Instructions for Paper Submissions to AISTATS 2023: Supplementary Materials

1 THORETICAL DERIVATIONS

1.1 Show that $\hat{\mu}(z_1, z_2) \approx \mu(z_1, z_2)$

We want to show that (a) $\hat{\mu}(z_1, z_2) = \frac{1}{|\mathcal{S}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}} f^s(\mathbf{x}^i)$ is an unbiased estimator of $\mu(z_1, z_2) = \frac{\int_{z_1}^{z_2} \mathbb{E}_{X_c|z}[f^s(z, X_c)] \partial z}{z_2 - z_1}$, under the assumption that that (a) z follows a uniform distribution in $[z_1, z_2)$, i.e., $z \sim \mathcal{U}(z_1, z_2)$ and (b) that the points are i.i.d. samples from the distribution $p(\mathbf{x}) = p(\mathbf{x_c}|z)p(z) = \frac{1}{z_2 - z_1} p(\mathbf{x_c}|z)$

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

Proof

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

1.2 Show that $\hat{\sigma}^2(z_1, z_2) \approx \sigma_*^2(z_1, z_2)$

We want to show that $\hat{\sigma}^2(z_1, z_2) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left(\frac{\partial f}{\partial x_s}(\mathbf{x}^i) - \hat{\mu}(z_1, z_2) \right)^2$ 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 assumption that that (a) z follows a uniform distribution in $[z_1, z_2)$, i.e., $z \sim \mathcal{U}(z_1, z_2)$ and (b) that the points are i.i.d. samples from the distribution $p(\mathbf{x}) = p(\mathbf{x_c}|z)p(z) = \frac{1}{z_2 - z_1} p(\mathbf{x_c}|z)$.

Description Same as before, we just use the the fact that the population variance is an unbiased estimator of the variance. We just show that $\sigma_*^2(z_1, z_2) = \mathbb{V}_{\tilde{X}}[f^s(\tilde{X})]$.

Proof

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

1.3 Proof Of 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 that:

$$\sigma_*^2(z_1, z_2) = \sigma^2(z_1, z_2) + \mathcal{E}^2(z_1, z_2) \tag{6}$$

Proof

$$\sigma_*^2(z_1, z_2) = \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{X_c|z} \left[(f^s(z, X_c) - \mu(z_1, z_2))^2 \right] \partial z \tag{7}$$

$$= \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{X_c|z} \left[(f^s(z, X_c) - \mu(z) + \rho(z))^2 \right] \partial z \tag{8}$$

$$= \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \mathbb{E}_{X_c|z} \left[(f^s(z, X_c) - \mu(z))^2 + \rho(z)^2 + 2 f^s(z, X_c)\mu(z)) \right] \tag{9}$$

$$= \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \left(\underbrace{\mathbb{E}_{X_c|z} \left[(f^s(z, X_c) - \mu(z))^2 \right]}_{\sigma^2(z)} + \underbrace{\mathbb{E}_{X_c|z} \left[\rho^2(z) \right]}_{\rho^2(z)} + 2(\underbrace{\mathbb{E}_{X_c|z} \left[(f^s(z, X_c) \right]}_{\mu(z)} - \mu(z))\rho(z)) \right) \partial z \tag{10}$$

$$= \underbrace{\frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \sigma^2(z) \partial z}_{\sigma^2(z_1, z_2)} + \underbrace{\frac{1}{z_2 - z_1} \int_{z_1}^{z_2} \rho^2(z) \partial z}_{\mathcal{E}^2(z_1, z_2)} = \sigma^2(z_1, z_2) + \mathcal{E}^2(z_1, z_2) \tag{11}$$

1.4 Proof Of Corollary

If a bin-splitting $\mathcal{Z}$ minimizes the accumulated error, then it also minimizes $\sum_{k=1}^{K} \sigma_*^2(z_1, z_2)\Delta z_k$

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

$$\mathcal{Z}^* = \arg\min_{\mathcal{Z}} \sum_{k=1}^{K} \sigma_*^2(z_{k-1}, z_k)\Delta z_k \tag{12}$$

$$= \arg\min_{\mathcal{Z}} \left[\sum_{k=1}^{K} (\sigma^2(z_{k-1}, z_k) + \mathcal{E}^2(z_{k-1}, z_k))\Delta z_k \right] \tag{13}$$

$$= \arg\min_{\mathcal{Z}} \left[\sum_{k=1}^{K} \left(\frac{\Delta z_k}{\Delta z_k} \int_{z_{k-1}}^{z_k} \sigma^2(z)\partial z) + \mathcal{E}^2(z_{k-1}, z_k)\Delta z_k \right) \right] \tag{14}$$

$$= \arg\min_{\mathcal{Z}} \left[\underbrace{\int_{z_0}^{z_K} \sigma^2(z)\partial z}_{\text{independent of } \mathcal{Z}} + \sum_{k=1}^{K} \mathcal{E}^2(z_{k-1}, z_k)\Delta z_k) \right] \tag{15}$$

$$= \arg\min_{\mathcal{Z}} \sum_{k=1}^{K} \mathcal{E}^2(z_{k-1}, z_k)\Delta z_k \tag{16}$$

