A Note on the Convergence of Muon and Further

Jiaxiang Li * Mingyi Hong †

Abstract

In this note, we inspect the convergence of a new optimizer for pretraining LLMs, namely the Muon optimizer. Such an optimizer is closely related to a specialized steepest descent method where the update direction is the minimizer of the quadratic approximation of the objective function under spectral norm (Bernstein and Newhouse, 2024). We provide the convergence analysis on both versions of the optimizer and discuss its implications.

1 Introduction

Recently, a new optimizer named Muon[1] has drawn attentions due to its success in training large models. Consider the following stochastic optimization problem:

$$\min_X f(X) = \mathbb{E}_\xi[F(x,\xi)] \tag{1.1}$$

where the variable $X \in \mathbb{R}^{m\times n}$ is a matrix. Without loss of generality we assume $m \geq n$. Muon optimizer writes:

$$\begin{aligned} G_t &\leftarrow \nabla F(X_t, \xi_t) \\ B_t &\leftarrow \mu B_{t-1} + G_t \quad \text{Nesterov momentum} \\ O_t &\leftarrow \operatorname*{argmin}_O\{\|O - B_t\|_F : O^\top O = I \text{ or } OO^\top = I\} \\ X_{t+1} &\leftarrow X_t - \eta_t O_t \end{aligned} \tag{1.2}$$

where the O step can be equivalently written as $O_t = UV^\top$ where $B_t = U\Sigma V^\top$ is the singular value decomposition.

On the other hand, a closely related formula of (1.2) is the following spectral optimizer:

$$\begin{aligned} G_t &\leftarrow \nabla F(X_t, \xi_t) \\ B_t &\leftarrow \mu B_{t-1} + G_t \\ \Delta_t &\leftarrow \operatorname*{argmin}_\Delta\{\mathrm{tr}(B_t^\top \Delta) + \frac{1}{2\eta_t}\|\Delta\|_2^2\} \\ X_{t+1} &\leftarrow X_t + \Delta_t \end{aligned} \tag{1.3}$$

When $\|\cdot\|_2$ is the matrix 2-norm (spectral norm), i.e. the largest singular value, the solution of the third line of (1.3) is $\Delta_t = -\eta_t\|B_t\|_* UV^\top$, where $B_t = U\Sigma V^\top$ is the singular value decomposition (see Bernstein and Newhouse (2024, Proposition 5)). In this case (1.3) only differs from (1.2) by the norm $\|B_t\|_*$.

*Department of Electrical and Computer Engineering, University of Minnesota. li003755@umn.edu
†Department of Electrical and Computer Engineering, University of Minnesota. mhong@umn.edu

[1] see https://kellerjordan.github.io/posts/muon/.

2 Convergence of Muon

Now we go back to the original Muon with a slight change:

$$\begin{aligned} G_t &\leftarrow \frac{1}{B}\sum_{i=1}^{B}\nabla F(X_t,\xi_{t,i}), \\ B_t &\leftarrow \beta B_{t-1}+(1-\beta)G_t, \\ O_t &\leftarrow U_tV_t^\top, \text{ with SVD } B_t=U_tS_tV_t^\top, \\ X_{t+1} &\leftarrow X_t-\eta_t O_t, \end{aligned} \tag{2.1}$$

where in the second line G_t is multiplied by $1-\beta$ as compared to (1.2). (2) is a heavy-ball method comparing to the Nesterov-type method (1.2).

Note that since we assumed $X\in\mathbb{R}^{m\times n}$ and $m\geq n$, the SVD satisfies $U_t\in\mathbb{R}^{m\times r}$, $S_t\in\mathbb{R}^{r\times r}$ and $V_t\in\mathbb{R}^{n\times r}$, where r is the rank of B_t, which we certainly need to assume to be positive. When $n=1$,i.e. the vector case, reduces to the normalized gradient descent with momentum. The convergence of normalized gradient descent with momentum is analyzed in Cutkosky and Mehta (2020).

