Supplementary material for the manuscript "Implicit Regularization in Stochastic Gradient Descent with momentum"

This appendix is organized in the following manner. In Section 1, we prove Theorem 5.1 for (SGD+M) in the main paper by showing the following two parts:

- In Theorem 1.1 we prove that the local error between the continuous piecewise-differentiable trajectory and the Heavy Ball momentum SGD in each iteration is $O(h^3)$. This Theorem depends on Lemma 1.3 which shows that higher order derivatives are bounded.
- Then in Theorem 1.4 we prove that the global error between the continuous trajectory and the Heavy-ball momentum update is $O(h^2)$.

In Section 2, we prove Theorem 4.1 in the main paper for (GD+M) by showing that this is sub-case of Theorem 5.1 (in the main paper) when the mini-batch loss E_k is the full-batch loss E for each of the piecewise differentiable trajectory in Corollary 2.1 (appendix). Finally, we presented the proof for Remark-5.2 and Remark-5.3 in Section 4 and Section 3.

1 Proof of Theorem 5.1

Throughout the appendix, $\|\cdot\|$ represents ℓ_2 norm for vectors and matrices, and Frobenius norm for 3 or 4-dimensional tensors.

Theorem 1.1. *[Bound on the local error] Let the loss for each mini-batch E_n be smooth and sufficiently (4-times) differentiable, and its zeroth to fourth order derivatives are bounded, then in each iteration, the Heavy Ball momentum SGD update*

$$\begin{cases} \mathbf{x}^{n+1} = \mathbf{x}^n - h\nabla E_n(\mathbf{x}^n) + \beta(\mathbf{x}^n - \mathbf{x}^{n-1}), \quad n = 1, 2, ..., N \\ \mathbf{x}^1 = \mathbf{x}^0 - h\nabla E_0(\mathbf{x}^0) \\ \mathbf{x}^0 = \mathbf{x}^{-1} = \mathbf{0} \end{cases} \tag{1.1}$$

is locally $O(h^3)$-close to the flow of the following modified ODE when updating the n^{th} mini-batch

$$\widetilde{\mathbf{x}}'(t) = -\nabla G_n(\widetilde{\mathbf{x}}(t)) - A_n(\widetilde{\mathbf{x}}(t)), \qquad \text{for } t_n \le t < t_{n+1} \tag{1.2}$$

with $t_n = nh$. Specifically, the following equality holds for each iteration

$$\widetilde{\mathbf{x}}(t_{n+1}) = \widetilde{\mathbf{x}}(t_n) - h\nabla E_n(\widetilde{\mathbf{x}}(t_n)) + \beta(\widetilde{\mathbf{x}}(t_n) - \widetilde{\mathbf{x}}(t_{n-1})) + O(h^3). \tag{1.3}$$

Here

$$G_n(\widetilde{\mathbf{x}}) = \sum_{k=0}^{n} \beta^{n-k} E_k(\widetilde{\mathbf{x}}), A_n(\widetilde{\mathbf{x}}) = \frac{h}{2}\sum_{k=0}^{n} \beta^{n-k} C_k(\widetilde{\mathbf{x}}),$$

and $C_k(\widetilde{\mathbf{x}}) = \nabla^2 G_k(\widetilde{\mathbf{x}})\nabla G_k(\widetilde{\mathbf{x}}) + \beta\nabla^2 G_{k-1}(\widetilde{\mathbf{x}})\nabla G_{k-1}(\widetilde{\mathbf{x}}))$ with initial condition $C_0(\mathbf{x}) = \nabla^2 G_0(\widetilde{\mathbf{x}})\nabla G_0(\widetilde{\mathbf{x}})$.

