Thank you reviewers for your positive and helpful suggestions!

R1: Figure 1 no OSDT (warm-start only): They provably have the same solution; the two points would be identical. Runtime benchmark against BinOCT and CART: Good idea. We can add BinOCT and CART execution profiles to figures like that shown in the paper's Figure 2. (We include examples in Figure 1 below.) In general, OSDT converges and *completes more quickly* than BinOCT; BinOCT *sometimes includes redundant leaves* leading to non-optimal solutions to our problem. CART frequently completes very quickly, but does not always yield results near optimal. For example, on monk-1, monk-2 and tic-tac-toe, the results of CART are far away from the optimal ones.
Show BinOCT and CART trees as well as OSDT trees (Fig 4): Sure! Figures 5 and 6 in the supplementary materials have example binOCT trees; we can add CART trees as well.
Revised title: We've made it efficient for real-valued features since the submission, owing to a good implementation of Theorem F.1. (see Figure 2 below). Introduce notation gradually: easy, will do. Including more background on CORELS: easy, will do. Thank you for all your suggestions! We appreciate it!

R2: Paper contributions: Sorry! (1) first practical optimal decision tree algorithm to achieve provably optimal solutions for nontrivial problems. (2) a series of new analytical bounds to reduce the space (Sec 3.2). (3) the first practical algorithmic use of a tree representation using only its leaves (Sec 3). (4) Implementation speedups saving 97% of run time (Sec 4 and appendix). (5) For the important COMPAS and FICO datasets (Sec 5), optimal trees have never previously been published. We present the first ones.
Utility for real and categorical features too: The COMPAS and FICO datasets used in the paper both have real-valued features, so it is applicable. Since the submission, our implementation of Thm F.1 is more efficient, allowing better scaling with real-valued features (e.g. on the Iris dataset in Figure 2, we reduced the number of tree evaluations by 45%). All possible split points are considered for each real-valued variable in our current implementation.

Scalability: We have been working on a parallel implementation and expect to have scalability results from it in time for the camera ready submission. Parallelization is *much* less difficult than what we already did.

R3: The authors would need to demonstrate convincingly that improved training accuracy also translates into improved test set generalization: The basis for all of statistical learning theory is that training accuracy *and simplicity* provably lead to better test accuracy, and OSDT's objective incorporates both accuracy and sparsity. The 10-fold cross validation results in the supplementary materials (Section J) illustrate this, showing in-sample and out-of-sample accuracy.
The results are less conclusive than for training: Since OSDT found solutions that were as accurate as other methods in testing, but were more sparse, we interpret that as conclusive evidence. We will discuss this in the main body.

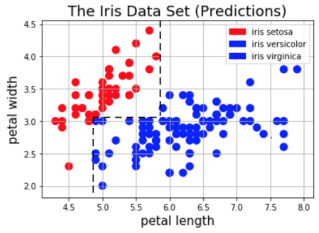

Figure 2: OSDT's result on Iris dataset. Covariates are real-valued. OSDT considered all possible splits to produce the optimal decision boundary (black).

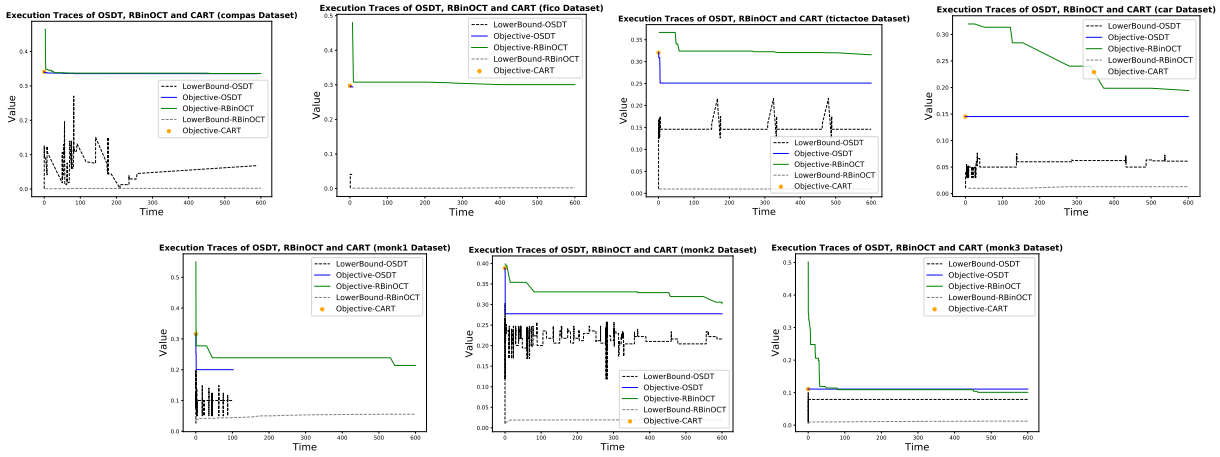

Figure 1: Execution traces of OSDT, CART and regularized BinOCT. OSDT converges much faster than RBinOCT generally. Traces end when execution completes for each algorithm separately. Runs were stopped after 10 min.