Understanding and Detecting Convergence for Stochastic Gradient Descent with Momentum

Supplementary material

A Section 2

We cite Yang et al. (2018) in Theorem 2.1 because their convergence rate consisting of a reducible and irreducible term with respect to the number of updates, and best matches our empirical observations. Though Loizou and Richtarik (2017) show a linear convergence rate with constant stepsize, their restriction on the momentum (β) makes their convergence rate difficult to realize in practice. For example, using the formula in their paper with min and max eigenvalues=0.5 and stepsize=0.1, $\beta < 0.2$, which is very restrictive.

B Section 3

Assumption 1. *The expected loss $f(\theta) = \mathbb{E}[\ell(\theta, \xi)]$ is strongly convex with constant c.*

Assumption 2. *The expected loss $f(\theta) = \mathbb{E}[\ell(\theta, \xi)]$ is Lipschitz-smooth with constant L.*

Assumption 3. *Theorem 2.1 (Yang et al., 2018) holds s.t. $\mathbb{E}[f(\theta_n) - f(\theta_\star)] \leq \gamma M$ for some $M > 0$ and large enough n.*

Assumption 4. *$\exists\ \sigma_0^2 > 0$ s.t. $\mathbb{E}[\|\nabla \ell(\theta, \xi)\|^2] > \sigma_0^2$.*

Assumption 5. *$\exists\ K > 1$ s.t. $\mathbb{E}[(\theta_n - \theta_{n-1})^\top(\theta_{n-1} - \theta_{n-2})] \geq -K\mathbb{E}[\|\theta_n - \theta_{n-1}\|^2]$ for large enough n.*

We first derive a lower bound on the expected distance between iterates.

Lemma B.1. *Suppose that Assumptions 4 and 5 hold. Then for large enough n,*

$$\mathbb{E}[\|\theta_n - \theta_{n-1}\|^2] \geq \gamma^2\sigma_0^2\left(\frac{1}{1 + 2K\beta + \beta^2}\right) \tag{1}$$

Proof. By re-arranging the update equation for SGDM in Equation (3)we get $\theta_n - \theta_{n-1} - \beta(\theta_{n-1} - \theta_{n-2}) = -\gamma\nabla\ell(\theta_{n-1}, \xi_n)$. For brevity, let $\Delta_n \equiv \theta_n - \theta_{n-1}$, and $\nabla\ell_n \equiv \nabla\ell(\theta_{n-1}, \xi_n)$. Applying squared norm and re-arranging terms,

$$\begin{aligned} \|\Delta_n - \beta\Delta_{n-1}\|^2 &= \| - \gamma\nabla\ell_n\|^2 \\ \|\Delta_n\|^2 &= \gamma^2\|\nabla\ell_n\|^2 + 2\beta\Delta_n^\top\Delta_{n-1} - \beta^2\|\Delta_{n-1}\|^2 \end{aligned}$$

Apply expectations to both sides and the second inequality assumption to define a recursive relation.

$$\begin{aligned}\mathbb{E}[\|\Delta_n\|^2] &= \gamma^2\mathbb{E}[\|\nabla\ell_n\|^2] + 2\beta\mathbb{E}[\Delta_n^\top\Delta_{n-1}] - \beta^2\mathbb{E}[\|\Delta_{n-1}\|^2] \\ &\geq \gamma^2\sigma_0^2 - \mathbb{E}[\|\Delta_{n-1}\|^2](2\beta K + \beta^2)\end{aligned}$$

Using this recursive relation we get

$$\begin{aligned}\mathbb{E}[\|\Delta_n\|^2] &\geq \gamma^2\sum_{i=0}^{n-1}(-1)^i(2\beta K + \beta^2)^i\sigma_0^2 \\ &= \gamma^2\sigma_0^2\left(\frac{1-(2\beta K+\beta^2)^n}{1+2\beta K+\beta^2}\right)\end{aligned}$$

Because we expect the number of iterations n to be large when entering the stationary phase, we make the approximation

$$\mathbb{E}[\|\Delta_n\|^2] \geq \gamma^2\sigma_0^2\left(\frac{1}{1+2\beta K+\beta^2}\right)$$

□

Remarks. If $\mathbb{E}[\Delta_n^\top\Delta_{n-1}] > 0$ then any negative lower bound is trivial. It is reasonable to assume K is not too large. We expect $\|\theta_n - \theta_{n-1}\|^2 \approx \|\theta_{n-1} - \theta_{n-2}\|^2$ because they are successive iterates.

Theorem 3.1. *Suppose that Assumptions 1, 3, and 4 hold. The test statistic for the convergence diagnostic in Algorithm 1for SGDM in Equation (3) is bounded*

$$\mathbb{E}[\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)] \leq (1+\beta)\left[M - \frac{c}{2}\gamma\sigma_0^2A_\beta\right] < 0$$

as $n\to\infty$. And thus the convergence diagnostic activates almost surely.