We now state the assumptions. Here the assumptions are with respect to the Frobenius norm. We need to assume the Lipschitzness of the gradient in Frobenius norm (a plain generalization of the Lipschitz smoothness in vector case), i.e.

$$\|\nabla f(X)-\nabla f(Y)\|_F\leq L\|X-Y\|_F \tag{2.2}$$

which implies (Beck, 2017, Lemma 5.7)

$$f(Y)\leq f(X)+\langle\nabla f(X),Y-X\rangle+\frac{L}{2}\|X-Y\|_F^2. \tag{2.3}$$

The inner product is just the Frobenius inner product $\langle X,Y\rangle=\mathrm{tr}(X^\top Y)$. Another assumption is that $\nabla F(X,\xi)$ is an unbiased estimator of $\nabla f(X)$ with variance bounded by σ^2, i.e. $\mathbb{E}\|\nabla F(X,\xi)-\nabla f(X)\|_F^2\leq\sigma^2$. From now on we use $\|\cdot\|$ to denote the Frobenius norm throughout this section.

Now we move to convergence analysis. The main body of the analysis follows Cutkosky and Mehta (2020), where we have a specific new descent lemma-type result for Frobenius norm.

Lemma 2.1. *For update (2), we have that*

$$f(X_{t+1})\leq f(X_t)-\frac{\eta_t}{4}\|\nabla f(X_t)\|+\frac{5}{2}\eta_t\|\nabla f(X_t)-B_t\|+\frac{\eta_t^2nL}{2}.$$

If we take $\eta_t=\eta$ a constant, we have

$$\sum_{t=1}^{T}\|\nabla f(X_t)\|\leq\frac{4(f(X_1)-f^*)}{\eta}+10\sum_{t=1}^{T}\|\nabla f(X_t)-B_t\|+2\eta nLT.$$

Proof. By Lipschitz smoothness we have

$$\begin{aligned} f(X_{t+1})\leq& f(X_t)+\langle\nabla f(X_t),X_{t+1}-X_t\rangle+\frac{L}{2}\|X_{t+1}-X_t\|^2 \\ =& f(X_t)-\eta_t\langle\nabla f(X_t),O_t\rangle+\frac{\eta_t^2nL}{2} \end{aligned} \tag{2.4}$$

since $\|X_{t+1}-X_t\|=\eta_t\|O_t\|=\eta_t\sqrt{n}$.

Now we analyze the term $-\eta_t\langle\nabla f(X_t), O_t\rangle$. Suppose $B_t = U_t S_t V_t^\top$ being the SVD decomposition of B_t so that $O_t = U_t V_t^\top$. Then

$$\begin{aligned}
-\langle\nabla f(X_t), O_t\rangle =& -\operatorname{tr}(\nabla f(X_t)^\top U_t V_t^\top) \\
=& -\operatorname{tr}(\nabla f(X_t)^\top U_t S_t V_t^\top V_t S_t^{-1} V_t^\top) \\
=& -\operatorname{tr}(\nabla f(X_t)^\top B_t V_t S_t^{-1} V_t^\top) \\
=& -\langle\nabla f(X_t) V_t S_t^{-1/2}, B_t V_t S_t^{-1/2}\rangle
\end{aligned} \tag{2.5}$$

For simplicity denote $\nabla_t = \nabla f(X_t)$, $\tilde{\nabla}_t = \nabla f(X_t) V_t S_t^{-1/2}$ and $\tilde{B}_t = B_t V_t S_t^{-1/2}$. We have

$$\begin{aligned}
&\|\nabla_t\| = \|\nabla_t V_t S_t^{-1/2} S_t^{1/2} V_t^\top\| \le \|\tilde{\nabla}_t\| \|S_t^{1/2}\|_2 \Rightarrow \|\tilde{\nabla}_t\| \ge \frac{\|\nabla_t\|}{\|S_t^{1/2}\|_2} \\
&\|\tilde{\nabla}_t - \tilde{B}_t\| = \|(\nabla_t - B_t) V_t S_t^{-1/2}\| \le \|(\nabla_t - B_t)\| \|S_t^{-1/2}\|_2
\end{aligned} \tag{2.6}$$