2 Dynamic Programming Analysis

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 $i \in \{0, \ldots, K_{max}\}$ denotes the point (z_i) and the index $j \in \{0, \ldots, K_{max}\}$ 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 $z_i = x_j$:

$$\mathcal{T}(i, j) = \min_{l \in \{0, \ldots, K_{max}\}} \left[\mathcal{T}(i-1, l) + \mathcal{B}(x_l, x_j) \right] \tag{17}$$

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

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 $[x_l, x_j)]$:

$$\mathcal{B}(x_l, x_j) = \begin{cases} \infty, & \text{if } x_j > x_l \text{ or } |\mathcal{S}_{(x_j, x_l)}| < N \\ 0, & \text{if } x_j = x_l \\ \hat{\sigma}^2(x_j, x_l), & \text{if } x_j \leq x_l \end{cases} \tag{18}$$

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

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 instancies, and $N = 19549$ remain. We provide their description with some basic descriptive statistics in Table 1 and their histogram in Figure 3.

In Figure 7 of the main paper, we provided the UALE 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 UALE with fixed-size approximation, for the same features. In Figure 2, 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 feautures leads us to similar conclusion. First, UALE and PDP-ICE plots compute similar effects and level of heterogeneity and UALE's approximation is (almost) as good as the best fixed-size approximation. More specifically, we observe that UALE'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.

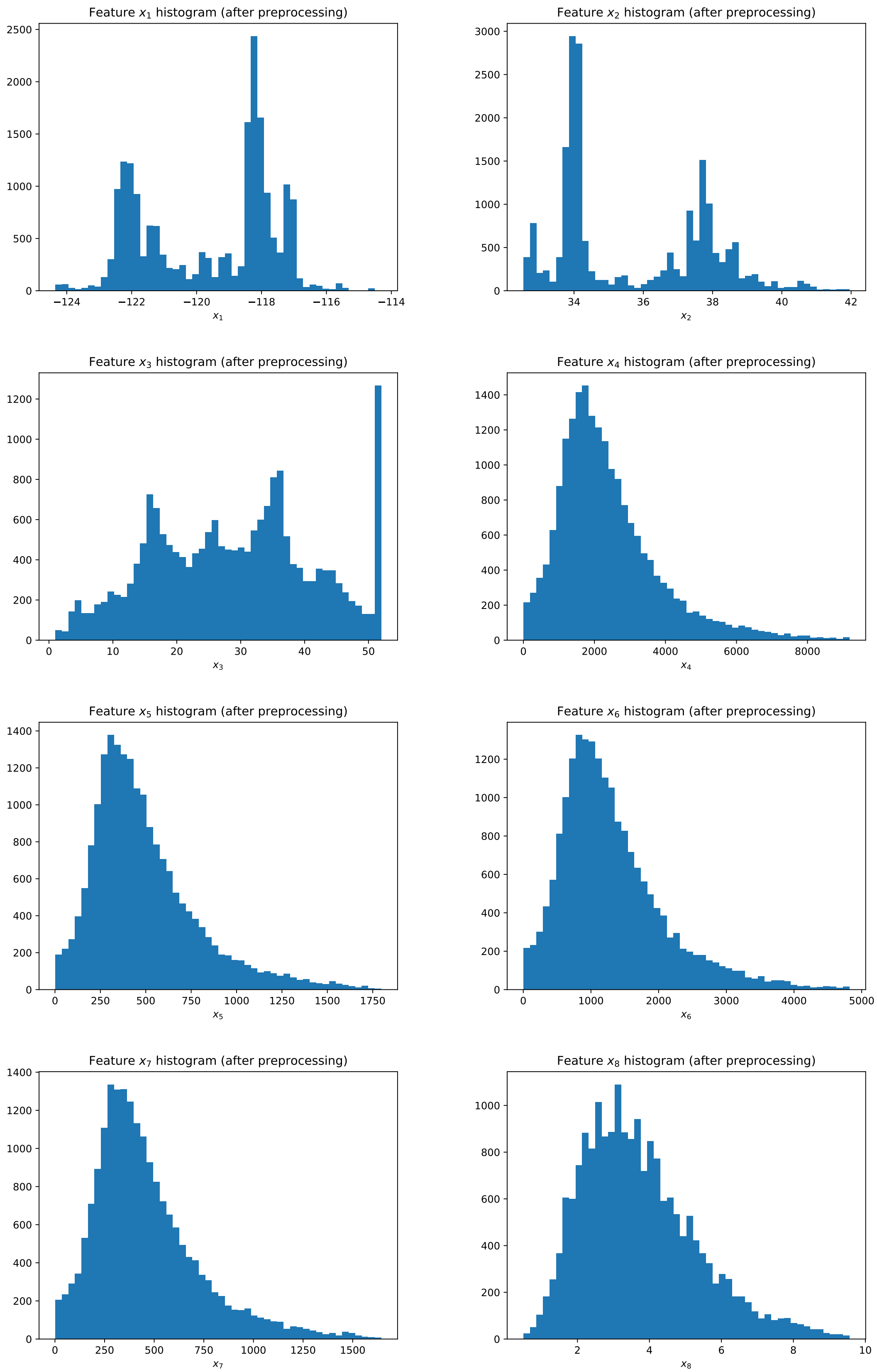

Figure 1: The Histogram of each feature in the California Housing Dataset.