Proof. First re-write the inner product with the decomposition in Equation (6).

$$\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n) = \frac{1}{\gamma}\nabla\ell(\theta_n,\xi_{n+1})^\top(\theta_{n-1}-\theta_n) + \frac{\beta}{\gamma}\nabla\ell(\theta_n,\xi_{n+1})^\top(\theta_{n-1}-\theta_{n-2})$$

Apply expectation to both sides, and then apply the strong convexity assumption.

$$\begin{aligned}\mathbb{E}[\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)] &\leq \frac{1}{\gamma}[f(\theta_{n-1}) - f(\theta_n) - \frac{c}{2}\|\theta_{n-1}-\theta_n\|^2] \\ &\quad + \frac{\beta}{\gamma}[f(\theta_{n-1}-\theta_{n-2}+\theta_n) - f(\theta_n) - \frac{c}{2}\|\theta_{n-1}-\theta_{n-2}\|^2] \\ &\leq \frac{1}{\gamma}[f(\theta_{n-1}) - f(\theta_\star) - \frac{c}{2}\|\theta_{n-1}-\theta_n\|^2] \\ &\quad + \frac{\beta}{\gamma}[f(\theta_{n-1}-\theta_{n-2}+\theta_n) - f(\theta_\star) - \frac{c}{2}\|\theta_{n-1}-\theta_{n-2}\|^2]\end{aligned}$$

Now extend the expectation over randomness in the trajectory, and apply the result from Theorem 2.1and Lemma B.1.

$$\begin{aligned}&\leq \frac{1}{\gamma}\left[\gamma M - \frac{c}{2}\gamma^2\sigma_0^2A_\beta\right] + \frac{\beta}{\gamma}\left[\gamma M - \frac{c}{2}\gamma^2\sigma_0^2A_\beta\right] \\ &= (1+\beta)[M - \frac{c}{2}\gamma\sigma_0^2A_\beta]\end{aligned}$$

□

Remarks. $A_\beta = 1/(1+2\beta K+\beta^2)$ from the previous Lemma B.1. A_β is a monotonically decreasing function, where $A_{\beta|\beta=0} = 1$, and $A_{\beta|\beta=1} = 1/(2+2K)$. Thus for the condition on the learning rate, higher β causes lower A_β which increases the condition on γ^2.

B.1 An alternative quantity to monitor stationarity

We now derive bounds similar to Theorem 3.1 for the alternative test statistic constructed in Equation (5) , and show that its expectation is highly sensitive to momentum. This increased sensitivity makes it a poor choice to use in a convergence diagnostic.

Theorem 3.2. *Suppose that Assumptions 1, 3, and 4 hold. Then,*

$$\begin{aligned}\mathbb{E}[(\nabla\ell(\theta_n,\xi_{n+1}) + \beta(\theta_n-\theta_{n-1}))^\top(\nabla\ell(\theta_{n-1},\xi_n)+\beta(\theta_{n-1}-\theta_{n-2}))]&\\ < \left(\frac{1}{\gamma}+\frac{\beta}{\gamma}+2\beta+\beta^2\right)\left[\gamma M - \frac{c}{2}\gamma^2\sigma_0^2 A_\beta\right] + \beta^3\gamma M&\end{aligned}$$

Proof. Decompose the terms in the inner product, apply expectation to both sides, and apply the strong convexity assumption.

$$\begin{aligned}&\mathbb{E}[(\nabla\ell(\theta_n,\xi_{n+1}) + \beta(\theta_n-\theta_{n-1}))^\top(\nabla\ell(\theta_{n-1},\xi_n)+\beta(\theta_{n-1}-\theta_{n-2}))]\\
&\quad= \mathbb{E}[\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)] + \mathbb{E}[\beta\nabla\ell(\theta_n,\xi_{n+1})^\top(\theta_{n-1}-\theta_{n-2})]\\
&\qquad+\mathbb{E}[\beta\nabla\ell(\theta_{n-1},\xi_n)^\top(\theta_n-\theta_{n-1})] + \mathbb{E}[\beta^2(\theta_{n-1}-\theta_{n-2})^\top(\theta_n-\theta_{n-1})]\\
&\quad\le \frac{1}{\gamma}\left[f(\theta_{n-1}) - f(\theta_n) - \frac{c}{2}\|\theta_{n-1}-\theta_n\|^2\right]\\
&\qquad+\frac{\beta}{\gamma}\left[f(\theta_{n-1}-\theta_{n-2}+\theta_n) - f(\theta_n) - \frac{c}{2}\|\theta_{n-1}-\theta_{n-2}\|^2\right]\\
&\qquad+\beta\left[f(\theta_{n-1}-\theta_{n-2}+\theta_n) - f(\theta_n) - \frac{c}{2}\|\theta_{n-1}-\theta_{n-2}\|^2\right]\\
&\qquad+\beta\left[f(\theta_{n-1}) - f(\theta_n) - \frac{c}{2}\|\theta_{n-1}-\theta_n\|^2\right]\\
&\qquad+\beta^2\left[\beta\|\theta_{n-1}-\theta_{n-2}\|^2 + f(\theta_{n-2}) - f(\theta_{n-1}) - \frac{c}{2}\|\theta_{n-1}-\theta_{n-2}\|^2\right]\end{aligned}$$

Where the last line uses $\theta_n - \theta_{n-1} = -\nabla\ell(\theta_{n-1},\xi_n + \beta(\theta_{n-1}-\theta_{n-2})$. Take expectation with respect to the trajectory, and apply Theorem 2.1, Lemma B.1, and Lemma C.1.

$$\le \left(\frac{1}{\gamma}+\frac{\beta}{\gamma}+2\beta+\beta^2\right)\left[\gamma M - \frac{c}{2}\gamma^2\sigma_0^2 A_\beta\right] + \beta^3\frac{\gamma M}{c}$$

□

Remarks. The expectation in Theorem 3.2 is much more sensitive to the momentum parameter β, as seen in the additional β terms. The upper bound assumption on $\mathbb{E}[\|\theta_n - \theta_{n-1}\|^2]$ is not restrictive since we have a bound on $\mathbb{E}[f(\theta_n) - f(\theta_\star)]$ from Theorem 2.1.

