We thank the reviewers for their feedback. We answer the questions from the reviewers in the order as they listed in the reviewer comments. Due to the space limit, we refer to the citations from the References in the original submission.

Reviewer 1

1. We apologize for not stating the intuition. Roughly speaking, let $Z = \sum_i^n w_i$ be the sum of the weights of the n training examples, and let the true edge of a weak rule h be γ. Then the expected (normalized) correlation between the predictions of h and the true labels is $E\left(\frac{w_i}{Z} y_i h(x_i)\right) = 2\gamma$. The variance of this correlation can be written as

$$\mathrm{Var}\left[\frac{w_i}{Z} y_i h(x_i)\right] = \frac{1}{n^2} \frac{E\big(w_i^2\big)}{E^2(w_i)} - 4\gamma^2,$$

which is very similar to the form of the "effective number of examples" given in Equation 6.

2. We experimented with the deeper trees (of depth 5) on the cover type dataset (Section 5.1). Since it takes too long to get the training time till convergence on deeper trees on large datasets, especially for XGBoost, we limited the tree depth to 2 on the splice and the bathymetry datasets.

3. Thanks for the suggestion! We will consider it in our future work.

4. In this paper, we focus on the training data size in memory. Using different subset of the features in each boosting round will not reduce the memory footprint.

5. Thanks for the suggestion!

6. Yes, all running time includes the data loading time, which we believe is a practical way of evaluation in the context that the training data mostly resides on disk but cannot fit in the memory.

7. LightGBM uses the histogram-based method for training [12]. In addition, we enabled the "two_round_loading" parameter in all experiments presented in the paper. We will clarify it in the final version of the paper.

8. Removing GOSS will increase the amount of memory needed by LightGBM. We decided to keep it since the memory footprint is the main thing we are trying to optimize.

9. The relationship between martingales, sequential analysis, and stopping rules is somewhat involved [18]. Briefly, when the advantage of a rule is smaller than γ, then the sequence is a supermartingale. If it is larger than γ, then it is a submartingale. **The only assumption is that the examples are sampled i.i.d.**. Theorem 1 guarantees two things about the stopping rule defined in Equation 8: (1) if the advantage is smaller than γ, the stopping rule will never fire (with high probability); (2) if the stopping rule fires, the advantage of the rule h is larger than γ.

10. At the time of writing the paper, we cannot successfully train LightGBM with GOSS on the cover type dataset because the software crashes for some reason that is unclear to us.

11. We believe the reviewer is referring to the first three rows of the top part of Table 1. Specifically, the top three training time of Sparrow. The explanation for the training times on the 8GB and 16GB instances is because the 16GB instance has fewer CPU cores than the 8GB instance. The difference between the 16GB and 32GB remains a mystery.

12. Yes, we ran all experiments on the AWS instances with an attached SSD storage.

Reviewer 2

1. Thank you for the reference! Indeed sequential analysis and stopping rules have a long history in statistics [18]. As discussed in the Related Work section, the "early stopping" technique has been investigated before (e.g. [4]), but they haven't considered using sampling to reduce memory footprint.

2. Our emphasis is on small memory sizes. LightGBM is sometimes faster when the memory size is *large enough to hold all training examples*, though the specific reason for why it is faster on the splice dataset is unclear to us.

3. Thanks for the suggestions!

Thanks for the improvement suggestions! We will address them to the best of our ability in our final version.

Reviewer 3

1. Thanks for the suggestion! We will try to make the explanation clearer.

2. We are using class imbalance (in Section 3.2) as one example in which resampling is beneficial. However, resampling is beneficial in a much wider set of situations, specifically, whenever the effective size of the sample is small. For example, this can happen when most of the examples are easy (have high margins).

Among the two big datasets we used in the experiment, the bathymetry dataset is balanced.