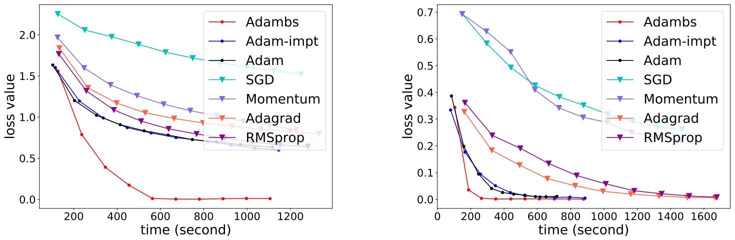
We sincerely thank our reviewers for their insightful comments. Below we address all their comments, numbered as C1-C7 (corresponding answers are numbered as A1-A7).

Reviewer 1

C1: It would be nice if the authors could conduct additional experiments to verify the effectiveness of the proposed methods over existing methods (e.g., SGD, momentum, Adagrad, RMSprop)
A1: We thank the reviewer for mentioning these methods. We have conducted experiments with these competitors. We didn't include them because they have worse performance than Adam in our experiments, which have been shown in the original Adam paper [16]. Due to the space limit, we include the results on two datasets (CIFAR10 and IMDB) in Figure 1 as below, and will happily include them in our camera-ready if accepted.



(a) CNN on CIFAR10

(b) RCNN on IMDB

Figure 1: Additional experiments

Reviewer 2

C2: It's unclear how important the doubly heavy tailed distribution assumption is to the result, and whether this assumption holds in practice.
A2: We thank the reviewer for the comment. The key intuition behind the design of our algorithm is to be adaptive to the difference among the examples. Our method is therefore particularly useful when the dataset is imbalanced. The doubly heavy tailed distribution is just one example of imbalanced data that we used for convergence analysis, as its simplicity facilitates the convergence analysis. Doubly-heavy tail distribution is not a single distribution, rather a family of distributions parameterized by the value of γ. As for its importance, doubly heavy tailed distribution is primarily responsible for getting the $\log(n)$ speedup. For other non-uniform distributions, the convergence will still be faster than Adam, but the speedup may be lower or higher than $\log(n)$ depending on the particular parameters of the distribution.

C3: If we are training models where there are more parameters than training examples, then the proposed method would not make significant difference compared to Adam.
A3: We thank the reviewer for bringing this up. We fully agree that our method shines when the number of training examples is massive. This is particularly aligned with the current trend in deep learning to use more and more data for training. As for the model size issue, we'd like to note that significant efforts in the community have been dedicated to reducing the model size because deep learning models are often over-parameterized. Our method is orthogonal, and thus we think combined with these model-size reduction methods, it will continue to retain its benefits due to its alignment with existing trends in the industry and the research community.

Reviewer 3

C4: is it that the proposed sampling approach is suitable only for ADAM-like procedures?
A4: The reviewer is correct. The idea behind our approach can be extended to other optimization procedures. We chose Adam due to its superior empirical performance and its well-understood theoretical properties. Exploring this extension to other minibatch-based optimizations will make an interesting direction of future research.

C5: What about curriculum learning? How does this approach compare to the state of the art in curriculum learning.
A5: We thank the reviewer for mentioning curriculum learning. Curriculum learning is indeed an important optimization strategy, which leverages a pre-trained teacher model to train the target model. The teacher model is critical in determining mini-batches. We will happily include a discussion and empirical experiments comparing with curriculum learning in our camera-ready if accepted.

C6: In Algorithm 2 Line 2, what is L? What is the influence of L and p_{min} in Algorithm 2?
A6: We apologize for not having made this clear. The L in Algorithm 2 is the upper bound on the gradient norm (formally defined in Lemma 1), which is commonly assumed in related literature. This assumption holds in most cases, especially when gradient clipping trick is applied. The influence of L and p_{min} is related to the convergence analysis of the bandit method. Simply speaking, they make sure the loss $l_{t,j}$ is always nonnegative, ensuring the correctness of distribution update.

C7: It seems that ADAMBS ... a lower loss value ... Is that a typical trend ...? If it is a common trend then was the influence of the sampling procedure on this phenomenon explored?
A7: We think this is a common trend, due to the sampling procedure of our method. The reason is that our method tends to sample examples with large gradient norms at each iteration, which effectively reduces gradient variance as discussed in Section 4.2. Some experiments have shown a similar trend for variance reduction techniques.