B.2 Quadratic Loss Model

Let's gain insight into the convergence diagnostic by first assuming quadratic loss $\ell(\theta,y,x) = \frac{1}{2}(y-x^\top\theta_\star)^2$, $\nabla\ell(\theta,y,x) = -(y-x^\top\theta)x$. Let $y = x^\top\theta_\star + \epsilon$, where ϵ are zero mean random variables

$\mathbb{E}[\epsilon|x] = 0$. If we are able to initialize $\theta_0 = \theta_\star$, i.e. within the stationary region, then the update equations are:

$$\begin{aligned}
\theta_1 &= \theta_\star + \gamma(y_1 - x_1^\top \theta_\star)x_1 \\
\theta_2 &= \theta_1 + \gamma(y_2 - x_2^\top \theta_1)x_2 + \beta(\theta_1 - \theta_\star) \\
\theta_3 &= \theta_2 + \gamma(y_3 - x_3^\top \theta_2)x_3 + \beta(\theta_2 - \theta_1)
\end{aligned}$$

With the gradients reducing to:

$$\begin{aligned}
\nabla\ell(\theta_\star, y_1, x_1) &= (y_1 - x_1^\top \theta_\star)x_1 \\
&= \epsilon_1 x_1 \\
\nabla\ell(\theta_1, y_2, x_2) &= (y_2 - x_2^\top \theta_1)x_2 \\
&= (y_2 - x_2^\top \theta_\star - \gamma(y_1 - x_1^\top \theta_\star)x_2^\top x_1)x_2 \\
&= (\epsilon_2 - \gamma\epsilon_1 x_2^\top x_1)x_2 \\
\nabla\ell(\theta_2, y_3, x_3) &= (y_3 - x_3^\top \theta_2)x_3 \\
&= (y_3 - x_3^\top \theta_1 - \gamma(y_2 - x_2^\top \theta_1)x_3^\top x_2 - \beta x_3^\top(\theta_1 - \theta_\star))x_3 \\
&= (y_3 - x_3^\top \theta_\star - \gamma(y_1 - x_1^\top \theta_\star)x_3^\top x_1 \\
&\quad - \gamma(y_2 - x_2^\top \theta_\star - \gamma(y_1 - x_1^\top \theta_\star)x_2^\top x_1)x_3^\top x_2 \\
&\quad - \beta x_3^\top(\gamma(y_1 - x_1^\top \theta_\star)x_1))x_3 \\
&= (\epsilon_3 - \gamma\epsilon_1 x_3^\top x_1 - \gamma\epsilon_2 x_3^\top x_2 + \gamma^2\epsilon_1(x_2^\top x_1)(x_3^\top x_2) - \beta\gamma\epsilon_1 x_3^\top x_1)x_3
\end{aligned}$$

The value of the expected test statistics is:

$$\begin{aligned}
\mathbb{E}[\nabla\ell(\theta_1, y_2, x_2)^\top \nabla\ell(\theta_2, y_3, x_3)] &= \mathbb{E}[-\gamma\epsilon_2^2(x_3^\top x_2)^2 + \gamma^2\epsilon_1^2(x_2^\top x_1)(x_3^\top x_1)(x_3^\top x_2) \\
&\qquad - \gamma^3\epsilon_1^2(x_2^\top x_1)^2(x_3^\top x_2)^2 + \beta\gamma^2\epsilon_1^2(x_2^\top x_1)(x_3^\top x_1)(x_3^\top x_2)] \\
&= -\gamma\mathbb{E}[\epsilon_2^2]\mathbb{E}[(x_3^\top x_2)^2] - \gamma^3\mathbb{E}[\epsilon_1^2]\mathbb{E}[(x_2^\top x_1)^2(x_3^\top x_2)^2] \\
&\quad + \gamma^2(1+\beta)\mathbb{E}[\epsilon_1^2]\mathbb{E}[(x_2^\top x_1)(x_3^\top x_1)(x_3^\top x_2)]
\end{aligned}$$

The following theorem gives us an idea how the test statistic evolves as the optimization procedure moves through parameter space.

Theorem 3.4. *Let the loss be quadratic, $\ell(\theta) = 1/2(y - x^\top\theta)^2$. Let x_n, x_{n+1} be two iid vectors from the distribution of x. Let $A = \mathbb{E}[(x_n x_{n+1}^\top)(x_n^\top x_{n+1})]$, $B = \mathbb{E}[(x_n x_n^\top)(x_n^\top x_{n+1})^2]$, $\sigma_{quad}^2 = \mathbb{E}[\epsilon_n^2]$, $d^2 = \mathbb{E}[(x_n^\top x_{n+1})^2]$. Then,*

$$\begin{aligned}
\mathbb{E}[\nabla\ell(\theta_n, \xi_{n+1})^\top \nabla\ell(\theta_{n-1}, \xi_n)|\theta_{n-1}, \theta_{n-2}] &= (\theta_{n-1} - \theta_\star)^\top(A - \gamma B)(\theta_{n-1} - \theta_\star) - \gamma\sigma_{quad}^2 d^2 \\
&\quad + (\theta_{n-1} - \theta_\star)^\top(\beta A)(\theta_{n-1} - \theta_{n-2})
\end{aligned}$$