Now back to (2.5), we have

$$\begin{aligned}
-\langle\nabla f(X_t), O_t\rangle =& -\langle\nabla f(X_t) V_t S_t^{-1/2}, B_t V_t S_t^{-1/2}\rangle = -\langle\tilde{\nabla}_t, \tilde{B}_t\rangle \\
=& \frac{1}{2}\left(\|\tilde{\nabla}_t - \tilde{B}_t\|^2 - \|\tilde{\nabla}_t\|^2 - \|\tilde{B}_t\|^2\right) \\
\overset{(2.6)}{\le}& \frac{1}{2}\left(\|(\nabla_t - B_t)\|^2 \|S_t^{-1/2}\|_2^2 - \frac{\|\nabla_t\|^2}{\|S_t^{1/2}\|_2^2}\right) \\
\le& \frac{\|(\nabla_t - B_t)\|^2 - \|\nabla_t\|^2}{2\|S_t^{1/2}\|_2^2} = \frac{\|(\nabla_t - B_t)\|^2 - \|\nabla_t\|^2}{2\|S_t\|_2} \\
=& \frac{\|\nabla_t - B_t\|^2 - \|\nabla_t\|^2}{2\|B_t\|_2}
\end{aligned} \tag{2.7}$$

where the last inequality is due to $\|S^{-1}\|_2 \le 1/\|S\|_2$. Now if $2\|\nabla_t - B_t\| \le \|\nabla_t\|$, we have (note that $\|B_t\|_2 \le \|B_t\|$)

$$\begin{aligned}
-\langle\nabla f(X_t), O_t\rangle \le& \frac{\|\nabla_t - B_t\|^2 - \|\nabla_t\|^2}{2\|B_t\|_2} \\
\le& -\frac{3}{8}\frac{\|\nabla_t\|^2}{\|B_t\|_2} \le -\frac{3}{8}\frac{\|\nabla_t\|^2}{\|B_t\|} \\
=& -\frac{3}{8}\frac{\|\nabla_t\|^2}{\|B_t - \nabla_t + \nabla_t\|} \\
\le& -\frac{3}{8}\frac{\|\nabla_t\|^2}{\|B_t - \nabla_t\| + \|\nabla_t\|} \\
\le& -\frac{3}{8}\frac{\|\nabla_t\|^2}{\frac{1}{2}\|\nabla_t\| + \|\nabla_t\|} = -\frac{1}{4}\|\nabla_t\|
\end{aligned}$$

Otherwise if $2\|\nabla_t - B_t\| > \|\nabla_t\|$, we have

$$\begin{aligned}
-\langle\nabla f(X_t), O_t\rangle \le& \|\nabla_t\| \|O_t\|_2 = \|\nabla_t\| \\
=& -\frac{1}{4}\|\nabla_t\| + \frac{5}{4}\|\nabla_t\| \\
\le& -\frac{1}{4}\|\nabla_t\| + \frac{5}{2}\|\nabla_t - B_t\|
\end{aligned}$$

So in either cases we have

$$-\langle \nabla f(X_t), O_t \rangle \leq -\frac{1}{4}\|\nabla_t\| + \frac{5}{2}\|\nabla_t - B_t\| \tag{2.8}$$

Plugging this back to (2.4) we get

$$f(X_{t+1}) \leq f(X_t) - \frac{\eta_t}{4}\|\nabla f(X_t)\| + \frac{5}{2}\eta_t\|\nabla f(X_t) - B_t\| + \frac{\eta_t^2 nL}{2} \tag{2.9}$$

and the second equation in the lemma statement is obtained by summing up (2.9) from $t = 1, ..., T$. □

Theorem 2.1. *Let $R = f(X_1) - f^*$. If we take $\beta = 1 - \alpha$ with $\alpha = \min(\frac{\sqrt{RL}}{\sigma\sqrt{T}}, 1)$, also $\eta_t = \eta = \sqrt{\frac{4R}{(10/\alpha+2n)TL}}$ and $B = 1$ (batch free convergence), then for update (2) we have*

This seems like a very arbitrary case

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f\left(X_t\right)\|\right] \leq \mathcal{O}\left(\frac{\sqrt{nRL}}{\sqrt{T}} + \frac{\sigma^2}{\sqrt{RLT}} + \frac{\sqrt{\sigma}(RL)^{1/4}}{T^{1/4}}\right).$$

If we take β as an arbitrary constant, then we will need to take $B = T$, so that

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\| \leq \mathcal{O}\left(\frac{\sqrt{nRL}}{\sqrt{T}} + \frac{\sigma}{T^{3/2}} + \frac{\sigma}{\sqrt{T}}\right).$$