Proof. Before proceeding with the proof, we state that the continuous trajectory $\widetilde{\mathbf{x}}(t)$ is differentiable in the whole domain $[0, T]$ except at the grid points $t_n, n = 0, 1, ..., N$, $N = \lfloor \frac{T}{h} \rfloor$. This is because in the interval $t_n \le t < t_{n+1}$, the continuous trajectory is given by $\widetilde{\mathbf{x}}'(t) = -\nabla G_n(\widetilde{\mathbf{x}}(t)) - A_n(\widetilde{\mathbf{x}}(t))$, whereas in the the interval $t_{n-1} \le t < t_n$, the trajectory is defined by $\widetilde{\mathbf{x}}'(t) = -\nabla G_{n-1}(\widetilde{\mathbf{x}}(t)) - A_{n-1}(\widetilde{\mathbf{x}}(t))$. We notice the following:

- the right-hand side and the left-hand side derivatives of the trajectories at any boundary point t_n are not equal, i.e, $\widetilde{\mathbf{x}}'(t_n^+) \neq \widetilde{\mathbf{x}}'(t_n^-), n = 0, 1, ..., N$. Here we define $\widetilde{\mathbf{x}}'(t_0^-) = \mathbf{0}$.
- based on Lemma 1.3, the norms of the first to third-order derivatives of the trajectory $\widetilde{\mathbf{x}}(t)$, $t \in [0, T]$ can be bounded by constants independent of h. Hence from 1.2 we can compute the left and right side derivatives at t_k (any boundary point $\forall k = 1, 2, ..., n$) as follows:

1. $\widetilde{\mathbf{x}}'(t_n^+) = -\nabla G_n(\widetilde{\mathbf{x}}(t_n)) - A_n(\widetilde{\mathbf{x}}(t_n))$ as it belongs to trajectory $t_n \leq t < t_{n+1}$
2. $\widetilde{\mathbf{x}}''(t_n^+) = \nabla^2 G_n(\widetilde{\mathbf{x}}(t_n))\nabla G_n(\widetilde{\mathbf{x}}(t_n)) + O(h)$
3. $\widetilde{\mathbf{x}}'(t_n^-) = -\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)) - A_{n-1}(\widetilde{\mathbf{x}}(t_n))$ as it belongs to trajectory $t_{n-1} \leq t < t_n$
4. $\widetilde{\mathbf{x}}''(t_n^-) = \nabla^2 G_{n-1}(\widetilde{\mathbf{x}}(t_n))\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)) + O(h)$

As $\|\widetilde{\mathbf{x}}'''(t)\|$ is bounded by a constant (Lemma 1.3), the Taylor expansion of the trajectory $\widetilde{\mathbf{x}}(t)$ is done at the point t_n on both sides:

$$\widetilde{\mathbf{x}}(t_{n+1}) = \widetilde{\mathbf{x}}(t_n) + \widetilde{\mathbf{x}}'(t_n^+)h + \frac{h^2}{2}\widetilde{\mathbf{x}}''(t_n^+) + O(h^3), \tag{1.4}$$

$$\widetilde{\mathbf{x}}(t_n) = \widetilde{\mathbf{x}}(t_{n-1}) + h\widetilde{\mathbf{x}}'(t_n^-) - \frac{h^2}{2}\widetilde{\mathbf{x}}''(t_n^-) + O(h^3). \tag{1.5}$$

Recall that the main objective of this theorem is to find how well the continuous trajectory $\widetilde{\mathbf{x}}(t)$ satisfies the H.B momentum update equation 1.1. Hence we plug $\widetilde{\mathbf{x}}(t_n)$ into 1.1 and examine the resulting error, which is also known as the Local Truncation Error (LTE) in the numerical ODE literature. The calculation is carried out as follows

