We thank the reviewers for their feedback and time! We are encouraged they found our theoretical results "impressive" (R1; score 6), "strong" (R2; score 7), and "interesting" (R4; score 3), noticed that our work "answers an important question" (R3; score 6), and emphasized that the paper is clearly written and well-organized (R1, R2, R3).

R1: ***1) the bounds add...*** Large batchsizes help us to obtain complexity guarantees beating the state-of-the-art ones. Moreover, as R1 pointed out, we also provide analysis with smaller batchsizes in Section F and empirically show that `clipped-SSTM` can work well in practice even with moderate batchsizes. ***2) the exp-al section is limited...*** We will add additional experiments with logistic regression on other datasets. We did not try SVM since it leads to non-smooth optimization while our paper focuses on smooth problems. ***3) On the triple-plots...*** We will modify the plots for the final version. ***4) I think some parts...*** We can add these details to the main body using an additional 9th page.

R2: ***1) The main weakness of this work is...*** We agree with this criticism. However, there are still a lot of important open problems in stochastic *convex* optimization that should be resolved as well. ***2) It seems that clipped-SSTM suffers from oscillations...*** Averaging is an interesting idea, but clipped-SSTM already suffers from oscillation significantly less than SSTM and, we guess than other accelerated stochastic methods. ***3) Some applications of transfer learning...*** We will try to test our methods on this task and investigate the heavy-tailedness of stochastic gradients for this problem.

R3: ***1) In [71] there are several theoretical...*** First of all, [71] contains the analysis of several versions of clipped-SGD establishing the rates of convergence *in expectation* while we focus on the *high-probability* complexity guarantees. Secondly, we consider convex and strongly convex cases while [71] provides an analysis in non-convex and strongly convex cases. Finally, [71] relies on the following assumption: there exist such $G > 0$ and $\alpha \in (1, 2]$ that the stochastic gradient $g(x)$ satisfies $\mathbb{E}\|g(x)\|_2^\alpha \leq G^\alpha$ for all $x \in \mathbb{R}^n$. This assumption implies the boundedness of the gradient of the objective function $f(x)$ on the whole space which is quite restrictive and contradicts to the strong convexity: there is no functions that are strongly convex and have bounded gradients on $\mathbb{R}^n$. One can argue that boundedness of the gradient is needed only on some compact, but this claim requires more refined analysis than one presented in [71]. In contrast, we assume the boundedness of the variance. Moreover, we consider *smooth* problems that allows us to accelerate clipped-SGD and obtain clipped-SSTM, while in [71] there is no analysis showing an acceleration of any algorithm. Taking all of this into account, we conclude that [71] is far from the setup we consider to be mentioned in Table 2, and it doesn't reduce the novelty of our results. ***2) A related result was provided in...*** Indeed, it is highly relevant, but Simsekli et al. focus on *non-convex* problems and rates of convergence *in expectation*. Anyway, we will mention this work in the final version. ***3) Throughout the paper, the authors frame...*** We agree that for non-convex optimization the heavy-tailed gradient noise can be beneficial for escaping bad local minima. We will add a remark about that in the final version. The reason why we treat heavy-tails as a negative outcome is that for a long time, it was not clear whether it is possible to obtain the convergence rates of AC-SA without light-tails assumption. ***4) I find the characterization...*** It is a good direction for further research. However, even in the setup, we consider there were important open questions addressed by our paper. ***5) I find the uniformity of the condition in Eq2...*** Indeed, this assumption is quite strong, but it is used in many papers on stochastic optimization. There are several extensions allowing σ to depend on x for the convergence in expectation, and it would be interesting to develop similar non-trivial extensions for the convergence with high-probability. ***6) This is a minor point: very recently...*** Thank you for the references. However, the examples of the problems they consider do not directly fit to the setup we focus on, but it is an interesting direction for further research. ***7) Overall, it's not clear to me what is...*** We use a classical clipping operator while in [47] authors apply different truncation operator ignoring the direction of the stochastic gradient if it is too big which makes the estimator less accurate in some sense. This is the main reason why in [47] boundedness of the domain is needed. Also, it seems that the direct acceleration of the method from [47] is not possible. ***8) Also as mentioned by the authors, the proof technique...*** In [23], authors consider the convergence in expectation and in [22] authors focus on the convergence with high-probability under "light-tails assumption" which offers them to apply Azuma–Hoeffding inequality. In our paper, we consider high-probability convergence without the "light-tails assumption" that forces us to apply Bernstein's inequality and makes the analysis more complicated. So, our analysis is novel while it reminds the approach used in [22] and [23]. ***9)*** y^0 ***and*** z^0 ***are not defined...*** Thanks for spotting this. The definition is $y^0 = z^0 = x^0$, we will add it.

R4: ***1) While it is not the first result for clipped SGD...*** We have never claimed this: we gave the first *high-probability* convergence results for clipped-SGD *without "light-tails assumption"* and also managed to *accelerate* clipped-SGD. ***2) One of the weaknesses is the paper misses discussions and comparisons with significantly related research, particularly [71]...*** We politely disagree. First of all, see R3 ***(1)***. Secondly, we do compare our results with all relevant research that we aware of. R4 mentioned only [71] as an example of "missing comparison" which is far from the setup that we consider to be discussed in detail. ***3) The experiment section is too weak...*** The main contribution is theoretical, so, our experimental part is not too weak: it justifies our theoretical findings. Figure 2: the presence of the outliers can be considered as a source of heavy-tailedness. See also R1 ***(2)***. ***3) Lastly, the paper is not reading smoothly...*** We politely disagree. Moreover, R1, R2, R3 found our work clearly written. Next, restarts technique is a classical tool in optimization (e.g., see "Gradient methods for minimizing composite objective function"). Finally, R4 claimed ***"Many of the essential discussions are not in the main paper"*** but did not provide any examples of such discussions. ***4) Some of their claims need... Missed many important papers such as [71].*** This is not true, see R4 ***(1)***.

Conclusion: we politely disagree with the main criticism of R4 and respectfully ask R4 to increase the score.