R1.1 *...the title mentioned "An Improved Analysis"...should be more clear in the main contributions...* Great suggestion! Our theory introduces a new Lyapunov function to analyze momentum SGD methods, it does not require uniformly bounded gradients and leads to *the first convergence bound for Multistage SGDM*. Will explain this in the contributions.

R1.2 *Important...Normally stochastic heavy ball method has no* $1-\beta$... The two are equivalent, as claimed in Page 1 footnote. Their update vectors $m^k = (1-\beta)\sum_{i=1}^k \beta^{k-i}\tilde{g}^i$ and $m^k = \sum_{i=1}^k \beta^{k-i}\tilde{g}^i$ only differ by a scaling.

R1.3 *...How the acceleration is justified?...a table is needed...* At the initial stages, the sublinear terms in the convergence bounds dominate, and Multistage SGDM allows for larger stepsizes, so it is faster during the initial stages. This may not hold for final stages, although we also observe acceleration in the final stages in our numerical tests. We apologize for the confusion and will clarify this in the revision! A table will be added to compare the convergence bounds.

R1.4 *...theorem 1 looks informal. For Theorem 2 it was assumed that* $k > k_0$.... We will include Assumption 1 in Thm 1. For Thm 2, $k_0 = \lfloor \frac{\log 0.5}{\log \beta} \rfloor$ is a fixed constant. we will highlight this fact in the paper.

R1.5 *...it is mentioned that item 2 is used...Can you elaborate...?* In fact by item 2 we have $\mathbb{E}_{\zeta^i}[\tilde{g}^i] = g^i$, where $\mathbb{E}_{\zeta^i}$ refers to taking expectation with respect to the minibatch $\zeta^i$ at the $i$th iteration. Therefore, when expanding $\mathbb{E}[\|\sum_{i=1}^k \beta^{k-i}(\tilde{g}^i - g^i)\|^2]$, we have $\mathbb{E}_{\zeta^1}\mathbb{E}_{\zeta^2}...\mathbb{E}_{\zeta^k}[\langle \beta^{k-i}(\tilde{g}^i - \mathbb{E}_{\zeta^i}[\tilde{g}^i]), \beta^{k-j}(\tilde{g}^j - \mathbb{E}_{\zeta^j}[\tilde{g}^j])\rangle] = 0$ when $i \neq j$.

R1.6 *Closely related references...* We have read these interesting papers! Will cite and discuss them in our paper.

R2.1 *...Since* $z_k$ *is not a convex combination of* $x_k$*, I am curious about how it will influence our results.* Interesting question! We do have similar convergence results for $\mathbb{E}[f(x^k)]$: since $x^k = (1-\beta)\sum_{i=2}^k \beta^{k-i}z^i + \beta^{k-1}z^1$, from which we can quickly get $\mathbb{E}[f(x^k) - f^*] = \mathcal{O}(r^k + \alpha\sigma^2)$, where $r = \max\{\beta, 1-\alpha\mu\}$. Will add this in the revision.

R3.1 *(7) requires that* $\alpha_i$ *and* $\beta_i$ *have to change...which makes the multistage setting less practical.* We respectfully disagree. Stagewise SGD with different stepsizes $\alpha_i$ is widely applied in practice. In addition to this, our Multistage SGDM only needs to compute $\beta_i$ using (7), which costs little. The update vectors follow a very simple recursion. We also demonstrated the effectiveness of Multistage SGDM in our numerical tests.

R3.2 *...all the theorems on the SGDM don't improve SGD, the results seems not interesting.* SGDM as well as its multistage variants are known to work well in training DNNs such as ResNet and DenseNet. But, their convergence properties remain largely unexplained. Our work narrows the gap between theory and practice. As pointed out by R1, R2, and R5, our analysis brings new insights on the effect of momentum and improves previous results.

The convergence rate of SGDM under strong convexity (Thm 2) already matches the lower bound in Prop. 3 of arXiv:1803.05591, so the worst-case guarantee cannot be improved. We believe that the same holds under nonconvexity.

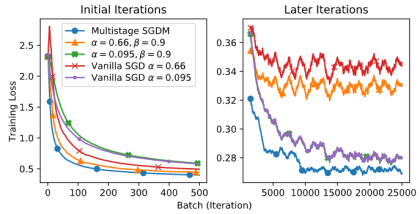
R3.3 *The proof of Lemma 1 seems wrong..might be fixable.* Thanks for catching the glitch! It is fixable. We have the same upper bound for $\mathbb{E}[\|m^k - (1-\beta)\sum_{i=1}^k \beta^{k-i}g^i\|^2]$, and the rest of the proof only needs minor changes.

R3.4 *some notations...* Agreed! $\mathbb{E}$ and Var should be $\mathbb{E}_{\zeta^k}$ and $\text{Var}_{\zeta^k}$, respectively. We will also use $g^k$ throughout.

R5.1 *...SGDM requires smaller stepsizes...* Agreed. It is possible that SGD is faster in certain cases. Will discuss this.

R5.2 *...is the rate of multi-stage SGDM better than SGDM?* Thank you for pointing this out. Multistage SGDM is faster during the initial stages (See R1.2). We will clarify this!.

R5.3 *... no experimental comparison to the SGD...* Thanks for the valuable suggestion! The comparison with SGD on MNIST can be found in the figure below, where SGD has a similar performance as SGDM when $\alpha = 0.095$, and is slightly slower when $\alpha = 0.66$. We will also add SGD in ResNet18 experiments.




*R5.4 Fig. 2 doesn't match with it's description...* We apologize that Fig.2 used a wrong file due to an inadvertent overwriting. The description in lines 207-209 corresponds to the figure on the left, where we have also two added vanilla SGD curves as requested. Multistage SGDM is faster in both the initial and final stages. The curves can be reproduced by our submitted code.

R5.5 *...why is it required to increase momentum coefficient in Multistage SGDM?* The convergence theory for Multistage SGDM requires the $\beta_i$ to satisfy eq (7). We agree that Multistage SGDM also works with a fixed momentum.

R5.6 *...none of the algorithms achieve SOTA performance...* We agree that using parameter choices with convergence guarantees may not give the SOTA performance. However, we just find that Multistage SGDM can also achieve a test accuracy of 93.0% with 200 epochs. Will mentioned this.