We thank each reviewer for their insightful and constructive feedback. It will surely improve the manuscript.

Experiments. Reviewers #2 and #4 suggested that we illustrate our theory with experiments. We wholeheartedly agree. Following this suggestion, our paper now includes experiments with various synthetic/real-world datasets and compares CART with k-NN and other kernel methods. In Fig. 1, we show the outcome of one such experiment. We sample $\{(\mathbf{X}_i, Y_i)\}_{1\leq i\leq n}$ i.i.d. with $n = 1000$ and consider a sparse additive model $Y = \sum_{j=1}^{d_0} g_j(X_j)$ with $d_0 = 5$ component functions, where each $g_j(X_j)$ equals $\pm X_j^2$ (alternating signs) and $\mathbf{X} \sim \text{Uniform}([0,1]^d)$. We then plot the test error of CART vs. k-NN as d ranges from 5 to 100. According to Theorems 3 and 4, the convergence rate of CART depends primarily on the sparsity d_0 and therefore its performance should not be adversely affected by growing d. Consistent with our theory, the prediction error of CART remains stable as d increases, whereas k-NN does not adapt to the sparsity.

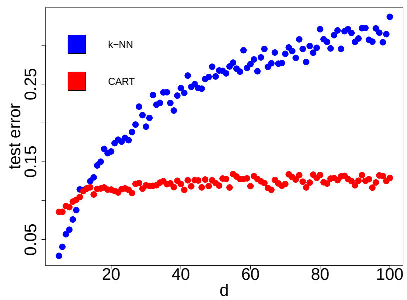

Figure 1: Prediction error (averaged over 10 independent replications) of pruned CART vs. k-NN (with cross-validated k) as a function of ambient dimension $d \in \{d_0, \dots, 100\}$ with fixed sparsity $d_0 = 5$. CART is impervious to increasing ambient dimensionality, whereas k-NN suffers and does not adapt to sparsity.

Distributional assumptions. Reviewer #2 inquired about the consequences of assuming the input variable $\mathbf{X} = (X_1, \dots, X_d)$ is uniformly distributed on $[0,1]^d$. For independent predictor variables X_j, there is no loss of generality in assuming uniform marginal distributions. Indeed, CART trees are invariant to strictly monotone transformations of the individual predictor variables. One such transformation is the marginal cumulative distribution function $F_{X_j}(\cdot)$ of the predictor variables, for which $F_{X_j}(X_j) \sim \text{Uniform}([0,1])$—and so the problem can be reduced to the uniform case. For dependent predictor variables, the proofs go through if the joint density of $\mathbf{X}$ is bounded above and below by a positive constant, though the convergence rates have worse dependence on the sparsity level.

Connection to ensemble models. Reviewer #1 would like to see a more in-depth discussion of how the analysis for CART trees can be carried over to random forests—one that goes beyond the ensemble principle. This is an excellent point and we agree that it deserves more attention. Therefore, we have included a new section that explicitly describes how our results can be used to show analogous adaptivity properties for random forests. Let us briefly mention that, while the efficacy of randomization in random forests (e.g., bagging or random feature selection) is not fully understood from a theoretical perspective, basic properties such as consistency often begin with a study of the individual trees [Scornet et al., 2015, Wager & Athey, 2017]. Finally, it is indeed true that the empirical soundness of CART has been well-documented over the past 30+ years—however—many of its theoretical properties (like adaptivity to sparsity) have not been made rigorous. Since CART is still widely used for its simplicity and interpretability, the primary goal of this paper is to put these empirical observations on a solid theoretical foundation.

Proxy for estimation and training error. Reviewer #4 asked for further clarification on how we avoid using the node diameters as a proxy for the approximation error and, instead, directly bound the training error. The extant approach typically bounds the expected prediction error (over the training data) by analyzing the approximation and estimation errors separately (the estimation error is usually less troublesome). Because CART outputs the average of the response values in a node, if the regression function is smooth, the approximation error is at most a constant multiple of the largest node diameter. Thus, one can use the node diameters as a proxy for the approximation error. In contrast, we use the fact that the prediction error is with high-probability (over the training data) bounded by the training error plus a complexity term. The connection between the Pearson correlation and the training error (see Lemma 1) facilitates our analysis and allows us to prove more fine-grained results.

Quantity that governs the training and prediction errors. Reviewer #4 asked for clarification on whether the quantity in Assumption 1 governs the convergence rate of both the training error and prediction error. Assumption 1 is merely a technical condition about the maximum number of data points in each node and enables one to use the correlation comparison inequality in Fact 2—it does not explicitly control the training and prediction errors. As mentioned in the discussion preceding Theorem 4, it is $\widehat{\rho}_{\mathcal{M}}$ (i.e., the largest correlation between the response data and a monotone function of an individual predictor variable) that explicitly controls the rate at which both errors tend to zero.

Additional reference. Reviewer #2 mentioned the reference [Nobel, 2002]. We thank the reviewer for pointing this out to us and have cited it in a revised version of the paper.

Accessibility to broader audience. Reviewer #3 expressed concern that the paper may seem difficult for readers without expertise in the subject area to follow along with the mathematical notation. We have carefully gone through the paper to explain and write all notation in a more accessible way that respects the reader's background. We have also carried out a thorough proof-reading to correct any typos or references that were not properly compiled (as mentioned by Reviewers #1 and #2).