Proof. Let $\nabla\ell(\theta_n) \equiv \nabla\ell(\theta_{n-1}, \xi_n)$. The inner product is

$$\begin{aligned}
\nabla\ell(\theta_n)^\top \nabla\ell(\theta_{n-1}) &= \left(y_{n+1} - x_{n+1}^\top\theta_n\right)\left(y_n - x_n^\top\theta_{n-1}\right)x_n^\top x_{n+1} \\
&= \Big[y_{n+1} - x_{n+1}^\top\theta_{n-1} - \gamma(y_n - x_n^\top\theta_{n-1})x_n^\top x_{n+1}
\end{aligned}$$

$$- \beta x_{n+1}^\top(\theta_{n-1} - \theta_{n-2})\Big] \Big(y_n - x_n^\top \theta_{n-1}\Big)\, x_n^\top x_{n+1}$$

Now distribute the second gradient term and trailing x_n, x_{n+1}

$$\begin{aligned}
&= \Big(y_{n+1} - x_{n+1}^\top \theta_{n-1}\Big)\Big(y_n - x_n^\top\theta_{n-1}\Big)\, x_n^\top x_{n+1} \\
&\quad - \gamma \Big(y_n - x_n^\top \theta_{n-1}\Big)^2 \Big(x_n^\top x_{n+1}\Big)^2 \\
&\quad - \beta \Big(y_n - x_n^\top\theta_{n-1}\Big)\, [x_{n+1}^\top(\theta_{n-1} - \theta_{n-2})] \Big(x_n^\top x_{n+1}\Big)
\end{aligned}$$

And now substitute for ϵ

$$\begin{aligned}
&= \Big[x_{n+1}^\top(\theta_\star - \theta_{n-1}) + \epsilon_{n+1}\Big]\Big[x_n^\top(\theta_\star - \theta_{n-1}) + \epsilon_n\Big] x_n^\top x_{n+1} \\
&\quad - \gamma \Big[x_n^\top(\theta_\star - \theta_{n-1}) + \epsilon_n\Big]^2 \Big(x_n^\top x_{n+1}\Big)^2 \\
&\quad - \beta \Big[x_n^\top(\theta_\star - \theta_{n-1}) + \epsilon_n\Big]\Big[x_{n+1}^\top(\theta_{n-1} - \theta_{n-2})\Big] \Big(x_n^\top x_{n+1}\Big)
\end{aligned}$$

Expand terms

$$\begin{aligned}
&= (\theta_{n-1} - \theta_\star)^\top \Big[(x_n x_{n+1}^\top)(x_n^\top x_{n+1})\Big](\theta_{n-1} - \theta_\star) + \epsilon_n W^{(1)} + \epsilon_{n+1} W^{(2)} + \epsilon_n \epsilon_{n+1} W^{(3)} \\
&\quad - (\theta_{n-1} - \theta_\star)^\top \Big[\gamma (x_n x_n^\top)(x_n^\top x_{n+1})^2\Big](\theta_{n-1} - \theta_\star) - \epsilon_n W^{(4)} - \gamma \epsilon_n^2 (x_n^\top x_{n+1})^2 \\
&\quad + (\theta_{n-1} - \theta_\star)^\top \Big[\beta (x_n x_{n+1}^\top)(x_n^\top x_{n+1})\Big](\theta_{n-1} - \theta_{n-2}) - \epsilon_n W^{(5)}
\end{aligned}$$

Take expectation with respect to $n-1, n-2$. Conditioning with respect to two terms is necessary to capture the momentum. The non square epsilon terms are eliminated because they are mean zero. □

C Section 4

We first give some technical Lemmas necessary to prove the main variance bound results.

Lemma C.1. *Suppose that Assumptions 1 and 3 hold for some positive $M' > 0$. Let $M = 8M'$. Then we can bound the L2 distance iterates,*

$$\mathbb{E}[\|\theta_n - \theta_{n-1}\|^2] \le \frac{\gamma M}{c}$$

Proof. Theorem 2.1presents Theorem 1 from Yang et al. (2018). We see that for large enough n, this is a valid condition. Under the strong convexity assumption, we can bound

$$\begin{aligned}
\frac{c}{2}\|\theta_n - \theta_\star\|^2 &\le f(\theta_n) - f(\theta_\star) - \nabla f(\theta_\star)^\top(\theta_n - \theta_\star) \\
&= f(\theta_n) - f(\theta_\star)
\end{aligned}$$

Applying expectation to both sides,

$$\frac{c}{2}\,\mathbb{E}[\|\theta_n-\theta_\star\|^2]\le\mathbb{E}[f(\theta_n)-f(\theta_\star)]$$

We now use the triangle inequality to obtain the desired bound.

$$\begin{aligned}\|\theta_n-\theta_{n-1}\|&\le\|\theta_n-\theta_\star\|+\|\theta_{n-1}-\theta_\star\|\\\|\theta_n-\theta_{n-1}\|^2&\le\|\theta_n-\theta_\star\|^2+\|\theta_{n-1}-\theta_\star\|^2+2\|\theta_n-\theta_\star\|\|\theta_{n-1}-\theta_\star\|\end{aligned}$$

