We thank all reviewers for the valuable comments! It appears that all reviewers saw the algorithmic and theoretical contributions of our work. Below we provide responses to the questions and concerns of the reviewers.

(R1) *Comparisons with more recent works on tree-based algorithms.* Thanks for your suggestions and the references! *Qualitative comparisons:* As we mentioned in line 69, [1] dealt with classification (not regression) by space partitioning and locally linear classifiers. [2] focused on the ensemble algorithms for *online regression*, which also built on trees with linear models in leaves, but using existing tree construction algorithms. [3] introduced two tree-based methods for *adaptive learning*, where DFT used fixed hard splits (not trained) and DAT used adaptive soft splits trained by stochastic gradient descent (SGD). In comparison, we use adaptive hard splits trained by looking into the rank correlation between the residuals and covariates, instead of directly minimizing the empirical error by SGD. [3] focused on the weighting in averaging all piecewise functions defined by the subtrees, while we focused on the split selections in tree constructions. *Experiments comparisons:* It turns out that [2] shared dataset 'Abalone' and [3] shared dataset 'Kinematics' with us. Table 3 in [2] reported that the ORF was the best with MSE $= 5.68$ by CV, while our Table 1 shows $\text{SLRT}_{\text{LASSO}}$ had lower MSE $= 4.562(2.136^2)$ and the ensemble version RF_{SLRT} had even smaller MSE $= 4.439(2.107^2)$. [3] reported DAT was the best with time accumulated error 0.0639 on the normalized data. To be comparable, we re-normalized 'Kinematics' by $\frac{y-(\max(y)+\min(y)/2}{(\max(y)-\min(y))/2}$. The MSEs by the 10-fold CV were 0.0326(SLRT_{LS}) and 0.0327($\text{SLRT}_{\text{LASSO}}$), respectively, which were significantly lower than DAT. Nevertheless, we admit the results were not entirely comparable since [3] calculated the time accumulated errors in the adaptive learning context while we used cross validation. On the two common datasets of both articles, SLRT shows promising performance despite our motivation is not to minimize the empirical error directly but to maximize the rank correlation between the residuals and covariates in splits selection. We've emailed the authors of [2] and [3] for experiment details and codes, and will make fairer comparisons in the final.

(R1 & R2) *Theoretical contributions and assumptions.* Theoretical justifications in Sec 2.2 show the selected splits converge in probability to the true splits, which motivated Alg1 (uncorrelated regressors) and Alg2 (correlated regressors). Although the convergence is in the asymptotic sense, it does provide mathematical assurance for consistent split selections, which is the force behind the promising performance in the simulations in Sec 5. Regarding assumptions in the supplement, ASMP 2.2-2.4 are quite general with ASMP 2.3 satisfied for symmetrically distributed X that can be realized by Box-Cox transformation, and ASMP 2.4 for identification purpose. We agree ASMP 2.1 on the correlated regressors is strong, which is the reason for providing theoretical results and Alg2 for the correlated regressors.

(R1 & R2) *Clarity issues.* We'll bring in the theoretical results in the main writeup, and add more explanations for the analysis in Sec 2.2 with illustrations by the example in Sec 5.1. We'll update the supplement to make it issues-free.

(R1 & R4) *Piecewise linearity concerns/ Generalization of models at the leaves.* To accommodate non-linearity at the leaves, higher order polynomial regressors can be added. This extension is essentially covered by the existing Alg2 for correlated regressors, with consistency of splits selection ensured by Th 2.2. We'll add the discussion in the final paper.

(R2) *How the algorithm optimizes P.* Sorry for not presenting clearly. The transformation matrix $P_{(j,a)}$ in line 137 is directly calculated, without optimization. Let S_X^L and S_X^R be the sample covariance matrices conditional on $X_j \le a$ and $X_j > a$, respectively. Then, $P_{(j,a)}$ satisfies: $P_{(j,a)} S_X^L P_{(j,a)}^T$ and $P_{(j,a)} S_X^R P_{(j,a)}^T$ are both diagonal matrices, which can be obtained by the spectral decomposition of positive definite matrices. We will add more details in the final paper.

(R2 & R3) *Experiment settings.*The ensemble size in the experiments was 50. We conduct the linear transformation on X by multiplying $(S_X^{1/2})^{-1}$ to remove the correlation, where S_X is the sample covariance matrix. We'll add the computational complexity analysis, RF and WRF ensembles with GUIDE and MARS as the base methods in the final version. Information about responses and covariates on the nine datasets were introduced in Sec 5 of the supplement.

(R2) *Distance between pairs.* The distance in line 128 is the Euclidean norm, where index j is regarded as an integer.

(R3) *How to determine the regularization parameter λ in LASSO?* By the cross validation over each node.

(R4) *A Theoretical discussion of the case when the data does not conform to the SLR setting.* In such cases, the induced tree would be large in size as the model complexity is adaptive to data, and the estimation tends to be a nonparametric first-order approximation. The theoretical discussion would be similar to the results in [4]. So, when the data conform to the SLR setting (approximately), SLRT provides interpretable results with a concise tree structure; otherwise, SLRT also provides fine approximations using elaborate partitions by a large tree. We'll add the discussion in the final paper.

(R4) *Detailed comments. Line 38:* "prevailing" means attaining the maximum of $\bar{C}(j,a)$. *Line 8 of Alg1:* Equations (2) and (4) were right. Sorry for the typo in Alg1. *How to determine whether to use Alg1 or Alg2?* This can be judged by testing on the covariance of X being diagonal or not using the sphericity test. *Line 193:* "subset or equal to".

Minor Issues. All minor issues raised by the reviewers will be rectified in full. Thanks for all constructive suggestions!

References

[1] J. Wang and V. Saligrama, "Local supervised learning through space partitioning," in *NeurIPS*, 2012, pp. 91–99.

[2] E. Ikonomovska *et al.*, "Online tree-based ensembles and option trees ... on evolving data streams," *Neurocomputing*, 2015.

[3] N. D. Vanli *et al.*, "A comprehensive approach ... regression based on trees," *IEEE Transactions On Signal Processing*, 2014.

[4] P. Chaudhuri *et al.*, "Piecewise-polynomial regression trees," *Statistica Sinica*, 1994.