Proof. The proof follows Cutkosky and Mehta (2020). Denote $\hat{\delta}_t = B_t - \nabla f(X_t)$, $\delta_t = G_t - \nabla f(X_t)$ and $S(X, Y) = \nabla f(X) - \nabla f(Y)$. Note that we have

$$\begin{aligned} &\mathbb{E}[\delta_t] = 0,\ \mathbb{E}\left[\|\delta_t\|^2\right] \leq \frac{\sigma^2}{m}, \\ &\mathbb{E}[\langle \delta_i, \delta_j \rangle] = 0,\ \forall i \neq j \\ &\|S(X, Y)\| \leq L\|X - Y\| \end{aligned} \tag{2.10}$$

Now following the update in (2), we get

$$\begin{aligned} \hat{\delta}_{t+1} &= \beta\hat{\delta}_t + (1-\beta)\delta_t + S(X_t, X_{t+1}) \\ &= \beta^t\hat{\delta}_1 + (1-\beta)\sum_{\tau=0}^{t-1}\beta^\tau\delta_{t-\tau} + \sum_{\tau=0}^{t-1}\beta^\tau S(X_{t-\tau}, X_{t+1-\tau}), \end{aligned}$$

therefore

$$\|\hat{\delta}_{t+1}\| \leq \beta^t\|\hat{\delta}_1\| + (1-\beta)\|\sum_{\tau=0}^{t-1}\beta^\tau\delta_{t-\tau}\| + \eta L\sum_{\tau=0}^{t-1}\beta^\tau.$$

Taking expectation we get (using the fact that $\hat{\delta}_1 = \delta_1$)

$$\begin{aligned} \mathbb{E}\|\hat{\delta}_{t+1}\| \leq& \beta^t\frac{\sigma}{\sqrt{B}} + (1-\beta)\sqrt{\sum_{\tau=0}^{t-1}\beta^{2\tau}\frac{\sigma^2}{B}} + \eta L\sum_{\tau=0}^{t-1}\beta^\tau \\ \leq& \frac{\sigma}{\sqrt{B}}\beta^t + \frac{\sigma}{\sqrt{B}}\frac{1-\beta}{\sqrt{1-\beta^2}} + \eta L\frac{1}{1-\beta} \\ \leq& \frac{\sigma}{\sqrt{B}}\beta^t + \frac{\sigma}{\sqrt{B}}\sqrt{1-\beta} + \eta L\frac{1}{1-\beta}. \end{aligned}$$

In conclusion we get

$$\sum_{t=1}^{T}\mathbb{E}\|\hat{\delta}_{t+1}\| \leq \frac{\sigma}{(1-\beta)\sqrt{B}} + T\sqrt{1-\beta}\frac{\sigma}{\sqrt{B}} + \frac{T\eta L}{1-\beta}.$$

Now utilize Lemma 2.1, we get

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla f(X_t)\| \leq & \frac{4(f(X_1)-f^*)}{\eta} + 10\sum_{t=1}^{T}\|\nabla f(X_t) - B_t\| + 2\eta nLT \\ \leq & \frac{4(f(X_1)-f^*)}{\eta} + 10\frac{\sigma}{(1-\beta)\sqrt{B}} + 10T\sqrt{1-\beta}\frac{\sigma}{\sqrt{B}} + 10\frac{T\eta L}{1-\beta} + 2\eta nLT \\ \leq & \frac{4R}{\eta} + 10\frac{\sigma}{(1-\beta)\sqrt{B}} + 10T\sqrt{1-\beta}\frac{\sigma}{\sqrt{m}} + 10\frac{T\eta L}{1-\beta} + 2\eta nLT.\end{aligned}$$

Now we just need to take $\eta = \sqrt{\frac{4R}{(10/(1-\beta)+2n)TL}}$ so that

$$\sum_{t=1}^{T}\|\nabla f(X_t)\| \leq 2\sqrt{(10/(1-\beta)+2n)RTL} + 10\frac{\sigma}{(1-\beta)\sqrt{B}} + 10T\sqrt{1-\beta}\frac{\sigma}{\sqrt{B}}.$$

Now we have two types of parameter choice. If we take $B = 1$ (batch size free), we need to take $1-\beta = \min(1, \frac{\sqrt{RL}}{\sigma\sqrt{T}})$ so that we have

$$\sum_{t=1}^{T}\|\nabla f(X_t)\| \leq 2\sqrt{2nRTL} + 10\sigma^2\sqrt{\frac{T}{RL}} + 2\sqrt{10\sigma}(RL)^{1/4}T^{3/4} + 10\sqrt{\sigma}(RL)^{1/4}T^{3/4},$$

thus

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\| \leq \mathcal{O}\left(\frac{\sqrt{nRL}}{\sqrt{T}} + \frac{\sigma^2}{\sqrt{RLT}} + \frac{\sqrt{\sigma}(RL)^{1/4}}{T^{1/4}}\right).$$

