DALE: Differential Accumulated Local Effects for accurate and efficient global effect estimation

Vasilis Gkolemis [1, 2] Theodore Dalamagas [1] Christos Diou [2]

[1]ATHENA Research Center [2]Harokopio University of Athens

TL;DR

DALE is a better approximation to ALE, the SotA feature effect method. By better, we mean faster and more accurate.

keywords: eXplainable AI, global, model-agnostic, deep learning

Motivation

Feature effect methods are simple and intuitive; they isolate the impact of a single feature x_s in the output y. Inspecting the feature effect plot a non-expert can easily understand whether a feature has positive/negative on the target variable. The task is difficult; isolating the effect of a single variable is tricky when features are correlated and the black-box function has learned complex. ALE is the only method that manages. However, ALE estimation, i.e., the approximation of ALE from the set of has efficiency and accuracy that we address with DALE.

DALE vs ALE

ALE definition

$$f(x_s) = \int_{x_{s,min}}^{x_s} \mathbb{E}_{\boldsymbol{x_c}|z} [\underbrace{\frac{\partial f}{\partial x_s}(z, \boldsymbol{x_c})}_{\text{point effect}}] \partial z$$

ALE defines the effect at $x_s = z$ as the expected change (derivative) on the output over the conditional distribution $\boldsymbol{x_c}|z$ and the feature effect plot as the integration of the expected changes.

ALE approximation

$$f(x_s) = \sum_k^{k_x} \underbrace{\frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \underbrace{[f(z_k, \boldsymbol{x_c^i}) - f(z_{k-1}, \boldsymbol{x_c^i})]}_{\text{point effect}}}_{\text{bin effect}}$$

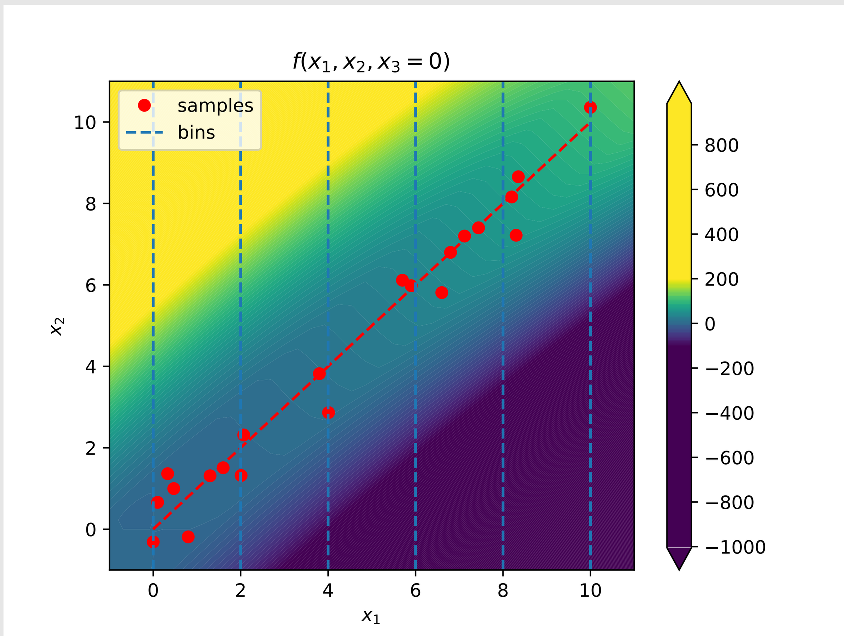
ALE approximation, i.e. estimating ALE from the training set $\mathcal{D}$, requires partitioning the s-th axis, i.e. $[x_{s,min}, x_{s,max}]$, in K equisized bins and computes the *local* point effects by evaluating the bin limits $[f(z_k, \boldsymbol{x_c^i}) - f(z_{k-1}, \boldsymbol{x_c^i})]$. This approach is slow and vulnerable to misestimatons. First, it is slow as the dimensionality of the dataset grows larger. Second, it demands predifing the bin limits. Finally, it may create out-of-distribution samples when bin size becomes large.

