We thank all the reviewers for their comments. We first address questions raised by multiple reviewers and next respond to individual questions.

Running time and efficiency of RGB. To address the RGB runtime questions raised by multiple reviewers, we provide a very simple explanation. Given: a) a fixed sample of S coordinates (Algoritm 1, line 2.) b) the same subroutine is used for single tree node splitting and pruning; then the runtime of one RGB round is equal to the runtime of S rounds of GB. If $S = 1$, then the runtime of RGB is equal to that of GB. Given S parallel workers, the time per worker of RGB is equal to that of GB. The measured runtime of our experiments confirms the explanation above and we are happy to include the runtime plots in the paper. Since our algorithm within a single boosting round allows to sample and evaluate each candidate in parallel, the training is efficient given multiple workers.

Experimental benchmarking of RGB against other algorithms. As some reviewers suggested it is definitely possible to compare RGB with ensemble algorithms other than GB (e.g. reinforcement learning, random forest, etc). We consciously made a choice to provide comparison with GB only, and specifically for the families of regression trees, since it most clearly illustrates our novel methodology. This is because both these algorithms explore similar hypotheses spaces $\mathcal{H}$ and we would like to show that RGB by using generalization bound and randomization makes use of $\mathcal{H}$ better than GB. It would not be a fair or a meaningful experiment to compare RGB with ensembles of neural networks for example, since the underlying spaces $\mathcal{H}$ are not comparable. It is important to stress that we don't claim that RGB beats all other ensemble algorithms which can vary in hypotheses class complexity and search methods, we rather claim that given the same hypotheses space $\mathcal{H}$, RGB does better than a standard gradient boosting algorithm by making the use of generalization properties of $\mathcal{H}$ and random search over $\mathcal{H}$.

Experimental benchmarking of RGB for other datasets. We are happy to include a variety of experiments on larger datasets in the final version. We already confirmed that the results on UCI and MNIST datasets carry over to Higgs dataset. Overall, since this submission presents novel theory, methodology and algorithm, this is not a purely empirical paper and the UCI as well as MNIST dataset experiments that we presented serve as an illustration of the power of our approach.

REV1. "Why do you restrict the analysis to the hypothesis families of regression trees...?". First, regression trees is the most widely used hypothesis family in the boosting literature; second, doing similar generalization analysis for other families such as neural networks would involve the same steps, but a different and nontrivial proofs for the Rademacher complexity bounds, which is beyond the scope of this work.

REV2. "My most concern is the speed of RGB for it will search multiple trees in each iteration. And I do not think Line 271-273 is correct". Indeed, one tree fitting for GB can be done in a multi-threaded way, but not necessarily since for example in XGBoost does not support exact tree splitting method distributed. However, for our experiments whenever we use a multi-threaded subroutine for single tree splitting, we use it both in RGB and GB, making the runtime comparison consistent. Thus, it is correct that one round of RGB takes as much time as S rounds of GB, which is also explained above.

"(Significance) Not clear, due to the experiments are in small toy data (UCI data), no results on large datasets." We addressed this above. We would like to add that this work has meaningful contributions in theory and methodology, not just experiments - as nicely summarized by REV3. For example, we provide the missing theory for regularized boosting, that explains the impact of regularization on its generalization properties. Thus, the purpose of the experiments that we provide is to illustrate the power of the novel methodology, but not to win the state of the art across many tasks.

"More experiments on large datasets, not UCI...". We would like to bring to the Reviewer's attention that we have presented experiments on larger MNIST datasets in addition to the UCI datasets. As suggested, we are happy to include a variety of large scale experiments in the final version and we have confirmed that the MNIST results carry over to the Higgs dataset.

REV3. "...the authors' claim in lines 307-311...". We would like to explain to the Reviewer, that the claim we make is more specific. We claim that compared to an algorithm that operates on a hypotheses space $\mathcal{H}$ of the same complexity, RGB will reduce over-fitting by leveraging the theory that we developed. This is precisely captured by our experiment presented. However, if we compare different algorithms on widely varying hypotheses spaces, we would not be able to draw a conclusion. For example, we can't make a claim about RGB on regression trees with depth up to 10 versus ensembles of neural networks with 10 layers.