<f_1(\mathbf{x})<1} \tag{52}$$

where $f_2(\mathbf{x}) = x_1x_3$ and $f_1(\mathbf{x}) = x_1 + x_2$. The feature effect of terms $f_1(\mathbf{x})\mathbb{1}_{f_1(\mathbf{x})\leq\frac{1}{2}} + (1 - f_1(\mathbf{x}))\mathbb{1}_{\frac{1}{2}<f_1(\mathbf{x})<1}$ are exactly the same with case (a). The term $f_2(\mathbf{x}) = x_1x_2$. For feature x_1 the effect is $\mathbb{E}_{x_3|x_1}[x_1x_3] = x_1\mathbb{E}_{x_3}[x_3] = 0$ and for feature x_2 the effect is $\mathbb{E}_{x_1|x_3}[x_1x_3] = x_3\mathbb{E}_{x_1}[x_1] = 0.5x_3$. Therefore, the ground truth feature effect is:

$$f^{\mathtt{GT}}(x_1) = x_1\mathbb{1}_{0\leq x_1\leq\frac{1}{4}} + \left(\frac{1}{4} - x_1\right)\mathbb{1}_{\frac{1}{4}<x_1<\frac{1}{2}} \tag{53}$$

$$f^{\mathtt{GT}}(x_2) = x_2\mathbb{1}_{0\leq x_2\leq\frac{1}{4}} + \left(\frac{1}{4} - x_2\right)\mathbb{1}_{\frac{1}{4}<x_2<\frac{1}{2}} \tag{54}$$

$$f^{\mathtt{GT}}(x_3) = \frac{1}{2}x_3 \tag{55}$$

For the same reason with cases (a) and (b), the terms $\mathbb{1}_{f_1(\mathbf{x})\leq\frac{1}{2}}$ and $\mathbb{1}_{\frac{1}{2}<f_1(\mathbf{x})\leq1}$, do not introduce heterogeneity. Since x_1, x_2 are independent the effect of x_1x_3 varies. For feature x_1, it varies following the standard deviation of x_3, i.e. $\sigma_3 = \frac{1}{2}$ and for feature x_3, it varies following the standard deviation of x_1, i.e. $\sigma_1 = \frac{1}{4}$.

Conclusion. The example confirms our previous knowledge that PDP-ICE provide erroneous effects in cases with correlated features. The feature effect computed by PDP and the heterogeneity illustrated by ICE are correct only for feature x_3, because it is independent from the other features. For features the correlated features x_1, x_2, both PDP and ICE provide misleading explanations. In contrast, RHALE handles well all cases, providing accurate estimations for the feature effects and the heterogeneity.

Table 1: Description of the features apparent in the California-Housing Dataset

	Description	*min*	max	μ	σ
x_1	longitude	-124.35	-114.31	-119.58	2
x_2	latitude	32.54	41.95	35.65	2.14
x_3	median age of houses	1	52	29.01	12.42
x_4	total number of rooms	2	9179	2390.79	1433.83
x_5	total number of bedrooms	2	1797	493.86	291
x_6	total number of people	3	4818	1310.91	771.78
x_7	total number of households	2	1644	460.3	267.34
x_8	median income of households	0.5	9.56	3.72	1.60
y	median house value	14.999	500000	206864.41	115435.67

B.3 Real World Experiment

In this section, we provide further details on the real-world example. The real-world example uses the California Housing Dataset, which contains 8 numerical features. We exclude instances with missing or outlier values. If we denote as μ_s (σ_s) the average value (standard deviation) of the s-th feature, we consider outliers the instances of the training set with any feature value over three standard deviations from the mean, i.e. $|x_s^i - \mu_s| > \sigma_s$. This preprocessing step discards 884 instances, and $N = 19549$ remain. We provide their description with some basic descriptive statistics in Table 1 and their histogram in Figure 5.

In Figure 7 of the main paper, we provided the RHALE vs PDP-ICE plots for features x_2 (latitude), x_6 (total number of people) and x_8 (median house value). In figure 8, we compared RHALE with fixed-size approximation, for the same features. In Figure 6, we provide the same information for the rest of the features; x_1 (longitude), x_3 (median age of houses), x_4 (total number of rooms), x_5 (total number of bedrooms) and x_7 (total number of households). The observation of these features leads us to similar conclusion. First, RHALE and PDP-ICE plots compute similar effects and level of heterogeneity and RHALE's approximation is (almost) as good as the best fixed-size approximation. More specifically, we observe that RHALE's variable size bin splitting correctly creates wide bins for features x_3, x_4, x_5, x_7, where the feature effect plot is (piecewise) linear, while using narrow bins for feature x_2 where the feature effect is not linear.