If we take β as an arbitrary constant in $(0,1)$, then we will need to take $B = T$, so that

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\| \leq \mathcal{O}\left(\frac{\sqrt{nRL}}{\sqrt{T}} + \frac{\sigma}{T^{3/2}} + \frac{\sigma}{\sqrt{T}}\right).$$

If we take β as an arbitrary constant in $(0,1)$ and take $B = T^\alpha$ where $\alpha \in (0,1)$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\| \leq \mathcal{O}\left(\frac{\sqrt{nRL}}{\sqrt{T}} + \frac{\sigma}{T^{\alpha/2+1}} + \frac{\sigma}{T^{\alpha/2}}\right).$$

The proof is completed. □

3 Convergence of the generic scheme in Bernstein and Newhouse (2024)

Now we analyze the convergence of (1.3), where we actually analyze the following heavy-ball and mini-batch scheme:

$$\begin{aligned} G_t &\leftarrow \frac{1}{B}\sum_{i=1}^{B}\nabla F\left(X_t,\xi_{t,i}\right) \\ B_t &\leftarrow \beta B_{t-1}+(1-\beta)G_t \\ \Delta_t &\leftarrow \operatorname*{argmin}_{\Delta}\{\operatorname{tr}(B_t^\top\Delta)+\frac{1}{2\eta}\|\Delta\|^2\} \\ X_{t+1} &\leftarrow X_t+\Delta_t \end{aligned} \tag{3.1}$$

where $\|\cdot\|$ denotes the spectral norm, and $\|\cdot\|_*$ is the dual norm, i.e. nuclear norm. Such notations carry on in this section.

We need to assume the Lipschitzness of the gradient, i.e.

$$\|\nabla f(X)-\nabla f(Y)\|_*\le L\|X-Y\| \tag{3.2}$$

which implies (Beck, 2017, Lemma 5.7)

$$f(Y)\le f(X)+\langle\nabla f(X),Y-X\rangle+\frac{L}{2}\|X-Y\|^2. \tag{3.3}$$

Here the inner product is just the Frobenius inner product $\langle X,Y\rangle=\operatorname{tr}(X^\top Y)$. Another assumption is that $\nabla F(X,\xi)$ is an unbiased estimator of $\nabla f(X)$ with variance bounded by σ^2, i.e.

$$\mathbb{E}\|\nabla F(X,\xi)-\nabla f(X)\|_*^2\le\sigma^2. \tag{3.4}$$

We have the following result:

Theorem 3.1. *Let $R=f(X_1)-f^*$. If we take $\eta\le\frac{1}{8L}\sqrt{\frac{1-2\beta}{2\beta}}$ and β as any constant in $(0,1)$, then for update (3.1) we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2\le\mathcal{O}\left(\frac{R}{T}+\frac{2n\sigma^2}{B}\right).$$

Proof. We first inspect the third line of (3.1). We know that (Bernstein and Newhouse, 2024, Proposition 1)

$$\operatorname*{argmin}_{\Delta}\{\operatorname{tr}(B^\top\Delta)+\frac{1}{2\eta}\|\Delta\|_2^2\}=-\eta\|B\|_*\operatorname*{argmax}_{\|A\|=1}\langle A,B\rangle$$

where $A^*:=\operatorname{argmax}_{\|A\|=1}\langle A,B\rangle=UV^\top$ is exactly the matrix such that $\langle A^*,B\rangle=\|B\|_*$, where $B=U\Sigma V^\top$ is the SVD. Therefore we have the following useful results:

$$\begin{aligned} \langle\Delta_t,B_t\rangle&=-\eta\|B_t\|_*^2 \\ \|\Delta_t\|&=\eta\|B_t\|_* \end{aligned} \tag{3.5}$$