$$\begin{aligned}
&\widetilde{\mathbf{x}}(t_{n+1}) - \widetilde{\mathbf{x}}(t_n) - \beta(\widetilde{\mathbf{x}}(t_n) - \widetilde{\mathbf{x}}(t_{n-1})) \\
=\ & h\widetilde{\mathbf{x}}'(t_n^+) + \frac{h^2}{2}\widetilde{\mathbf{x}}''(t_n^+) - \beta(h\widetilde{\mathbf{x}}'(t_n^-) - \frac{h^2}{2}\widetilde{\mathbf{x}}''(t_n^-)) + O(h^3) \\
=\ & -h\nabla G_n(\widetilde{\mathbf{x}}(t_n)) - hA_n(\widetilde{\mathbf{x}}(t_n)) + \frac{h^2}{2}\nabla^2 G_n(\widetilde{\mathbf{x}}(t_n))\nabla G_n(\widetilde{\mathbf{x}}(t_n)) + \beta h\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)) + \beta h A_{n-1}(\widetilde{\mathbf{x}}(t_n)) \\
& + \beta\frac{h^2}{2}\nabla^2 G_{n-1}(\widetilde{\mathbf{x}}(t_n))\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)) + O(h^3) \\
=\ & -h\underbrace{(\nabla G_n(\widetilde{\mathbf{x}}(t_n)) - \beta\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)))}_{\nabla E_n(\widetilde{\mathbf{x}}(t_n))} - h\underbrace{(A_n(\widetilde{\mathbf{x}}(t_n)) - \beta A_{n-1}(\widetilde{\mathbf{x}}(t_n)))}_{\frac{h}{2}C_n} \\
& + \frac{h^2}{2}\underbrace{(\nabla^2 G_n(\widetilde{\mathbf{x}}(t_n))\nabla G_n(\widetilde{\mathbf{x}}(t_n)) + \beta\nabla^2 G_{n-1}(\widetilde{\mathbf{x}}(t_n))\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)))}_{C_n(\text{by definition})} + O(h^3) \\
=\ & -h\nabla E_n(\widetilde{\mathbf{x}}(t_n)) - h\frac{h}{2}C_n(\widetilde{\mathbf{x}}(t_n)) + \frac{h^2}{2}C_n(\widetilde{\mathbf{x}}(t_n)) + O(h^3) \\
=\ & -h\nabla E_n(\widetilde{\mathbf{x}}(t_n)) + O(h^3).
\end{aligned}$$

The equation has been simplified using the following two identities which is easy to verify:

$$\nabla G_n(\widetilde{\mathbf{x}}(t_n)) - \beta\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n))) = E_n(\widetilde{\mathbf{x}}(t_n))$$

and

$$A_n(\widetilde{\mathbf{x}}(t_n)) - \beta A_{n-1}(\widetilde{\mathbf{x}}(t_n)) = \frac{h}{2}C_n.$$

Hence we proved that the solution of the continuous trajectory satisfies the discrete H.B. momentum updates up to an error of order $O(h^3)$.

□

Corollary 1.2. *The piecewise ODE* (1.1) *in Theorem 1.1, can be equivalently written as:*

$$\begin{aligned} &\widetilde{\mathbf{x}}'(t) = -\nabla \hat{E}_n(\widetilde{\mathbf{x}}(t)) \quad \text{for } t_n \le t < t_{n+1}, \\ &\text{where,} \quad \hat{E}_n(\widetilde{\mathbf{x}}) = G_n(\widetilde{\mathbf{x}}) + \frac{h}{4}(\|\nabla G_n(\widetilde{\mathbf{x}})\|^2 + 2\sum_{r=0}^{n-1} \beta^{n-r}\|\nabla G_r(\widetilde{\mathbf{x}})\|^2), \end{aligned} \tag{1.6}$$

which is what we used in the statement of Theorem 5.1 in the main paper.