Apply expectation to both sides,

$$\begin{aligned}\mathbb{E}[\|\theta_n-\theta_{n-1}\|^2]&\le\mathbb{E}[\|\theta_n-\theta_\star\|^2]+\mathbb{E}[\|\theta_{n-1}-\theta_\star\|^2]+2\mathbb{E}[\|\theta_n-\theta_\star\|]\mathbb{E}[\|\theta_{n-1}-\theta_\star\|]\\&\le\mathbb{E}[\|\theta_n-\theta_\star\|^2]+\mathbb{E}[\|\theta_{n-1}-\theta_\star\|^2]+2\sqrt{\mathbb{E}[\|\theta_n-\theta_\star\|^2]\mathbb{E}[\|\theta_{n-1}-\theta_\star\|^2]}\\&\le 8\frac{\gamma M'}{c}\end{aligned}$$

In the stationary phase the variance of the stochastic gradient dominates and no more progress is made towards $\theta_\star$, so then $\|\theta_n-\theta_\star\|^2$ and $\|\theta_{n-1}-\theta_\star\|^2$ are independent in the stationary phase. The second inequality is by Jensen's. □

Lemma C.2. *Consider the SGDM procedure in Equation (3). Suppose that Assumptions 2 and 3 hold. Then we can bound*

$$\mathbb{E}\left[\left(\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)\right)^2\right]\ge(1+2\beta+\beta^2)\left[M'-\frac{L}{2}\gamma\sigma_0^2A_\beta\right]^2$$

Proof. Apply expectation with respect to ξ_{n+1}, Jensen's inequality, and the decomposition in Equation (6).

$$\begin{aligned}\mathbb{E}[(\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n))^2]&\ge\mathbb{E}[(\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n))]^2\\&=\left(\mathbb{E}[\frac{1}{\gamma}\nabla\ell(\theta_n,\xi_{n+1})^\top(\theta_{n-1}-\theta_n)]+\mathbb{E}[\frac{\beta}{\gamma}\nabla\ell(\theta_n,\xi_{n+1})^\top(\theta_{n-1}-\theta_{n-2})]\right)^2\end{aligned}$$

Define $f(\theta)\equiv\mathbb{E}[\ell(\theta,\xi)]$. Apply the Lipschitz-smoothness assumption,
Now apply the Lipschitz smoothness condition.

$$\begin{aligned}&\ge\Bigg(\frac{1}{\gamma}[\ell(\theta_{n-1})-\ell(\theta_n)-\frac{L}{2}\|\theta_{n-1}-\theta_n\|^2]\\&\quad+\frac{\beta}{\gamma}[\ell(\theta_{n-1}-\theta_{n-2}+\theta_n)-\ell(\theta_n)-\frac{L}{2}\|\theta_{n-1}-\theta_{n-2}\|^2]\Bigg)^2\\&=\frac{1}{\gamma^2}\left[\ell(\theta_{n-1})-\ell(\theta_n)-\frac{L}{2}\|\theta_{n-1}-\theta_n\|^2\right]^2\\&\quad+2\frac{\beta}{\gamma^2}\left[\ell(\theta_{n-1})-\ell(\theta_n)-\frac{L}{2}\|\theta_{n-1}-\theta_n\|^2\right]\\&\qquad\left[\ell(\theta_{n-1}-\theta_{n-2}+\theta_n)-\ell(\theta_n)-\frac{L}{2}\|\theta_{n-1}-\theta_{n-2}\|^2\right]\\&\quad+\frac{\beta^2}{\gamma^2}\left[\ell(\theta_{n-1}-\theta_{n-2}+\theta_n)-\ell(\theta_n)-\frac{L}{2}\|\theta_{n-1}-\theta_{n-2}\|^2\right]^2\end{aligned}$$

For brevity of notation define $\Delta\ell_{n-1,n} \equiv \ell(\theta_{n-1}) - \ell(\theta_n)$.

$$
\begin{aligned}
&= \frac{1}{\gamma^2}\left[\Delta\ell_{n-1,n}^2 - 2\Delta\ell_{n-1,n}\frac{L}{2}\|\theta_{n-1}-\theta_n\|^2 + \frac{L^2}{4}\|\theta_{n-1}-\theta_n\|^4\right] \\
&\quad + 2\frac{\beta}{\gamma^2}\left[\Delta\ell_{n-1,n}\Delta\ell_{n-1n-2n,n} - \Delta\ell_{n-1n-2n,n}\frac{L}{2}\|\theta_{n-1}-\theta_n\|^2\right. \\
&\qquad \left. - \Delta\ell_{n-1,n}\frac{L}{2}\|\theta_{n-1}-\theta_{n-2}\|^2 + \frac{L^2}{4}\|\theta_{n-1}-\theta_n\|^2\|\theta_{n-1}-\theta_{n-2}\|^2\right] \\
&\quad + \frac{\beta^2}{\gamma^2}\left[\Delta\ell_{n-1n-2n,n}^2 - 2\Delta\ell_{n-1n-2n,n}\frac{L}{2}\|\theta_{n-1}-\theta_{n-2}\|^2 + \frac{L^2}{4}\|\theta_{n-1}-\theta_{n-2}\|^4\right]
\end{aligned}
$$