DALE approximation

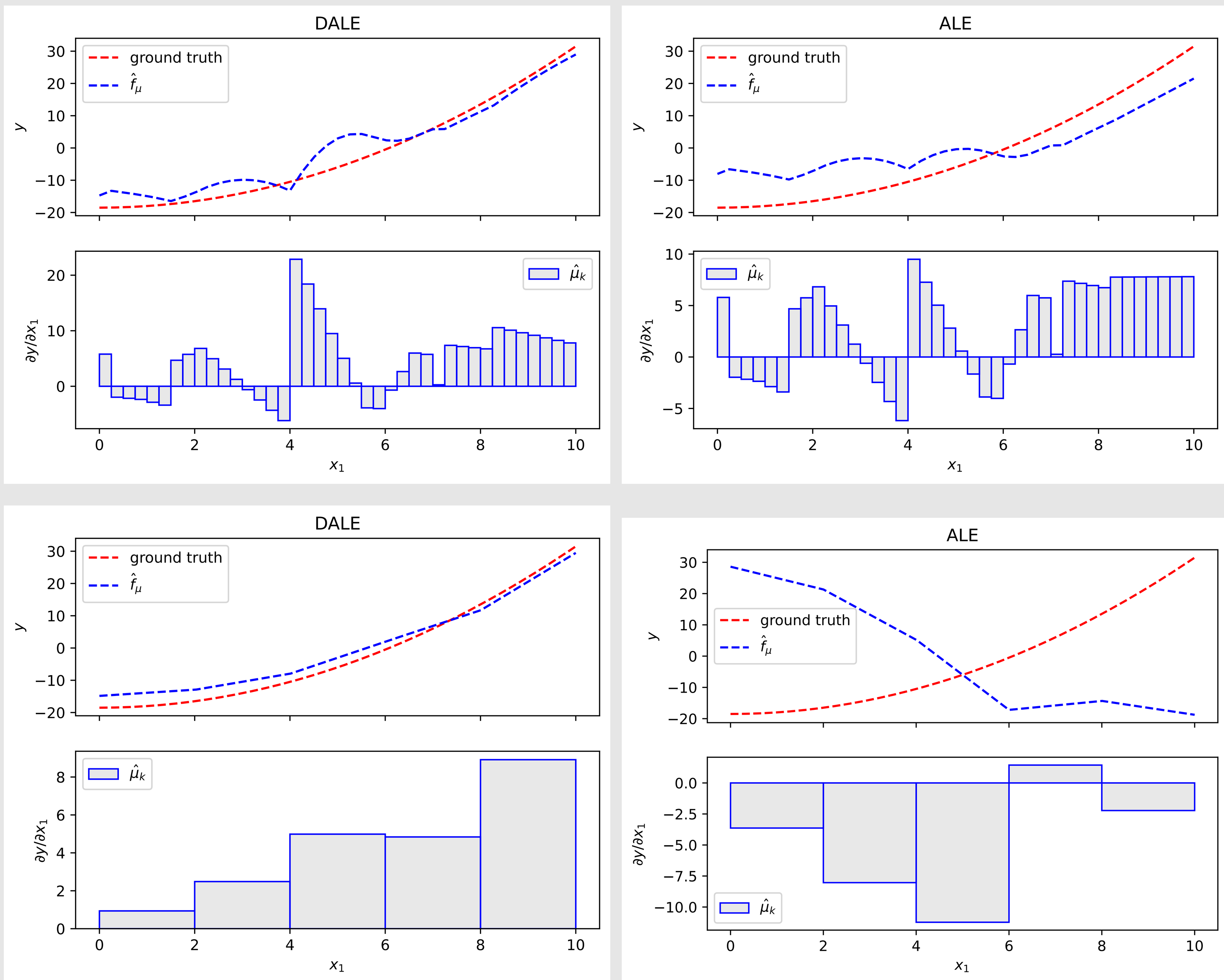
$$f(x_s) = \Delta x \sum_k^{k_x} \underbrace{\frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \underbrace{[\frac{\partial f}{\partial x_s}(x_s^i, \boldsymbol{x_c^i})]}_{\textbf{point effect}}}_{\text{bin effect}}$$

DALE approximation addresses these issues. The main difference is the use of automatic differentiation, instead of evaluating at the bin limits.