Proof. By the definition of the ODE in Theorem 1.1, we can see

$$\begin{aligned} A_n(\widetilde{\mathbf{x}}(t)) &= \frac{h}{2}\sum_{k=1}^{n} \beta^{n-k}\left(\nabla^2 G_k(\widetilde{\mathbf{x}})\nabla G_k(\widetilde{\mathbf{x}}) + \beta\nabla^2 G_{k-1}(\widetilde{\mathbf{x}})\nabla G_{k-1}(\widetilde{\mathbf{x}})\right) + \frac{h}{2}\beta^n \nabla^2 G_0(\widetilde{\mathbf{x}})\nabla G_k(\widetilde{\mathbf{x}}) \\ &= \frac{h}{2}\sum_{k=0}^{n} \beta^{n-k}\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\nabla G_k(\widetilde{\mathbf{x}}(t)) + \frac{h}{2}\sum_{k=1}^{n} \beta^{n-k+1}\nabla^2 G_{k-1}(\widetilde{\mathbf{x}}(t))\nabla G_{k-1}(\widetilde{\mathbf{x}}(t)) \\ &= \frac{h}{2}\nabla^2 G_n(\widetilde{\mathbf{x}}(t))\nabla G_n(\widetilde{\mathbf{x}}(t)) + h\sum_{k=0}^{n-1} \beta^{n-k}\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\nabla G_k(\widetilde{\mathbf{x}}(t)). \end{aligned}$$

Therefore we can rewrite (1.2) as

$$\widetilde{\mathbf{x}}'(t) = -\nabla G_n(\widetilde{\mathbf{x}}(t)) - \frac{h}{4}\nabla\left(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2 + 2\sum_{k=0}^{n-1}\beta^{n-k}\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2\right) := \nabla \hat{E}_n(\widetilde{\mathbf{x}}(t)), \tag{1.7}$$

where $\hat{E}_n(\widetilde{\mathbf{x}}) = G_n(\widetilde{\mathbf{x}}) + \frac{h}{4}\left(\|\nabla G_n(\widetilde{\mathbf{x}})\|^2 + 2\sum_{k=0}^{n-1}\beta^{n-k}\|\nabla G_k(\widetilde{\mathbf{x}})\|^2\right)$. □

Lemma 1.3. *Under the assumption of Theorem 1.1, let* $\widetilde{\mathbf{x}}(t)$ *be defined as in* (1.2)*, then the first to third order derivatives of* $\widetilde{\mathbf{x}}$ *with respect to time are bounded, i.e., there exists constants* c_1, c_2, c_3 *such that* $\|\widetilde{\mathbf{x}}'(t)\| \le c_1, \|\widetilde{\mathbf{x}}''(t)\| \le c_2$, $\|\widetilde{\mathbf{x}}'''(t)\| \le c_3$, *for all* $t \in [0, T]$.

Proof. Although the continuous trajectory $\widetilde{\mathbf{x}}(t)$ is defined piece-wise, with the a step-size h, we want to obtain a constant (i.e., h-independent) upper bound for its derivatives so as to faithfully truncate the Taylor-expansion (1.4) to get an error of $O(h^3)$. By the assumption of Theorem 1.1, we have boundedness of $\|\nabla^{(\alpha)} E_k(\mathbf{x})\|$ for $0 \le \alpha \le 4$, i.e., with some constant c_0, for all $1 \le n \le N$,

$$\sup \|\nabla^{(\alpha)} E_n(\mathbf{x})\| \le c_0, \quad 0 \le \alpha \le 4,$$

where $\|\cdot\|$ denotes the Frobenius norm of the tensors.

Then we can immediately bound the derivatives of G_k and C_k, for any k, $\mathbf{x}$, and $0 \le \alpha \le 4$,

$$\|\nabla^{(\alpha)} G_k(\widetilde{\mathbf{x}})\| = \left\|\sum_{k=0}^{n} \beta^{n-k}\nabla^{(\alpha)} E_k(\widetilde{\mathbf{x}})\right\| \le \sum_{k=0}^{n}\beta^{n-k}\|\nabla^{(\alpha)} E_k(\widetilde{\mathbf{x}})\| \le \frac{c_0}{1-\beta}, \tag{1.8}$$

$$\|C_k(\widetilde{\mathbf{x}})\| \le \|\nabla^2 G_k(\widetilde{\mathbf{x}})\|\|\nabla G_k(\widetilde{\mathbf{x}})\| + \beta\|\nabla^2 G_{k-1}(\widetilde{\mathbf{x}})\|\|\nabla G_{k-1}(\widetilde{\mathbf{x}})\| \le \frac{1+\beta}{(1-\beta)^2}c_0^2. \tag{1.9}$$

