Reviewer 1: We very much thank the reviewer for the excellent and thorough feedback!
1. "Typos and inaccuracies". Minor edits such as awkward wording and incorrect subscripts (e.g. $\eta_k = \eta = ...$ will be corrected following the reviewer's suggestion. We next address the major technical concerns raised.
Detailed explanation of remark 1: Though the function we are optimizing $f(x) = x^2/2$ is deterministic, we use a **stochastic gradient oracle** $g_k = g(x_k) = \nabla f(x) + \xi = x + \xi$, where $\xi \in \mathbb{R}^d$ is a random variable with $\mathbb{E}\|\xi\|^2 = \infty, \mathbb{E}\|\xi\|^\alpha = \sigma^\alpha, \mathbb{E}[\xi] = \vec{0}$. It is the noise in this gradient oracle which causes divergence. Specifically, $\mathbb{E}[\|\nabla f(x_{k+1})\|^2] = \mathbb{E}[\|x_{k+1}\|^2] = \mathbb{E}\|x_k - \eta_k g_k\|^2 = \mathbb{E}\|x_k - \eta_k(x_k + \xi)\|^2 = \mathbb{E}\|(1-\eta_k)x_k - \eta_k\xi\|^2 = \mathbb{E}\|(1-\eta_k)x_k\|^2 - 2(1-\eta_k)\eta_k x_k^\top \mathbb{E}[\xi] + \eta_k^2\mathbb{E}\|\xi\|^2 \geq \eta_k^2\mathbb{E}\|\xi\|^2 = \infty$. Note that this holds for **any** fixed $\eta_k > 0$ even if allowed to depend on the statistics of the noise distribution (such as σ or α).
Detailed explanation of line 405: We agree with the reviewer that more details and minor corrections are needed in the appendix, but our claims remain unaffected. We hope the reviewer, if convinced by the derivation below, could confirm the correctness of our proof in the discussion. The intuition for case 2 is that when gradient is very large($\tau = \mathcal{O}(\eta^{-1/\alpha}) = \mathcal{O}(K^{1/(3\alpha-2)})$), noise gets dominated by the gradient. Recall that $\|\nabla f(x_k)\| > \tau/2$, $\tau = \sigma(\eta L)^{-1/\alpha}$, $p = \mathbb{P}\{\|g_k\| \leq \tau\}$, and that expectation is taken wrt g_k. We use ∇f as a shorthand for $\nabla f(x_k)$:

$$\begin{aligned}
&\mathbb{E}[\langle \nabla f, g_k\rangle \mathbb{1}_{\{\|g_k\|\leq\tau\}}] \geq \mathbb{E}[(\|\nabla f\|^2 - \|\nabla f\|\|g_k - \nabla f\|)\mathbb{1}_{\{\|g_k\|\leq\tau\}}] \\
\geq &\mathbb{E}[\|\nabla f\|^2\mathbb{1}_{\{\|g_k\|\leq\tau\}} - \tfrac{1}{2}\|\nabla f\|^2\mathbb{1}_{\{\|g_k\|\leq\tau, \|g_k-\nabla f\|\leq\tau/4\}} - \|\nabla f\|\|g_k - \nabla f\|\mathbb{1}_{\{\|g_k\|\leq\tau, \|g_k-\nabla f\|\geq\tau/4\}}] \\
\geq &\tfrac{p}{2}\|\nabla f\|^2 - \|\nabla f\|\mathbb{E}[\|g_k - \nabla f\|\mathbb{1}_{\{\|g_k-\nabla f\|\geq\tau/4\}}] \geq \tfrac{p}{2}\|\nabla f\|^2 - \|\nabla f\|\tfrac{\sigma^\alpha}{(\tau/4)^{\alpha-1}}
\end{aligned}$$

The first inequality uses $\langle \nabla f, g_k\rangle = \|\nabla f\|^2 + \langle \nabla f, g_k - \nabla f\rangle$. The second line follows by $\|\nabla f\| > \tau/2$ and $\|g_k - \nabla f\| < \tau/4 \implies -\|\nabla f\|\|g_k - \nabla f\| \geq -\|\nabla f\|^2/2$. Then, we use the fact that $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$ for any random events A and B. The last inequality follows by $\sigma^\alpha \geq \mathbb{E}[\|g_k - \nabla f\|^\alpha] \geq \mathbb{E}[\|g_k - \nabla f\|(\frac{\tau}{4})^{\alpha-1}\mathbb{1}_{\|g_k-\nabla f\|\geq\tau/4\}}]$. With the above, we go back to line 405,