Now we analyze $B_t-\nabla f(X_t)$, by convexity

$$\begin{aligned} &\|B_t-\nabla f(X_t)\|_*^2 \\ =&\|(1-\beta)(G_t-\nabla f(X_t))+\beta(B_{t-1}-\nabla f(X_t))\|_*^2 \\ \le&(1-\beta)\|G_t-\nabla f(X_t)\|_*^2+\beta\|B_{t-1}-\nabla f(X_t)\|_*^2 \\ \le&(1-\beta)\|G_t-\nabla f(X_t)\|_*^2+2\beta\|B_{t-1}-\nabla f(X_{t-1})\|_*^2+2\beta\|\nabla f(X_t)-\nabla f(X_{t-1})\|_*^2. \end{aligned} \tag{3.6}$$

Now if we only take expectation w.r.t ξ_t (random variables of iteration t) we get

$$\begin{aligned}\mathbb{E}\|G_t-\nabla f(X_t)\|_*^2 \leq & n\mathbb{E}\|G_t-\nabla f(X_t)\|_F^2\\ =& n\mathbb{E}\|\frac{1}{B}\sum_{i=1}^{B}\left(\nabla F(X_t,\xi_{t,i}-\nabla f(X_t)\right)\|_F^2\\ \overset{(i)}{=}& \frac{n}{B}\mathbb{E}\|\left(\nabla F(X_t,\xi_{t,1}-\nabla f(X_t)\right)\|_F^2\\ \leq & \frac{n}{B}\mathbb{E}\|\left(\nabla F(X_t,\xi_{t,1}-\nabla f(X_t)\right)\|_*^2\leq \frac{n}{B}\sigma^2,\end{aligned} \tag{3.7}$$

where (i) is due to the independency of the samples $\xi_{t,i}$.

Therefore

$$\begin{aligned}&\mathbb{E}\|B_t-\nabla f(X_t)\|_*^2\\ \leq&(1-\beta)\frac{n\sigma^2}{B}+(1+\gamma)\beta\|B_{t-1}-\nabla f(X_{t-1})\|_*^2+(1+\frac{1}{\gamma})\beta L^2\|X_t-X_{t-1}\|^2\\ =&(1-\beta)\frac{n\sigma^2}{B}+(1+\gamma)\beta\|B_{t-1}-\nabla f(X_{t-1})\|_*^2+(1+\frac{1}{\gamma})\beta L^2\|\Delta_t\|^2\\ \leq&(1+\gamma)\beta\|B_{t-1}-\nabla f(X_{t-1})\|_*^2+(1-\beta)\frac{n\sigma^2}{B}+(1+\frac{1}{\gamma})\beta L^2\eta^2\|B_t\|_*^2,\end{aligned}$$

where $\gamma>0$ is a constant to be determined.

Therefore we have

$$\begin{aligned}&(1-(1+\gamma)\beta)\sum_{t=1}^{T}\mathbb{E}\|B_t-\nabla f(X_t)\|_*^2\\ &\qquad\leq(1-\beta)\frac{n\sigma^2}{B}T+(1+1/\gamma)\beta L^2\eta^2\sum_{t=1}^{T}\mathbb{E}\|B_t\|_*^2.\end{aligned}$$

Now since

$$\sum_{t=1}^{T}\mathbb{E}\|B_t\|_*^2\leq 2\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2+2\sum_{t=1}^{T}\mathbb{E}\|B_t-\nabla f(X_t)\|_*^2$$

we have

$$\begin{aligned}&(1-(1+\gamma)\beta-2(1+1/\gamma)\beta L^2\eta^2)\sum_{t=1}^{T}\mathbb{E}\|B_t-\nabla f(X_t)\|_*^2\\ \leq&(1-\beta)\frac{n\sigma^2}{B}T+2(1+1/\gamma)\beta L^2\eta^2\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2.\end{aligned} \tag{3.8}$$

Now by (3.3) we have

$$\begin{aligned}f(X_{t+1})\leq& f(X_t)+\langle\nabla f(X_t),\Delta_t\rangle+\frac{L}{2}\|\Delta_t\|^2\\ \overset{(3.5)}{=}&f(X_t)-\eta\|B_t\|_*^2+\frac{\eta^2L}{2}\|B_t\|_*^2+\langle\nabla f(X_t)-B_t,\Delta_t\rangle\\ \leq&f(X_t)-(\eta-\frac{\eta^2L}{2})\|B_t\|_*^2+\frac{c}{2}\|\nabla f(X_t)-B_t\|_*^2+\frac{1}{2c}\eta^2\|B_t\|_*^2\end{aligned}$$

so we get

$$(\eta - \frac{\eta^2 L}{2} - \frac{\eta^2}{2c})\|B_t\|_*^2 \leq f(X_t) - f(X_{t+1}) + \frac{c}{2}\|\nabla f(X_t) - B_t\|_*^2.$$