Now apply expectation with respect to the trajectory, Theorem 2.1, and Lemma B.1. Let $A_\beta = (\frac{1}{1+2K\beta+\beta^2})$.

$$
\begin{aligned}
&\geq \frac{1}{\gamma^2}\left[\gamma^2 M'^2 - M'L\gamma^3\sigma_0^2 A_\beta + \frac{L^2}{4}\gamma^4\sigma_0^4 A_\beta^2\right] \\
&\quad + 2\frac{\beta}{\gamma^2}\left[\gamma^2 M'^2 - M'L\gamma^3\sigma_0^2 A_\beta + \frac{L^2}{4}\gamma^4\sigma_0^4 A_\beta^2\right] \\
&\quad + \frac{\beta^2}{\gamma^2}\left[\gamma^2 M'^2 - M'L\gamma^3\sigma_0^2 A_\beta + \frac{L^2}{4}\gamma^4\sigma_0^4 A_\beta^2\right] \\
&= \left(\frac{1+2\beta+\beta^2}{\gamma^2}\right)\left[\gamma M' - \frac{L}{2}\gamma^2\sigma_0^2 A_\beta\right]^2
\end{aligned}
$$

To bound $\mathbb{E}[\Delta_{n-1,n}^2]$, $\mathbb{E}[\Delta_{n-1,n}^2] \geq \mathbb{E}[\Delta_{n-1,n}]^2$ by Jensen's inequality, and $\mathbb{E}[\Delta_{n-1,n}] \geq \mathbb{E}[\Delta_{\star,n}] \geq -\gamma M'$ through our assumption. To bound $\mathbb{E}[\Delta_{n-1,n}\|\theta_{n-1}-\theta_n\|^2]$, we first use $\Delta_{n-1,n} \geq \Delta_{\star,n}$. We then use the Cauchy-Schwarz and then Jensen's inequality to bound $\mathbb{E}[\Delta_{n-1,n}\|\theta_{n-1}-\theta_n\|^2] \geq \sqrt{\mathbb{E}[\Delta_{\star,n}^2]}\sqrt{\mathbb{E}[\|\theta_{n-1}-\theta_n\|^4]} \geq \mathbb{E}[\Delta_{\star,n}]\mathbb{E}[\|\theta_{n-1}-\theta_n\|^2] \geq -\gamma M'\gamma^2\sigma_0^2 A_\beta$, and use the Theorem 2.1 bound and Lemma B.1. We bound $\mathbb{E}[\|\theta_{n-1}-\theta_n\|^4] \geq \mathbb{E}[\|\theta_{n-1}-\theta_n\|^2]^2$ with Jensen's inequality, and can then use Lemma B.1. $\Delta_{n-1,n}$ and $\Delta_{n-1n-2n,n}$ are independent in the stationary region, and thus $\mathbb{E}[\Delta_{n-1n-2n,n}\Delta_n] \geq \gamma^2 M'^2$. To bound $\frac{L^2}{4}\|\theta_{n-1}-\theta_n\|^2\|\theta_{n-1}-\theta_{n-2}\|^2$, we first note that the two terms are positively correlated due to momentum. For two random variables X, Y with positive covariance, $\mathbb{E}[XY] = Cov(X,Y) + \mathbb{E}[X]\mathbb{E}[Y] \geq \mathbb{E}[X]\mathbb{E}[Y]$. Thus, $\mathbb{E}[\|\theta_{n-1}-\theta_n\|^2\|\theta_{n-1}-\theta_{n-2}\|^2] \geq \mathbb{E}[\|\theta_{n-1}-\theta_n\|^2]\mathbb{E}[\|\theta_{n-1}-\theta_{n-2}\|^2]$. □

Lemma C.3. *Consider the SGDM procedure in Equation (3). Suppose that Assumptions 2 and 3 hold. Then we can lower bound*

$$
\mathbb{E}[\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)] \geq -(1+\beta)M'(1+4L/c)
$$

Proof. First apply the decomposition in Equation (6).

$$
\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n) = \frac{1}{\gamma}\nabla\ell(\theta_n,\xi_{n+1})^\top(\theta_{n-1}-\theta_n) + \frac{\beta}{\gamma}\nabla\ell(\theta_n,\xi_{n+1})^\top(\theta_{n-1}-\theta_{n-2})
$$

Then apply expectation with respect to ξ_{n+1} and use the Lipschitz smoothness assumption to create the desired inequality

$$
\mathbb{E}[\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)] \geq \frac{1}{\gamma}[f(\theta_{n-1}) - f(\theta_n) - \frac{L}{2}\|\theta_{n-1}-\theta_n\|^2]
$$

$$+ \frac{\beta}{\gamma}[f(\theta_{n-1} - \theta_{n-2} + \theta_n) - f(\theta_n) - \frac{L}{2}\|\theta_{n-1} - \theta_{n-2}\|^2]$$

Now apply expectation with respect to the trajectory, Theorem 2.1, and Lemma C.1.

$$\begin{aligned}
&\geq \frac{1}{\gamma}[-\gamma M' - L\gamma 8M'/c] + \frac{\beta}{\gamma}[-\gamma M' - L\gamma 8M'/c] \\
&= -(1+\beta)M'(1 + 8L/c)
\end{aligned}$$