$$\begin{aligned}
\mathbb{E}[\langle \nabla f, \hat{g}_k\rangle] &= \mathbb{E}[\langle \nabla f, g_k\rangle \mathbb{1}\{\|g_k\| \leq \tau\}] + \mathbb{E}[\langle \nabla f, g_k/\|g_k\|\rangle \mathbb{1}\{\|g_k\| \geq \tau\}] \\
&\geq \tfrac{p}{2}\|\nabla f\|^2 - \|\nabla f\|\tfrac{\sigma^\alpha}{(\tau/4)^{\alpha-1}} + (1-p)\|\nabla f\|/3 - \tfrac{8}{3}\mathbb{E}[\|\nabla f - g_k\|] \\
&\geq \|\nabla f\|/3 - \|\nabla f\|\tfrac{\sigma^\alpha}{(\tau/4)^{\alpha-1}} - \tfrac{8\sigma}{3} \geq \|\nabla f\|/3 - \|\nabla f\|/12 - \|\nabla f\|/6
\end{aligned}$$

The second line follows by Lemma 11. Since $\tau \geq 2$ (informally, recall $\tau = \mathcal{O}(\eta^{-1/\alpha}) = \mathcal{O}(K^{1/(3\alpha-2)})$ increases with total steps K), and $\|\nabla f\| \geq \tau/2$, we have $\frac{p}{2}\|\nabla f\|^2 \geq p\|\nabla f\|/3$. Then, by imposing $\eta \leq \frac{1}{L(48\sigma)^{\alpha/(\alpha-1)}}$ and since $\tau = \sigma(\eta L)^{-1/\alpha}$ by defn., we have $\frac{\sigma^\alpha}{(\tau/4)^{\alpha-1}} \leq \frac{1}{12}$. Finally, combining with line 407 we gethe desired result in the last inequality above. Note that the proof above corrects some mistakes and hence differs slightly from the submission.
2. "Strongly convex and bounded second moment implies bounded domain". This is true and is the classical setting for all stochastic subgradient methods under strongly convex assumption (e.g. Chapter 6.1 Bubeck, Sébastien. "Convex optimization: Algorithms and complexity. "). Since this is a standard setting, we briefly discussed the bounded domain in line 369. We will move these comments to Thm 4 as the reviewer suggested.

Reviewer 2: We thank the reviewer for detailed comments. We address the reviewer's question as follows and will edit the paper accordingly. **0. "Novelty of Clipping"** As reviewer said, the idea of clipping is not new but the understanding has been heuristic (to avoid gradient explosion). We provide solid theoretical justification for its advantage by formalizing clipping's faster convergence under the experiment motivated by heavy-tailed condition. This help explain the mismatch between theory (SGD is optimal; adaptive methods give worse convergence bound) and practice (adaptive methods converge much faster in many settings). **1. "Experimental setting of Figure 1":** We fix the model parameters and keep sampling minibatches without updating the model (see line 80) to plot Figure 1(b)(f). **2. "Remark 1":** Please see point 1 in response to Reviewer 1. **3. "Thm 2":** K is the number of iterations, and x_k is the variable at iteration k. Please take a look at line 130. The convergence is in the average sense as we are summing up all the iterate and then divide by K. **4. "Thm 4":** Assump. 4 assumes strong convexity and is in the appendix due to limited space. We will remove some experiments from the main part and move the optimization setting (hence clarifying Figure 1) to the main text. See the "explanation of line 405" in response to reviewer 1 for more proof details. **5. "Line 146":** $\mathbb{E}[|g(x)|^\alpha] \leq 1$ refers to the function that we used to prove the lower bound. In other words, the counter-example presented in appendix F satisfies $\mathbb{E}[|g(x)|^\alpha] \leq 1$. Note that this is a stricter condition than that used by our upper-bound, making our lower bound even stronger. **6. "Update of the threshold":** We use the exponential moving average estimator motivated by ADAM, which also uses it to estimate moments. **7. "Other":** Evaluation loss = validation set, and wd = weight decay. Hence, equal number of parameters are tuned for ADAM and ACClip.

Reviewer 3: 1. We used the statistical estimator studied in [Simsekli, et al ICML 2018]. Results are in Figure 8. 2. Please refer to Reviewer 2 point 1 for experiment setup. 3. Please refer to Reviewer 1 point 1 for proof of remark 1.

Reviewer 4: We appreciate the reviewer's comment. We will address the typos and further improve readability by incorporating all reviewer's comments.