.dit/*
Informatics & Telematics

RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

Vasilis Gkolemis [1, 2] Theodore Dalamagas [1] Eirini Ntoutsi [3] Christos Diou[2]

[1]ATHENA Research Center [2]Harokopio University of Athens [3]Bundeswehr University of Munich

ATHENA Research & Innovation Information Technologies

TL;DR

RHALE: Robust and **Heterogeneity-aware** ALE
- Robust: auto-bin splitting
- Heterogeneity: $\pm$ from the average

keywords: eXplainable AI, ALE, Heterogeneity, Feature Effect

Motivation - ALE limitations

ALE (Apley et. al) is a SotA feature efect method but it has two limitations:
- it does not quantify the heterogeneity, i.e., deviation of the instance-level effects from the main (average) effect
- it is vulnerable to poor approximations, due to the bin-splitting step

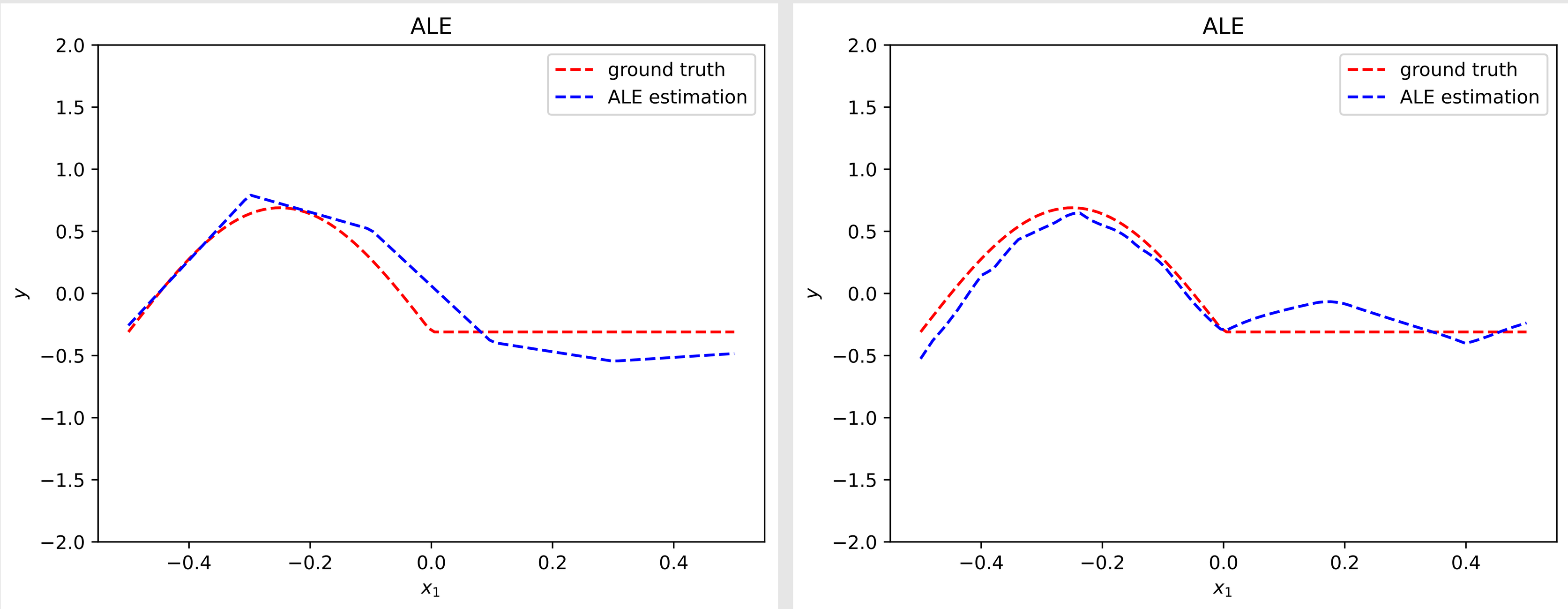

Figure 1. Left: ALE approximation with a narrow bin-splitting (5 bins). Right: ALE approximation with a dense bin-splitting (50 bins)

- Both approximations are bad:
  - Narrow bin-splitting hides fine-grain details
  - Dense bin-splitting is noisy (low samples per bin rate)
- No information regarding the heterogeneity

Simple approach: ALE + Heterogeneity

ALE main effect definition

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

ALE main effect approximation

$$\hat{f}^{\mathrm{ALE}}(x_s) = \Delta x \sum_{k}^{k_x} \underbrace{\frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[\frac{\partial f}{\partial x_s}(x_s^i, \boldsymbol{x_c^i})\right]}_{\text{bin effect}:\hat{\mu}(z)}$$

