We thank the reviewers for their valuable feedback that will significantly improve our paper. We will address all the feedback such as notations, wording, additional plots, and the missing references in the final version. For the major comments, we organize our responses as follows.

Reviewer 1: Limitations of the results? what if A1 and A2 are not satisfied? While the uniform marginal distribution assumption in A1 is relatively easy to satisfy (by transforming the features via its inverse CDF), we agree that the independence assumption in A1 is quite strong. Correlated features are commonly encountered in practice and difficult for any feature selection method. This is indeed a limitation of Theorem 1. We will point this out in our revision. On the other hand, although we assume noisy features to be independent, Theorem 1 allows relevant features to be correlated. The CHIP data included in our simulation studies shows that MDI-oob works in this setting. We would like to deal with more general feature correlations in our future work.

Reviewer 1: A controlled experiment where $G_0(T)$ can be computed exactly to empirically appreciate the tightness of the bound.

Our theorem states that for fixed dimension p, disregarding the $\log n$ terms, the bound is approximately inversely proportional to the minimum node size m_n. In Figure 1 of our submission, we plot the MDI importance of each feature as a function of m_n. If we plot the MDI importance against $1/m_n$ as in Fig. 1, then we observe that the MDI of noisy features is close to a linear function w.r.t. $1/m_n$, which verifies that the $1/m_n$ rate is tight. We plan to add this plot in our supplementary material. The constants in the proof are not tight and can be improved with a more careful analysis.

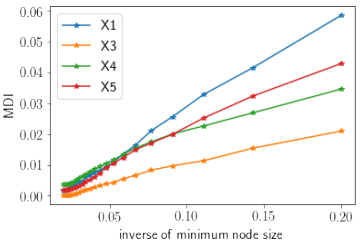

Fig. 1. MDI for noisy features

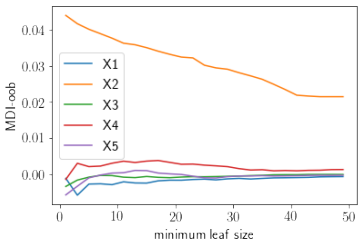

Fig. 2. MDI-oob for the first sim..

Reviewer 2: Link between Part 2 (MDI-oob) and Part 1 (Theoretical analysis of the bias) The link between 2) and 1) is more from a practical point of view: 1) points out the deep tree regime where MDI has the biggest problem, and 2) offers an empirical solution to alleviate the issue. Empirical evidence on how MDI-oob reduces the bias can be seen in Fig. 2.

Reviewers 2 and 3: Give theoretical/empirical evidence that MDI-oob can "debias" MDI. Unfortunately, we do not have theories for MDI-oob yet. Empirically, we compute the MDI-oob for the first simulation. The result is shown in Fig. 2, which shows that MDI-oob indeed reduces the bias of MDI. We will add Fig. 2 in our future manuscript.

Reviewer 2: Why MDA performs badly given the fact that it also uses OOB samples? We think this could be due to the low signal-to-noise ratio in our simulation setting. After we train the RF model on the training set, we evaluated the model's accuracy on a test set. It turns out that the accuracy of the model is quite low. In that case, MDA measure struggles because the accuracy difference between permuting a relevant feature and permuting a noisy feature is small. If we increase the signal-to-noise ratio, the MDA gets better.

Reviewers 2 and 3: Why not directly computing the impurity and the feature importance using the OOB samples? Directly computing the impurity using OOB samples may indeed lower the bias. We will add this point in our revision. However, unless the responses of all the OOB samples falling into a node are constant, the impurity decrease at that node is still always positive. In this case, an argument similar to the proof of Theorem 1 can show that the bias of directly computing impurity using OOB samples could still be large for deep trees.

Reviewer 3: "G0(T) is typically non-negligible in real data", how could one ever know this? it requires knowing the true distribution, which we don't know for any real data. Our statement requires knowing which features are noisy in a given prediction problem. While we agree that this is generally difficult, based on our experience, there are many applications where negative controls are measured, particularly in the biological sciences. In those problems, we do know that such noisy features exist and thus $G_0(T)$ is often non-negligible.

Reviewer 3: Comment on SHAP. SHAP originates from game theory and offers a novel perspective when we analyze the existing methods. While it is desirable to have 'consistency, missingness and local accuracy', our analysis indicates that there are other theoretical properties that are also worth taking into account. As shown in our simulation, the feature selection bias of SHAP increases with the depth of the tree, and we believe SHAP can also use OOB samples to improve feature selection performance.

All reviewers: Related work on using a validation set to compute the MDI.

We will provide a review of related literature in our future manuscript.