Now, using (1.8) and (1.9) and the boundedness of h (i.e., $h \le T$), we can show that for some constant c_1,

$$\begin{aligned} \|\widetilde{\mathbf{x}}'(t)\| &\le \|G_n(\widetilde{\mathbf{x}}(t))\| + \frac{h}{2}\left\|\sum_{k=0}^{n}\beta^{n-k}C_k(\widetilde{\mathbf{x}})\right\| \\ &\le \sum_{k=0}^{n}\beta^{n-k}\max_{\widetilde{\mathbf{x}},k}\|E_k(\widetilde{\mathbf{x}})\| + \frac{h}{2}\sum_{k=0}^{n}\beta^{n-k}\max_{\widetilde{\mathbf{x}},k}\|C_k(\widetilde{\mathbf{x}})\| \\ &\le \frac{c_0}{(1-\beta)^2} + \frac{c_0^2 h(1+\beta)}{2(1-\beta)^3} := c_1. \end{aligned} \tag{1.10}$$

Next we show that $\|\widetilde{\mathbf{x}}''(t)\|$ is uniformly bounded,

$$\begin{aligned}\widetilde{\mathbf{x}}''(t) = &-\sum_{k=0}^{n}\beta^{n-k}\nabla^2 E_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\\ &-\frac{h}{2}\sum_{k=0}^{n}\beta^{n-k}\underbrace{\left(\nabla^3 G_k(\widetilde{\mathbf{x}}(t))[\widetilde{\mathbf{x}}'(t)]\nabla G_k(\widetilde{x}(t)) + \nabla^2 G_k(\widetilde{\mathbf{x}}(t))\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\right)}_{(I)}\\ &-\frac{h}{2}\beta\sum_{k=0}^{n}\beta^{n-k}\underbrace{\left(\nabla^3 G_{k-1}(\widetilde{\mathbf{x}}(t))[\widetilde{\mathbf{x}}'(t)]\nabla G_{k-1}(\widetilde{\mathbf{x}}(t)) + \nabla^2 G_{k-1}(\widetilde{\mathbf{x}}(t))\nabla^2 G_{k-1}(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\right)}_{(II)}.\end{aligned}$$

Hence we have:

$$\|\widetilde{\mathbf{x}}''(t)\| \le \|\sum_{k=0}^{n}\beta^{n-k}\nabla^2 E_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\| + \frac{h}{2(1-\beta)}\|(I)\| + \frac{h\beta}{2(1-\beta)}\|(II)\|. \tag{1.11}$$

Individually examining $\|(I)\|$ and $\|(II)\|$, we have

$$\begin{aligned}\|(I)\| &\le \|\nabla^3 G_k(\widetilde{\mathbf{x}}(t))[\widetilde{\mathbf{x}}'(t)]\nabla G_k(\widetilde{\mathbf{x}}(t))\| + \|\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\|\\ &\le \|\nabla^3 G_k(\widetilde{\mathbf{x}}(t))\|\|\widetilde{\mathbf{x}}'(t)\|\|\nabla G_k(\widetilde{\mathbf{x}}(t))\| + \|\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\|^2\|\widetilde{\mathbf{x}}'(t)\| \le 2c_0^2 c_1.\end{aligned}$$

Here in the last inequality we used the fact that $\|\widetilde{\mathbf{x}}'\| \le c_1$ as in (1.10). Similarly, $\|(II)\| \le 2c_0^2 c_1$. Putting these inequalities into 1.11, we have for some constant c_2:

$$\|\widetilde{\mathbf{x}}''(t)\| \le \|\sum_{k=0}^{n}\beta^{n-k}\nabla^2 E_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\| + \frac{h}{2(1-\beta)}\|(I)\| + \frac{h\beta}{2(1-\beta)}\|(II)\| \le c_2.$$