ALE heterogeneity definition

$$\sigma(x_s) = \sqrt{\int_{x_{s,\min}}^{x_s} \underbrace{\mathbb{E}_{X_c|X_s=z}\left[\left(f^s(z, X_c) - \mu(z)\right)^2\right]}_{\sigma^2(z)} \partial z}$$

ALE heterogeneity approximation

$$\mathrm{STD}(x_s) = \sqrt{\sum_{k=1}^{k_x} (z_k - z_{k-1})^2 \underbrace{\frac{1}{|\mathcal{S}_k| - 1} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left(f^s(\mathbf{x}^i) - \hat{\mu}(z_1, z_2)\right)^2}_{\sigma^2(z)}}$$

Simple but wrong: ALE + Heterogeneity

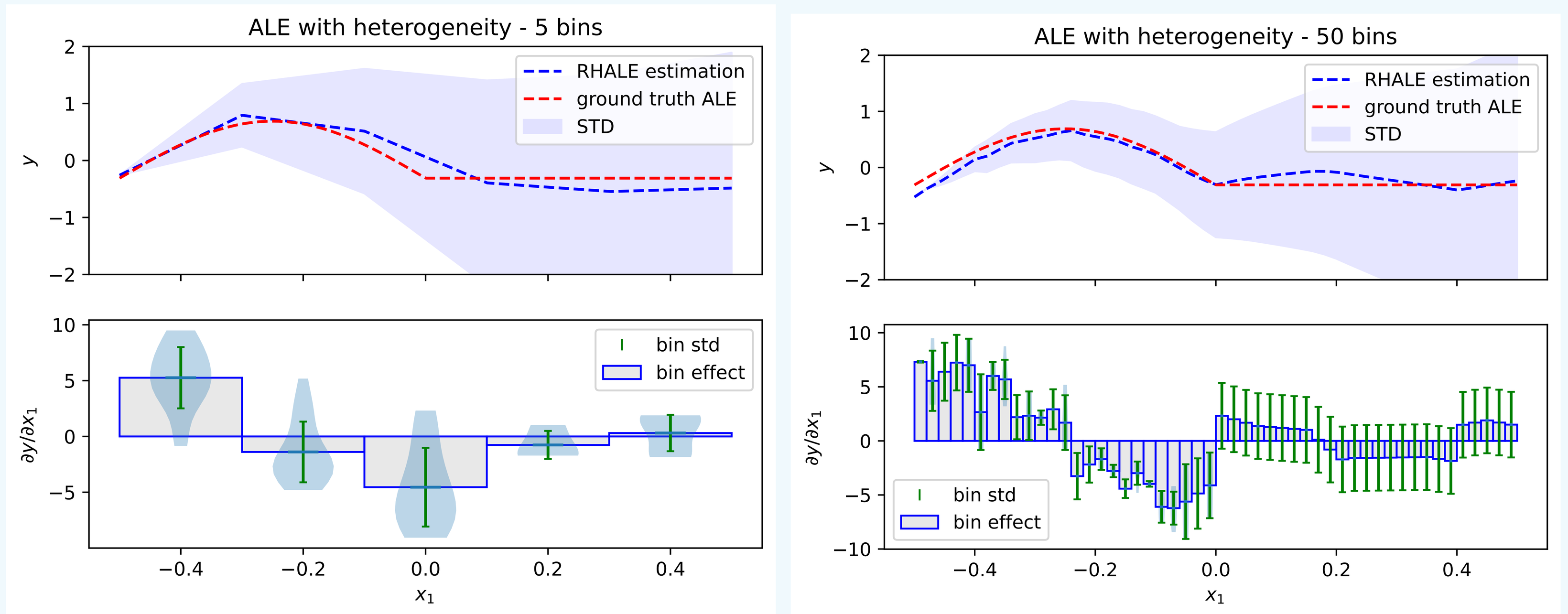

Figure 2. Left: approximation with narrow bin-splitting (5 bins) and (Right) with dense-bin splitting