References

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.

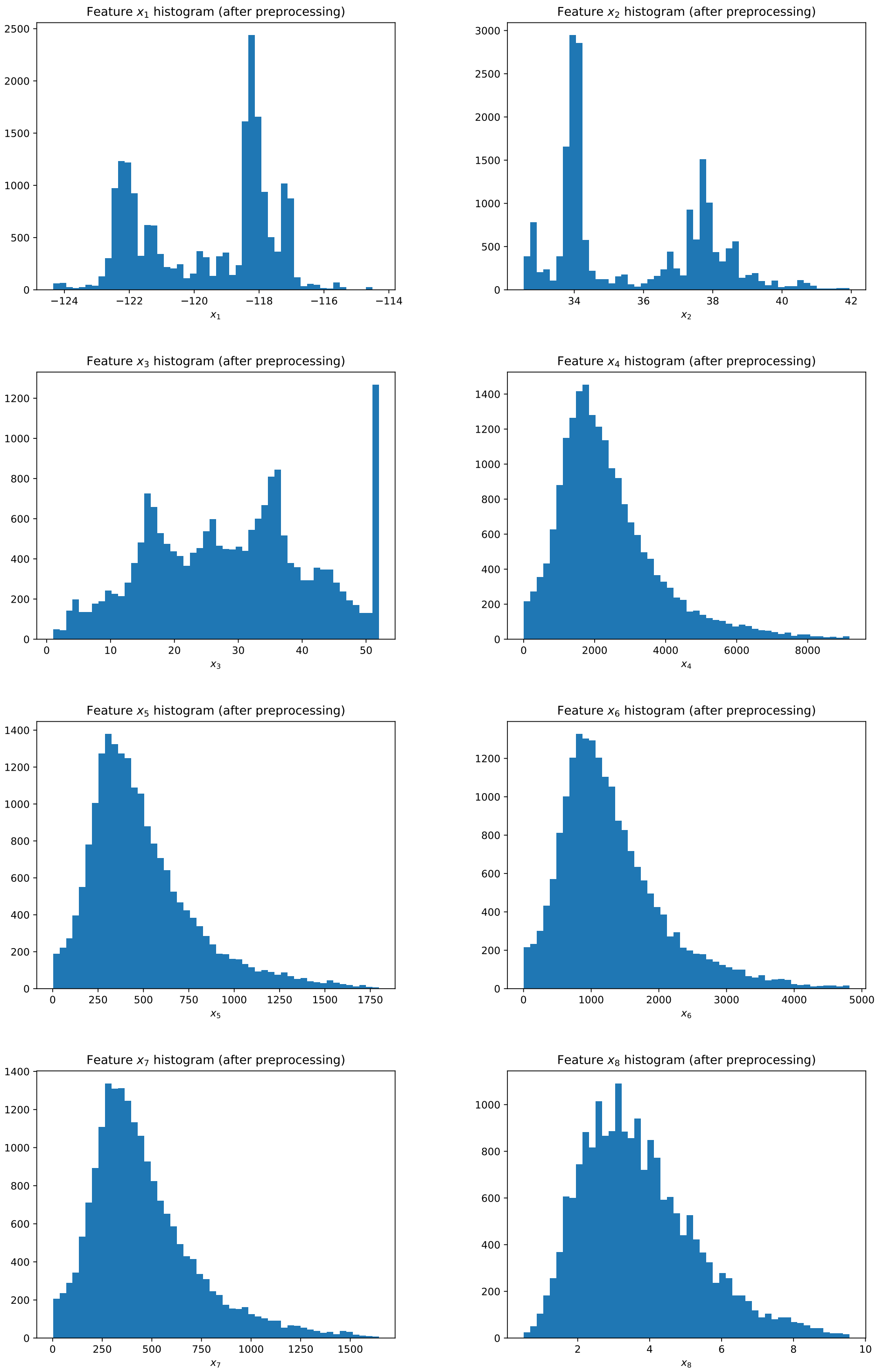


Figure 5: The Histogram of each feature in the California Housing Dataset.

