We thank the reviewers for their thoughtful and valuable feedback. We appreciate their time and effort, especially given the current uncertain times.

Response to Reviewer 1: We begin by responding to Reviewer #1's remark that the notion of estimating learnability is interesting but unsurprising:

"*The results are not surprising at all. That does not mean that it's easy to prove, but it is still not surprising that it is possible to estimate how well a learning algorithm will do on a set S by observing only a small part of the set S.*"

There are many other algorithms for learning decision trees, based on generic algorithmic paradigms such as polynomial regression [LMN93, KKMS08] and bottom-up construction [EH89, MR02]. We in fact believe that for these other approaches, it is impossible to estimate learnability with the sample complexity achieved in this work: exponentially smaller than the information-theoretic minimum required for learning. This highlights a unique advantage of the top-down algorithms that we study in this work: one can build a tiny part of the hypothesis corresponding to a specific input, without constructing the entire hypothesis.

Regarding the notion of estimating learnability more generally, although it is still relatively new, there is already a growing body of work (appearing at recent NeurIPS, COLT, and AISTATS conferences; see lines 41-48 of our submission for references), studying it for a variety of learning problems. These works highlight novel connections between this notion and other areas of interest in both the theory (sublinear time algorithms, property testing, etc.) and practice (data selection, hyperparameter tuning, etc.) of machine learning. Our work is the first is to study this notion in the context of decision tree learning.

"*Also, the paper makes strong monotonicity assumptions, but does not discuss the implications of it on the strength (and relevance to application) of the results.*"

We thank the reviewer for raising this point. The focus of our work is on formal performance guarantees, and such guarantees for top-down algorithms are only known for monotone target functions. There are simple examples of non-monotone target functions for which top-down algorithms fare very poorly in the sense of building a tree that is no more accurate than a trivial classifier (unless we allow them to grow a huge tree). Monotonicity is a natural way of excluding these adversarial functions, and for this reason it is one of the most common assumptions in learning theory. Results for monotone functions tend to be good proxies for the performance of learning algorithms on real-world datasets, which also do not exhibit these adversarial structures. Just as ID3 and CART do, we expect our algorithm will work well in practice for most real-world datasets, even if they are not perfectly monotone. We will revise our paper to discuss this.

Response to Reviewer 2: We thank Reviewer #2 for suggestions for improving our presentation. We agree with them, and will incorporate these suggestions in our next revision.

Response to Reviewer 4: Regarding Reviewer's #4 point about the distinction between our work and [BLT20]: that work focuses on proving that top-down heuristics successfully learn monotone functions, whereas our focus is different. We have access to an unlabeled dataset, and wish to estimate how well those top-down heuristics would perform on the labeled dataset by only labeling a few points. Our design and analysis of mini-batch top-down is in service of our main goal, which is to give an algorithm for the aforedescribed learnability estimation task.

We thank the reviewer for their question about overall complexity. The runtime of our algorithm can be upper bounded by the product of the size of the dataset and the sample complexity of our learnability procedure. In particular, taking a batch sample from a particular leaf can be done in a single sweep through the dataset to determine which inputs are consistent with the leaf and then randomly sampling one of them. We will revise our paper to incorporate the runtime.

References

[BLT20] Guy Blanc, Jane Lange, and Li-Yang Tan. Provable guarantees for decision tree induction: the agnostic setting. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[EH89] Andrzej Ehrenfeucht and David Haussler. Learning decision trees from random examples. *Information and Computation*, 82(3):231–246, 1989.

[KKMS08] Adam Kalai, Adam Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

[MR02] Dinesh Mehta and Vijay Raghavan. Decision tree approximations of boolean functions. *Theoretical Computer Science*, 270(1-2):609–623, 2002.