□

Theorem 4.2. *Consider the SGDM procedure in Equation (3). Suppose that Assumptions 1, 2, 3, 4, and 5 hold. Define* $IP = \nabla \ell(\theta_n, \xi_{n+1})^\top \nabla \ell(\theta_{n-1}, \xi_n)$*. Then,*

$$\frac{Var[IP]}{\mathbb{E}[IP]^2} \geq \frac{(M' - L\gamma\sigma_0^2 A_\beta)^2}{M'^2(1 + 8L/c)^2} - 1$$

Proof.

$$\begin{aligned}
\frac{Var[IP]}{\mathbb{E}[IP]^2} &= \frac{\mathbb{E}[IP^2] - \mathbb{E}[IP]^2}{\mathbb{E}[IP]^2} \\
&= \frac{\mathbb{E}[IP^2]}{\mathbb{E}[IP]^2} - 1
\end{aligned}$$

We have a lower bound on $\mathbb{E}[IP^2]$ from Lemma C.2, and an upper bound on $\mathbb{E}[IP]^2$ from Lemma C.3. We use Lemma C.3 and not the bound from Theorem 3.1because $|(1+\beta)M'(1 + 8L/c)| \geq |(1+\beta)[M' - \frac{c}{2}\gamma\sigma_0^2 A_\beta]|$.

$$\begin{aligned}
\frac{Var[IP]}{\mathbb{E}[IP]^2} &= \frac{\mathbb{E}[IP^2]}{\mathbb{E}[IP]^2} - 1 \\
&\geq \frac{(1+\beta)^2(M' - \frac{L}{2}\gamma\sigma_0^2 A_\beta)^2}{(1+\beta)^2 M'^2(1 + 8L/c)^2} - 1
\end{aligned}$$

□

Corollary 4.3. *Consider the SGDM procedure in Equation (3). Fix a scaling factor* $\lambda > 2$*. Set the learning rate* $\gamma = 2tM'/L\sigma_0^2 A_\beta$ *with* $t \geq 1 + \sqrt{\lambda}(1 + 4L/c)$*. Then the lower bound in Theorem 4.2 is bounded*

$$\frac{Var[IP]}{\mathbb{E}[IP]^2} \geq \lambda - 1$$

Proof. Setting $\gamma = 2tM'/L\sigma_0^2 A_\beta$, we see that

$$\frac{Var[IP]}{\mathbb{E}[IP]^2} \geq \frac{(1-t)^2}{(1 + 8L/c)^2} - 1.$$

The lower bound is obtained by solving the inequality,

$$(1-t)^2 \geq \lambda(1 + 8L/c)^2$$

$$t^2 - 2t + 1 - \lambda(1 + 8L/c)^2 \geq 0$$

We solve the quadratic inequality. Because it is convex, we find

$$\begin{aligned} t &\geq \left|\frac{2 \pm \sqrt{4 - 4(1 - \lambda(1 + 8L/c)^2)}}{2}\right| \\ &= |1 \pm (1 + 8L/c)\sqrt{\lambda}| \end{aligned}$$

