We want to thank reviewers for insightful comments, and we will improve the paper accordingly. In the following we focus on technical questions.

[To reviewer 1]

[Q1] ... the argument in the proof of Theorem 1 where the unit cube in Rˆd is treated as a discrete set by summing over its elements ... This kind of hand-waving would be more acceptable if there was a rigorous proof in the appendix.
[A1] We will clarify that Theorem 1 holds for discrete and continuous sets in Rˆd, while we use the summing over its elements in the proof of Theorem 1 for simplicity and readers' understanding. We will present the detailed proof in the appendix based on expectation and probability, rather than the summing over its elements, which is suitable for discrete and continuous distributions.

[Q2] The assumption on the distribution in Theorem 4 is in terms of the algorithm itself and remains quite obscure. Some elucidation of the assumption would be in order.
[A2] We will clarify that the assumption in Theorem 4 is relevant to algorithm, and it also holds for some irrelevant cases, for example, Algorithm 1 in our work satisfies such assumption when the data is separable and the separable hyperplane is parallel to axis.

We will improve this work according to your suggestions. Thank you.

[To Reviewer 2]

[Q1] No experiments, which could be helpful to verify the derived rates/identify if they are pessimistic.
[A1] We will clarify that this is a theory paper and the main arguments are supported by proofs that are rigorous and more reliable (experiments may be less reliable because many factors, such as parameter tuning and sampling, may influence the result). In a future longer version, we will consider the design of experiments.

We will improve this work according to your suggestions. Thank you.

[To Reviewer 3]

[Q1] There is no argument suggesting that those theorems could extend nicely to the multi-class case.
[A1] We will clarify that this work focuses on binary classification, and it is interesting to extend our work to multi-class learning, where the challenges lie in the theoretical analysis of predictions $f(x,y) - \max_{i\neq y} f(x,i)$ and Lipshcitz assumptions over multiple class-conditional distributions. Chen & Sun (JMLR 2016) and Ramaswamy et al. (ArXiv 2015) may shed some lights on this direction.

We will improve this work according to your suggestions. Thank you.

[To Reviewer 4]

[Q1] The studied algorithms remain quite far from real random forests (no bootstrap sampling, split choices are fully independent of the data, trees are pruned, etc.)
[A1] We will clarify that this work takes one step towards convergence rate of random forest for classification, and Algorithm 1 follows Breiman's random forests but with different splitting dimension and position. We will also clarify that it is still a long way to fully understand random forests and relevant mechanisms such as bootstrap sampling, data-dependence tree structure, tree pruning, etc. We leave those to future work.

[Q2] As in other results in the literature, convergence rates for forests are by-product of convergence rate of individual trees (using Lemma 1) ... This should be discussed in the paper I think.
[A2] We will clarify that the convergence rates of random forests are obtained from the expectation of convergence rates of individual trees based on Lemma 1, which can be viewed as the average of convergence rates of all of individual random trees.

[Q3] No real conclusion is drawn from the theoretical results that would help better understand standard RF or suggest modification to these methods.
[A3] We will clarify that our work is beneficial to understand the splitting mechanisms for random forests, such as the selections of splitting leaves, dimensions and positions, which may motivate new methods. For example, we get better convergence rates for pure random forests by using midpoint splits than random split (Thms 1 and 2); we also get better convergence rates by considering different selections of splitting leaves and dimensions in Algorithm 1 (Thms 3 and 4).

[Q4] One problem ... definition of the random variables Y_i and U_i between lines 254 and 261 ... They should thus be indexed by j in addition to i ... needed for the proof of Lemma 8 in Appendix C that considers a fixed dimension j.
[A4] We will add indexer i and j to random variables Y and U, and move relevant definitions before Lemma 8.

We will reorganize this work, add more discussions, and improve this work according to your suggestions. Thank you.