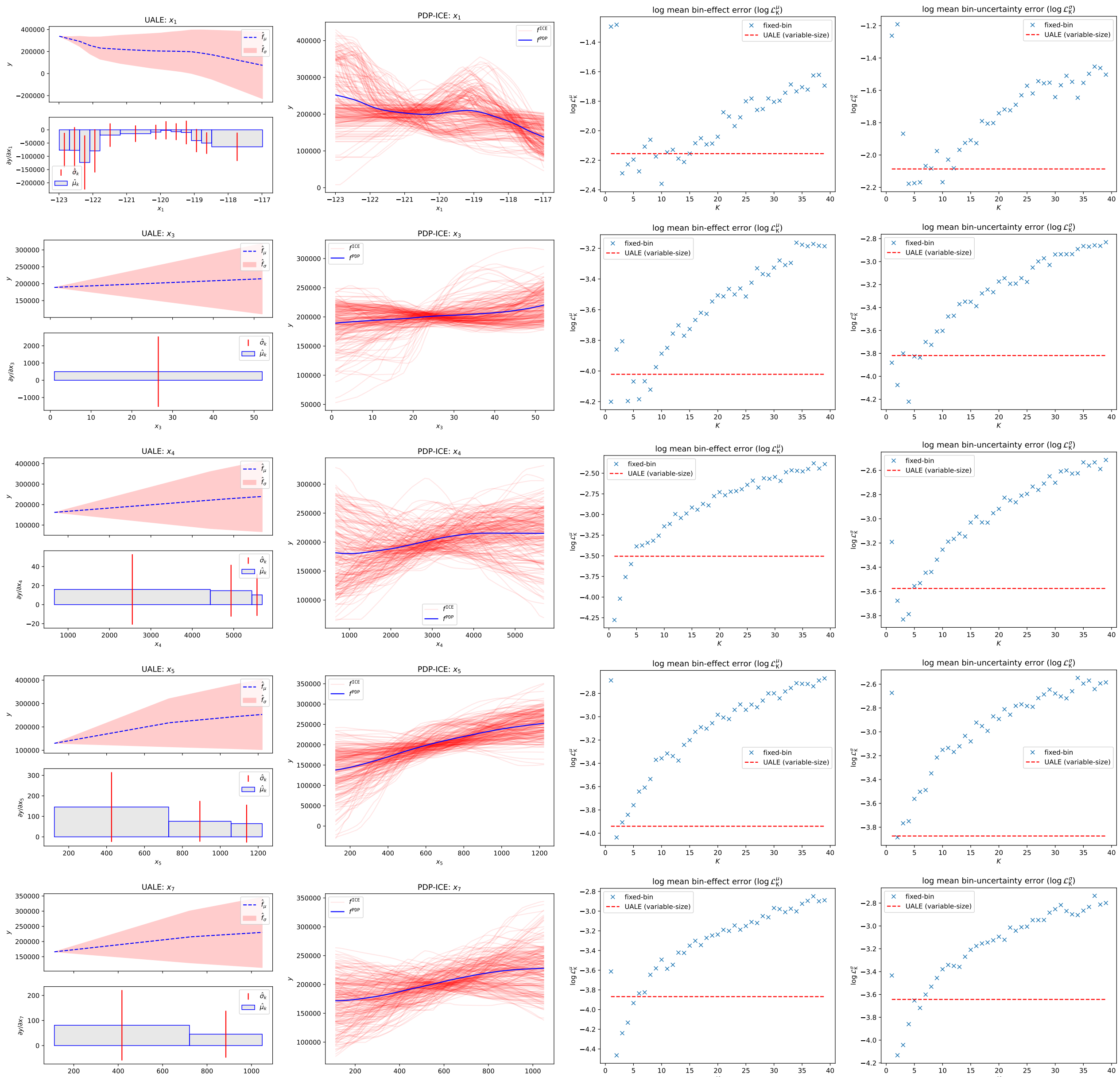

Figure 2: From left to right: (a) UALE plot, (b) PDP-ICE plot, (c) UALE vs fixed-size $\mathcal{L}^{\mu}$ and (d) UALE vs fixed-size $\mathcal{L}^{\sigma}$. From top to bottom, features $x_1, x_3, x_4, x_5, x_7, x_8$.