Thus $t = 1 + \sqrt{\lambda}(1 + 8L/c)$ because we need $\gamma > 0$. □

D Additional Experiments

Figure 1: LR Reduction on MNIST

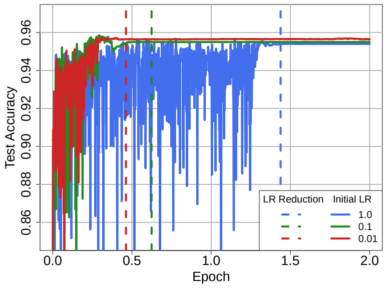


Figure 2: Momentum Config on MNIST

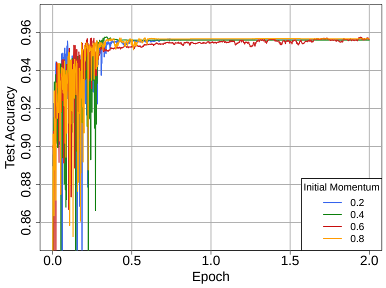


Figure 3: Test Acc for SGDM with lr=0.001 on MNSIT

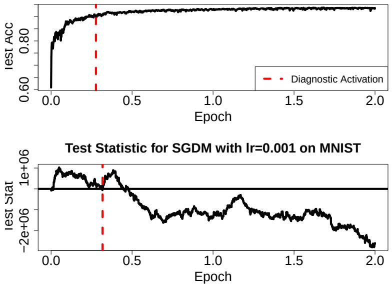

Test Statistic for SGDM with lr=0.001 on MNIST

Figure 4: Mom Reduce Close to Stationary MNIST

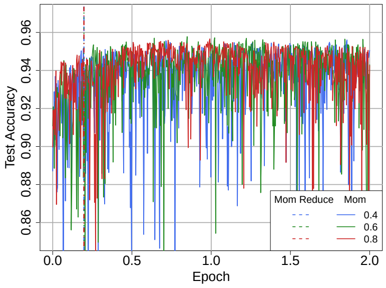


Figure 5: Constant SGDM on MNIST

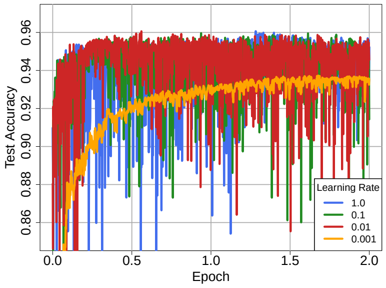
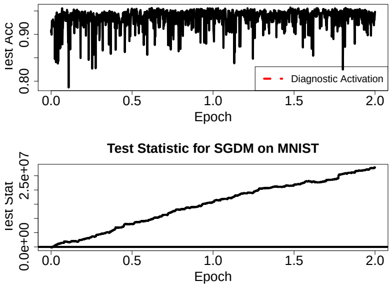
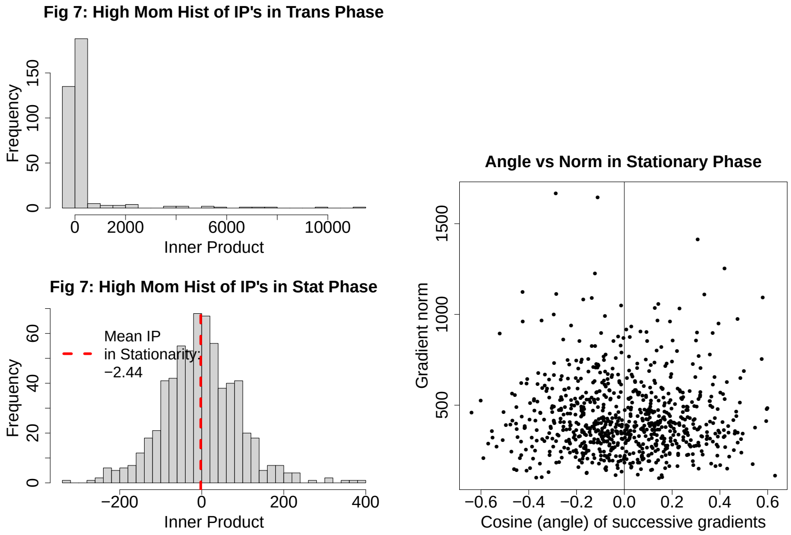
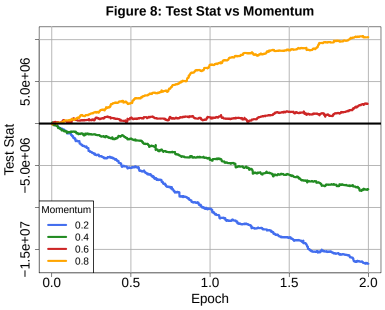


Fig 6: High Mom Test Acc & No Mom Red (MNSIT)


Test Statistic for SGDM on MNIST





The function $h()$ used in Algorithm 1 was the mean squared distance between successive iterates. However, any convergence heuristic can be used for $h()$ in Algorithm 1.

Figure 1 marks where the learning rate is reduced on the test accuracy vs epoch plot; each was reduced by a factor of 10 only once during training from the initial value. Figure 2 plots additional momentum values. Although there is variation in the transient phase the final test accuracy values are very similar. Figure 3 plots the proposed test statistic for SGDM with constant learning rate on MNIST. We see that the test statistic activates when the test accuracy has begun to flatten out, indicating the convergence diagnostic is a good indicator for stationarity. Figure 4 shows that the momentum is reduced close to the stationary phase. We plot the test accuracy of constant SGDM on MNIST and mark the momentum reduction with a vertical line. We see that momentum is reduced just before the test accuracy plateaus. Again, the learning rate has been tuned. This switch can be viewed as a rough initial estimate to stationarity, with the convergence diagnostic as a final, more accurate estimate. We found that varying the random seed had no effect, and we excluded a figure due to space constraints.

Comparison of adaptive step-size SGDM to decaying and constant stepsize. The application in Section 6 is for an automatic learning rate, which by definition should require little hand tuning. The comparison to decaying stepsize in Sec 6 was to show that, regardless of the reason for its worse performance, decaying stepsize requires far more hand tuning to get to work well. Figure 5 plots constant stepsize SGDM. With higher learning rate the maximum final test accuracy is comparable to Algorithm 2, however the range of fluctuation is quite large, about 10. Algorithm 2 is able to achieve as high test accuracy compared to constant SGDM, but with effectively no fluctuations in test accuracy.

Necessity of decreasing momentum for small batch sizes. We conducted an ablation study; by removing the momentum reduction component with a high momentum (0.8) on MNIST Algorithm 1 did not work at all. The convergence diagnostic never activated because the expectation of the test statistic was positive. Figure 6 plots the test accuracy and test statistic for this setup. We see from the cumulative sum of the test statistic (bottom plot) that the expectation is positive, and thus with non momentum reduction the diagnostic will never activate.

Proposition 4.1: is negative expectation due to curvature or noise? We have observed that with high momentum the key inner products disappear, for small batch sizes. This indicates that they key iterates are due to loss curvature and not noise. Figure 7 is a follow up to Figures

1 and 2 from the paper, now with high momentum (0.8). We see in the histogram that the left tail is no longer there, in fact now the tail of the inner product distribution extends further to the positive end. In the plot of angle vs norm, the plot looks roughly symmetric and thus the key iterates have disappeared. Figure 8 plots the test statistic as a cumulative sum vs epoch for a number of momentum values on MNIST. No momentum reduction is used. We see that as the momentum increases, the slope of the test statistic increases, meaning a higher expected value. The learning rate has been tuned as well.

References

Nicolas Loizou and Peter Richtarik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. In *Workshop on Optimization for Machine Learning*, 2017.

Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. In *International Joint Conferences on Artificial Intelligence*, 2018.