Finally, we bound the thrid order derivative

$$\begin{aligned}\widetilde{\mathbf{x}}'''(t) = &-\sum_{k=0}^{n}\beta^{n-k}\nabla^3 E_k(\widetilde{\mathbf{x}}(t))[\widetilde{\mathbf{x}}'(t)]\widetilde{\mathbf{x}}'(t) - \sum_{k=0}^{n}\beta^{n-k}\nabla^2 E_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}''(t)\\ &-\frac{h}{2}\sum_{k=0}^{n}\beta^{n-k}\frac{d(I)}{dt} - \frac{h}{2}\beta\sum_{k=0}^{n}\beta^{n-k}\frac{d(II)}{dt}.\end{aligned}$$

Bounding the norm on $\widetilde{\mathbf{x}}'''(t)$ based on this expression is straightforward.

□

Theorem 1.4 (Bound on the global error)**.** *Let $\widetilde{\mathbf{x}}(t)$ be the solution to* (1.2) *and assume the conditions in Theorem 1.1 hold. Then the global error $\|\mathbf{e}^n\| = \|\widetilde{\mathbf{x}}(t_n) - \mathbf{x}^n\|$ is of order $O(h^2)$, where $\widetilde{\mathbf{x}}(t_n)$ is the solution of the ODE in* (1.2) *at the n^{th} boundary point and $\mathbf{x}^n$ is the discrete H.B Momentum update.*

Proof. In Theorem 1.1, we already showed that the solution of the piecewise ODE $\widetilde{\mathbf{x}}(t_n)$ satisfies

$$\widetilde{\mathbf{x}}(t_{n+1}) = \widetilde{\mathbf{x}}(t_n) - h\nabla E_n(\widetilde{\mathbf{x}}(t_n)) + \beta(\widetilde{\mathbf{x}}(t_n) - \widetilde{\mathbf{x}}(t_{n-1})) + O(h^3), \tag{1.12}$$

and by definition, the discrete H.B momentum update satisfy

$$\mathbf{x}^{n+1} = \mathbf{x}^n - h\nabla E_k(\mathbf{x}^n) + \beta(\mathbf{x}^n - \mathbf{x}^{n-1}). \tag{1.13}$$

Let the error at the n^{th} update is denoted by $\mathbf{e}_n = \widetilde{\mathbf{x}}(t_n) - \mathbf{x}_n$, then subtracting the two updates, we have:

$$\mathbf{e}_{n+1} = \mathbf{e}_n + \beta(\mathbf{e}_n - \mathbf{e}_{n-1}) - h\underbrace{(\nabla E_n(\mathbf{x}^n) - \nabla E_n(\widetilde{\mathbf{x}}(t_n)))}_{M} + O(h^3). \tag{1.14}$$

By the assumption in Theorem 1.1, there exists some constant c_1 such that $\|\nabla^2 E_n\|_\infty := \max_{\mathbf{x}} \|\nabla^2 E_n(\mathbf{x})\| \leq c_1$. Then M can be bounded by

$$\|M\| = \|\nabla E_n(\mathbf{x}^n) - \nabla E_n(\widetilde{\mathbf{x}}(t_n))\| \leq \|\nabla^2 E_n\|_\infty \|\mathbf{e}_n\| \leq c_1 \|\mathbf{e}_n\|. \tag{1.15}$$

Now taking the norm on $\mathbf{e}_{n+1} - \mathbf{e}_n$ and applying triangular inequality on the right hand side of (1.14), we have for some constant c:

$$\|\mathbf{e}_{n+1} - \mathbf{e}_n\| \leq \beta \|\mathbf{e}_n - \mathbf{e}_{n-1}\| + h\|M\|_2 + ch^3 \leq \beta \|\mathbf{e}_n - \mathbf{e}_{n-1}\| + hc_1 \|\mathbf{e}_n\| + ch^3. \tag{1.16}$$