DALE saves from OOD



Consider the following case; (a) we have limited samples (b) high variance and (c) the black-box function changes abruptply outside the data manifold. For example, $f(x_1, x_2, x_3) = x_1 x_2 + x_1 x_3 \pm g(x)$, with $x_1 \in [0, 10]$, $x_2 = x_1 + \epsilon$ and $x_3 \sim \mathcal{N}(0, \sigma^2)$. The term $x_1 x_3$ makes estimations from limited samples noisy, so we need to grow the bins larger (more $\frac{\text{points}}{\text{bin}}$). But as we grow bins, ALE creates OOD samples,



DALE is much more efficient

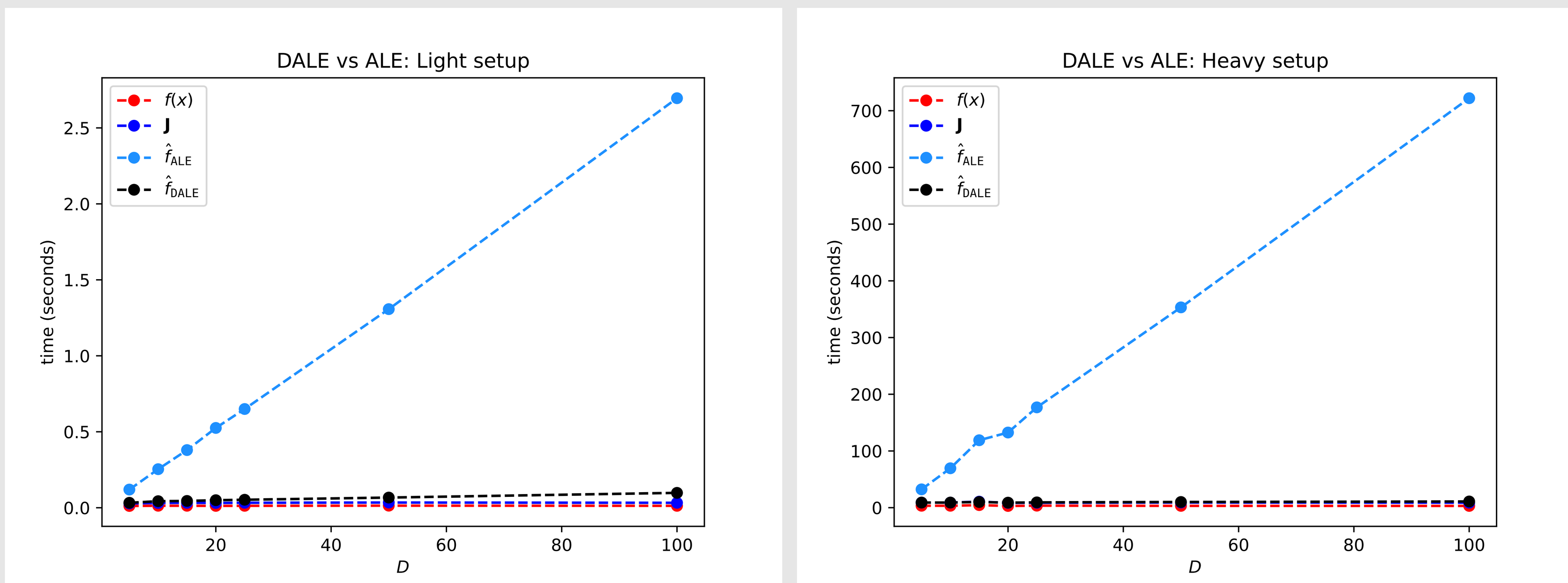


Figure 1. Light setup; small dataset ($N = 10^2$ instances), light f. Heavy setup; big dataset ($N = 10^5$ instances), heavy f

A block title

This poster was made by modifying the Gemini Beamer Poster Theme **Athalye2018** and the Beamer `seagull` Color Theme. Some block contents, followed by a diagram, followed by a dummy paragraph. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi ultricies eget libero ac ullamcorper. Integer et euismod ante. Aenean vestibulum lobortis augue, ut lobortis turpis rhoncus sed. Proin feugiat nibh a lacinia dignissim. Proin scelerisque, risus eget tempor fermentum, ex turpis condimentum urna, quis malesuada sapien arcu eu purus.

References

 vgkolemis@athenarc.gr