Therefore the following holds:

$$\begin{aligned}
&(\eta - \frac{\eta^2 L}{2} - \frac{\eta^2}{2c})\|\nabla f(X_t)\|_*^2 \\
\leq& 2(\eta - \frac{\eta^2 L}{2} - \frac{\eta^2}{2c})\|B_t\|_*^2 + 2(\eta - \frac{\eta^2 L}{2} - \frac{\eta^2}{2c})\|\nabla f(X_t) - B_t\|_*^2 \\
\leq& 2(f(X_t) - f(X_{t+1})) + \left(c + (2\eta - \eta^2 L - \frac{\eta^2}{c})\right)\|\nabla f(X_t) - B_t\|_*^2.
\end{aligned}$$

Taking $c = \eta$ we get

$$(\frac{\eta}{2} - \frac{\eta^2 L}{2})\|\nabla f(X_t)\|_*^2 \leq 2(f(X_t) - f(X_{t+1})) + \left(2\eta - \eta^2 L\right)\|\nabla f(X_t) - B_t\|_*^2. \tag{3.9}$$

Now sum up above inequality we get

$$\begin{aligned}
&(\frac{\eta}{2} - \frac{\eta^2 L}{2})\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2 \\
\leq& 2(f(X_0) - f^*) + \left(2\eta - \eta^2 L\right)\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t) - B_t\|_*^2.
\end{aligned}$$

Now apply (3.8) to above equation we get

$$\begin{aligned}
&(\frac{\eta}{2} - \frac{\eta^2 L}{2})\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2 \\
\leq& 2(f(X_0) - f^*) + \frac{2\eta - \eta^2 L}{(1 - (1+\gamma)\beta - 2(1+1/\gamma)\beta L^2\eta^2)}\left((1-\beta)\frac{n\sigma^2}{m}T + 2(1+1/\gamma)\beta L^2\eta^2\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2\right).
\end{aligned}$$

That is, the following holds:

$$\begin{aligned}
&\left(\frac{\eta}{2} - \frac{\eta^2 L}{2} - \frac{2(2\eta - \eta^2 L)(1+1/\gamma)\beta L^2\eta^2}{(1 - (1+\gamma)\beta - 2(1+1/\gamma)\beta L^2\eta^2)}\right)\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2 \\
\leq& 2(f(X_0) - f^*) + \frac{(1-\beta)(2\eta - \eta^2 L)}{(1 - (1+\gamma)\beta - 2(1+1/\gamma)\beta L^2\eta^2)}\frac{n\sigma^2}{m}T.
\end{aligned}$$

Taking $\gamma \leq 1/(2\beta) - 1$ so that $1 - (1+\gamma)\beta \geq 1/2$, also taking $\eta \leq \frac{1}{8L}\sqrt{\frac{1-2\beta}{2\beta}}$ and β as any constant in $(0, 1)$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2 \leq \mathcal{O}\left(\frac{R}{T} + \frac{2n\sigma^2}{B}\right).$$

This completes the proof. □

Remark 3.1. *Note that we are not able to achieve a batch free convergence for (3.1). The difficulty stems from (3.6), where we used convexity of the quadratic and norm functions. A better bound is available in Ghadimi et al. (2020, Theorem 3.5) yet it only works for vector norm to our knowledge.*

Remark 3.2. *If we replace the norm in this section by an arbitrary norm, (3.7) no longer holds. In stead we can only say* $\mathbb{E}\|G_t - \nabla f(X_t)\|_*^2 \leq \sigma^2$ *and the final bound becomes*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(X_t)\|_*^2 \leq \mathcal{O}\left(\frac{R}{T} + \sigma^2\right).$$

In this case there is an unavoidable constant variance term even with mini-batch updates.

Acknowledgment

JL and MH thank Jeremy Bernstein, also Zhiqi Bu for helpful discussions.

References

A. Beck. *First-order methods in optimization*. SIAM, 2017. (Cited on pages 2 and 6.)

J. Bernstein and L. Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024. (Cited on pages 1 and 6.)

A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020. (Cited on pages 2 and 4.)

S. Ghadimi, A. Ruszczynski, and M. Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020. (Cited on page 8.)