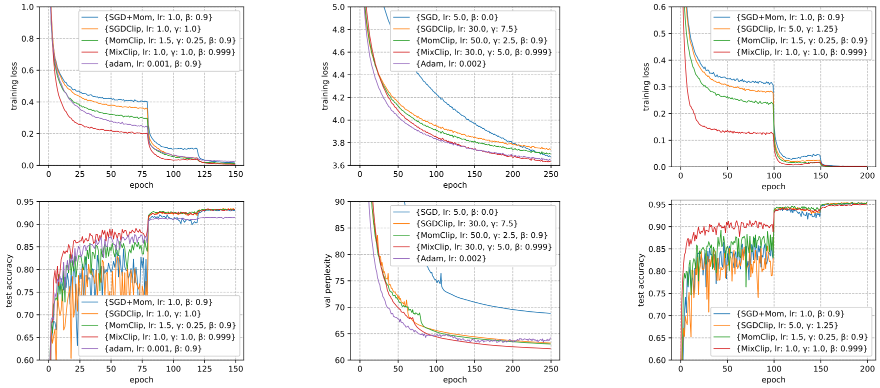
We would like to thank the entire review team for their efforts and comments. In particular, we would like to thank Reviewer 1 and 2 for the positive comments and Reviewer 3 for sharing the concerns.

To Reviewer 1 We would like to thank the suggestions from Reviewer 1. We will follow these suggestions in the revised version. Specifically, we will move the imagenet experiments to the main text and repeat the experiments in 10 times. We will change the term "energy funcion" to "Lyapunov function", and add more discussions of the mixed clipping method. The test errors/validation perplexities of all the algorithms are shown in the following table.

	CIFAR-10 test acc	PTB validation ppl	ImageNet validation top1 acc
SGD	93.0	68.87	76.1
SGD Clip	93.3	63.25	75.9
Mom Clip	93.2	63.05	76.1
Mix Clip	93.2	62.17	76.1

To Reviewer 2 We would like to thank the suggestions from Reviewer 2. We have added experiments using Adam optimizer with best hyper-parameters (see Figures (a, b)). Results show that the training speed using Adam is faster than using baseline SGD. However, Adam generalizes worse. We also add experiments using the same ResNet architecture (ResNet20) and the same hyper-parameters as Zhang et al. [2020] (see Figure (c)). All algorithms can achieve 95% test accuracy (as reported in Zhang et al. [2020]), and the training curve is similar to Figure (a). For PTB dataset, the validation loss is 4.13 using mixed clipping. We will plot standard deviation as shaded area in the revised version of our paper. We turn the step size and clipping hyper-parameters by grid-search.



(a) CIFAR-10 (b) PTB (c) CIFAR10 (ResNet20)

To Reviewer 3 about the noise assumption We would like to thank Reviewer 3 for raising this concern. We justify the reasonability of our assumption below:

(A) Our paper follows the research line typically from Zhang et al. [2020]. This research line attempts to understand the strength of clipping methods for non-smooth (in the traditional sense) objective functions. Note that Zhang et al. [2020] have made the *same* assumptions, in that they also assume the noise is bounded. We improve their complexity under the same conditions.

(B) We are aware that Ghadimi and Lan [2013] obtains the upper bound complexity under a weaker assumption. However, shown in Section 3.3 (line 234), there is a hard objective function that satisfies our (stronger) assumption, for which the Stochastic Gradient Descent algorithm must take $\Omega\left(\Delta L \sigma^2 \epsilon^{-4}\right)$ to find a first-order stationary point. Hence Ghadimi's result *cannot* be further improved under our (stronger) assumption. In contrast, we show that clipping methods enjoy a better complexity of $\mathcal{O}\left(\Delta L_0 \sigma^2 \epsilon^{-4}\right)$.

(C) We think that the bounded (or Sub-Gaussian tail) noise assumption is quite *common* in the non-convex stochastic optimization field. Many analyses adopt this assumption to prove convergence, especially for adaptive gradient methods (e.g. Li and Orabona [2019]) and escaping saddle points (e.g. Fang et al. [2018]).

To Reviewer 3 about the clipping parameters and step sizes. Sorry for the confusion about clipping parameters. Taking Theorem 3.2 as an example. The clipping threshold of gradient norm is actually $\gamma/\eta = 5\sigma = \Theta(1)$, which is *at constant magnitude* and independent of ϵ. It is true that the step size provided in this result is $\mathcal{O}(\epsilon^2)$. This step size is common in the analyses of non-convex stochastic algorithms (e.g. Ghadimi and Lan [2013]). In practice, we may choose a relatively large step size at beginning and gradually decrease it. Our analysis can also be extended to this setting. However, in the final state, the step size still needs to be $\mathcal{O}(\epsilon^2)$.

To Reviewer 3 about the best choice of β. Theorem 3.2 implies that for sufficiently small ϵ, the number of iterations needed to reach an ϵ-stationary point *does not* depend on β.