We want to thank all reviewers for carefully reviewing our paper and for their insightful remarks. We will make sure to address them all for the camera-ready version of the paper, if it is accepted for publication. Below, we discuss selected issues raised by the reviewers.

Reviewer 1 - Issue 1. I was hopping to see some more discussion on why the Monte-Carlo approximation fails.

Thank you for the remark. This point was not explicitly specified in the Introduction, where we first mention the *Monte-Carlo approximation failure*. Instead, we discuss it further in the first paragraph of Section 4.3, while in Section 5.1.2, we demonstrate it through a synthetic example. For the camera-ready version, we will enhance Section 4.3 with a more precise description of the reasons behind Monte-Carlo's failure and provide the relevant references in the Introduction.

Reviewer 3 - Issue 1. DALE requires the differentiability of the black-box model. However, DALE is suitable for tabular data where features have a semantic meaning. As state of the art black-box methods are (ensemble) tree-based models, that are unfortunately non-differentiable, the impact of DALE is therefore very limited.

Thank you for the remark. DALE indeed requires the differentiability of the black-box model, which, unfortunately, does not generally hold for tree-based methods. There are, though, many differentiable models with near SotA results in tabular data, e.g. TabNet. Furthermore, there is lately a growing interest in methods that use tree-based ideas in a differentiable setting, like the Tree Ensemble Layer. For these reasons, we believe that DALE can be a useful choice in many scenarios.

Reviewer 3 - Issue 2. There is no evidence that DALE improves ALE's approximation in real world datasets.

It is indeed difficult to provide definite evidence for DALE improvement in terms of accuracy in this case, because in real-world datasets we lack the ground truth effect. Having a definite comparison between the two methods was the actual reason for designing the synthetic example of 5.1.2. Therefore, for the Bike-Sharing dataset, we limit ourselves to an example where we show that the computed effect through DALE is accurate independently of the bin size, whereas ALE's estimation deteriorates when altering the bin size.

Reviewer 4 - Issue 1. This paper introduces two version of DALE, i.e. the first-order DALE and the second-order DALE. What is the main differences on the performance? which one is better? Can you add some discussion?

Thank you for the remark. The second-order effect is naturally slower than the first-order in both DALE and ALE (it requires the computation of derivatives wrt to two variables and the creation of a grid of bins). However, it is more meaningful to compare them in pairs; first-order DALE vs. ALE and second-order DALE vs. ALE. In both cases, DALE is faster by a factor of D compared to ALE's approximation, where D is the dimensionality of the dataset. The speed-up is due to the auto-differentiation step, as discussed in Section 4.2, and not any consequent step. This point is confirmed in the synthetic example (Section 5.1.1, Figures 2, 3) and the Real Dataset (Section 5.2, Table 2). We will update the manuscript by adding a discussion at the end of Section 4.2 to clarify that the same computational benefit (factor of D) also holds for the second-order case, since it is a direct consequence of auto-differentiation.

Reviewer 4 - Issue 2. The experiment part can be enhanced. The proposed DALE is only evalued on simple synthetic and real dataset. More diverse experiments are expected. Are there any results support that DALE has better scalability to high-dimensional data than ALE? How does the proposed DALE compared with other feature effect methods, e.g. PDPlots, and MPlots?

Regarding the scalability part, we tried to cover it experimentally through the synthetic example (Section 5.1.1) and the Real Dataset (Section 5.2). In the synthetic example, we evaluate the runtime after altering: (a) the dimensionality of the dataset D - Figure 3, the two first images - and (b) the model size L (i.e., how long it takes to evaluate the model) - Figure 3, the last two images. We also mention that the dataset's number of samples N does not play a significant role in the runtime. Therefore, as we mention in the last sentences of 5.1.1, the computational benefit of DALE generalizes to high-dimensional data as long as there is enough memory for the dataset/model to be stored. The only reason we limited ourselves to cases up to $D = 100$ is that tabular data typically do not exceed this dimensionality. Following the remark, we will clarify that the same conclusion holds for even larger dimensionality.

Regarding the comparison between DALE (and ALE) with the rest of the feature effect methods (PDPlots, MPlots), we quickly discuss some differences in Section 3 (Background), where we introduce the reader to the advantages of ALE compared to PDPlots and MPlots. The original ALE paper thoroughly discusses the differences between the three methods and the advantages of ALE in terms of accuracy and efficiency. In our paper, we mainly focus on the advantages of DALE compared to ALE's approximation; therefore, we do not explicitly compare DALE with PDPlots and MPlots. However, realizing that this is not evident in the manuscript, we will add an explanation stating that DALE is the more efficient approach since PDPlots and MPlots are even slower than ALE's approximation, which is confirmed by the original ALE paper.