1. Fixed-size bin splitting can ruin the estimation of the heterogeneity

RHALE: Robust and Heterogeneity-aware ALE

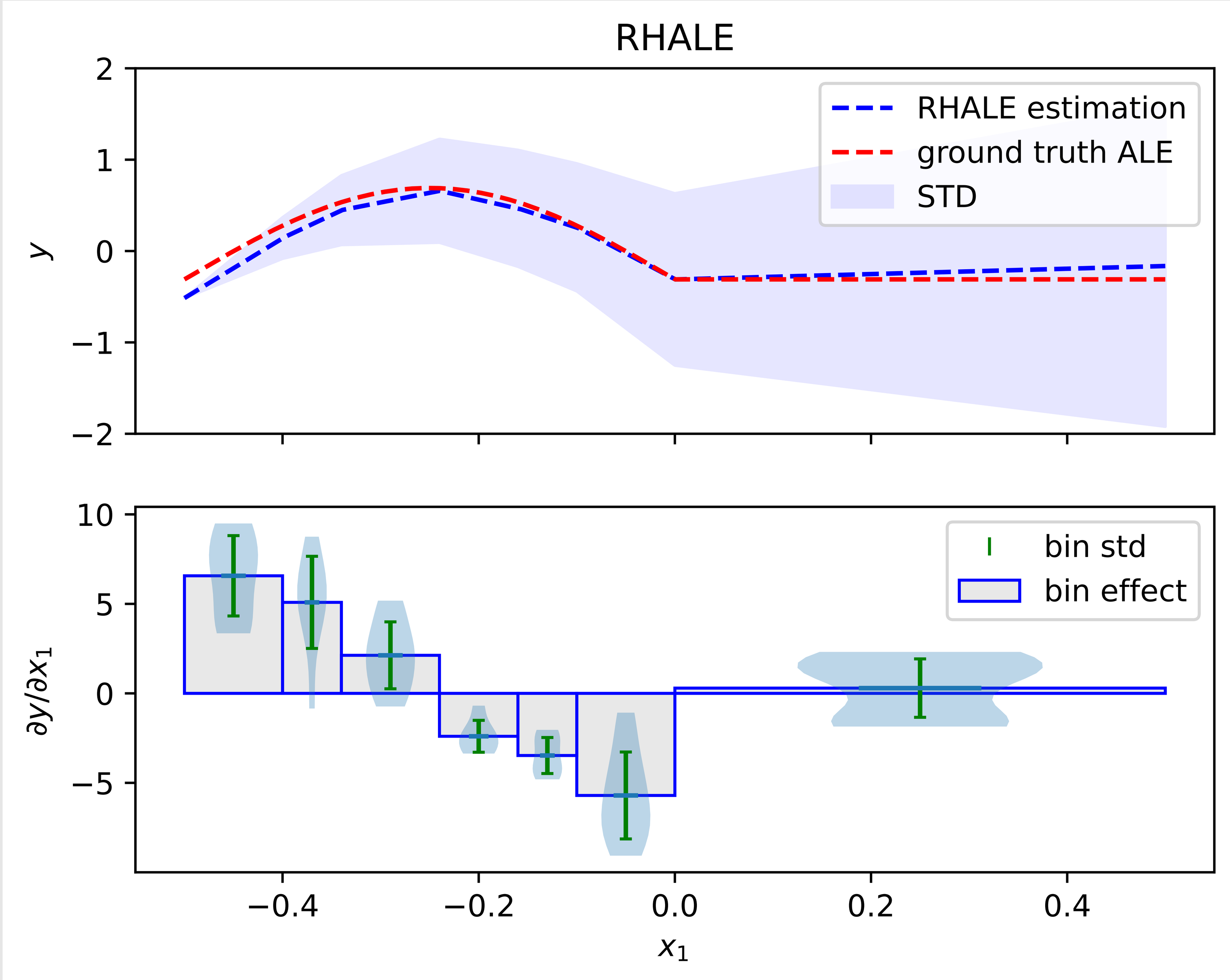

Figure 3. Narrow bins ($K = 40$) $\Rightarrow$ limited $\frac{\text{samples}}{\text{bin}}$ $\Rightarrow$ both plots are noisy

Simple but correct:
- Automatically finds the **optimal** bin-splitting
- Optimal $\Rightarrow$ best approximation of the average (ALE) effect
- Optimal $\Rightarrow$ best approximation of the heterogeneity

Optimal bin-splitting

In the paper, we **formally prove**:
1. the conditions under which a bin does not hide the finer details of the main effect
2. the conditions under which a bin is an unbiased approximator of the heterogeneity
3. that given (1) and (2), increasing bin size leads to a reduction in estimation variance

Based on the above, we formulate bin-splitting as an optimization problem and propose an efficient solution using dynamic programming.

Conclusion

In case you work with a differentiable model, as in Deep Learning, use RHALE to:
- quantify the heterogeneity of the ALE plot, i.e., the deviation of the instance-level effects from the average effect
- get a robust approximation of (a) the main ALE effect and (b) the heterogeneity, using automatic bin-splitting

References

- Paper repo: git@github.com:givasile/RHALE.git
- Personal site: givasile.github.io
- Twitter: twitter.com/givasile1