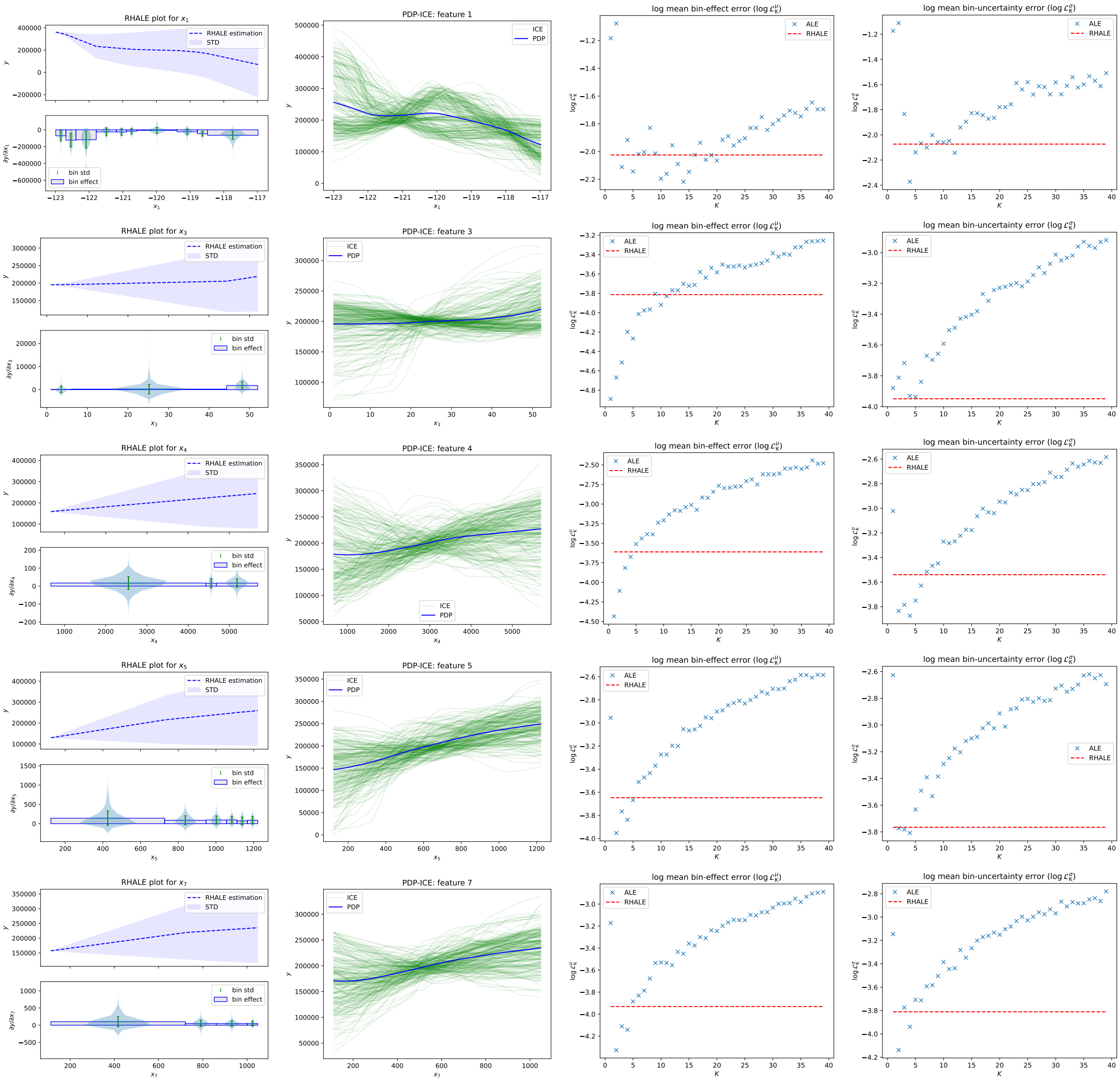


Figure 6: From left to right: (a) RHALE plot, (b) PDP-ICE plot, (c) RHALE vs fixed-size $\mathcal{L}^{\mu}$ and (d) RHALE vs fixed-size $\mathcal{L}^{\sigma}$. From top to bottom, features $x_1, x_3, x_4, x_5, x_7, x_8$.