Defining three quantities $d_1 = \frac{c}{c_1}$, $d_2 = \frac{2c_1}{1-\beta}$, and $d_3 = \frac{2c}{1-\beta}$, now we prove the following statement using principle of induction

$$\|\mathbf{e}_n\| \leq d_1 \mathbf{e}^{d_2 hn} h^2, \quad \|\mathbf{e}_{n+1} - \mathbf{e}_n\| \leq d_3 \mathbf{e}^{d_2 hn} h^3, \quad n \geq 0. \tag{1.17}$$

We first show that base case. When $n = 0$, by definition we have $\|\mathbf{e}_0\| = \|\widetilde{\mathbf{x}}(0) - \mathbf{x}^0\| = 0$. And by (1.14), $\|\mathbf{e}_1 - \mathbf{e}_0\| \leq ch^3 < d_3 h^3$, hence the induction base holds.

Assume the proposition(1.17) holds for $(n-1)$, that is, $\|\mathbf{e}_{n-1}\| \leq d_1 \mathbf{e}^{d_2 h(n-1)} h^2$, and $\|\mathbf{e}_n - \mathbf{e}_{n-1}\| \leq d_3 \mathbf{e}^{d_2 h(n-1)} h^3$, then in the nth case,

$$\begin{aligned}
\|\mathbf{e}_n\| &\leq \|\mathbf{e}_{n-1}\| + \|\mathbf{e}_n - \mathbf{e}_{n-1}\| \\
&\leq d_1 \mathbf{e}^{d_2 h(n-1)} h^2 + d_3 \mathbf{e}^{d_2 h(n-1)} h^3 \\
&= d_1 (1 + \frac{d_3 h}{d_1}) \mathbf{e}^{d_2 h(n-1)} h^2 = d_1 (1 + d_2 h) \mathbf{e}^{d_2 h(n-1)} h^2 \\
&\leq d_1 \mathbf{e}^{d_2 hn} h^2
\end{aligned}$$

and by (1.16),

$$\begin{aligned}
&\|\mathbf{e}_{n+1} - \mathbf{e}_n\| \leq \beta d_3 \mathbf{e}^{d_2 h(n-1)} h^3 + hc_1 d_1 \mathbf{e}^{d_2 hn} h^2 + ch^3 \\
&\leq d_3 \underbrace{(\frac{d_1 c_1}{d_3} + \beta + \frac{c}{d_3})}_{=1} \mathbf{e}^{d_2 hn} h^3 = d_3 \mathbf{e}^{d_2 hn} h^3.
\end{aligned}$$

Then we have proven that the claim (1.17) also holds for the n^{th} case.

□

2 Proof of Theorem 4.1 (IGR-M)

Corollary 2.1. *Let the loss E for full-batch gradient descent be smooth and 4-times differentiable, then GD-momentum update*

$$\mathbf{x}^{n+1} = \mathbf{x}^n - h\nabla E(\mathbf{x}^n) + \beta(\mathbf{x}^n - \mathbf{x}^{n-1}), \qquad n = 1, 2, ..., N$$

is $O(h^2)$ close to the flow of the piecewise ODE

$$\widetilde{\mathbf{x}}'(t) = -\frac{1}{1-\beta} \nabla \hat{E}_n(\widetilde{\mathbf{x}}(t)), \quad t \in [t_n, t_{n+1}], \tag{2.1}$$

where the modified loss is given as

$$\hat{E}_n(\widetilde{\mathbf{x}}(t)) = (1 - \beta^{n+1}) \nabla E(\mathbf{x}(t)) + \frac{h}{2(1-\beta)} \left[\frac{(1+\beta)}{(1-\beta)} (1 - \beta^{2n+2}) - 4(n+1)\beta^{n+1} \right] \nabla^2 E(\mathbf{x}(t)) \nabla E(\mathbf{x}(t)).$$

For later intervals when $k \gg \frac{\log h}{\log \beta}$ (we only ruled out a very small number of the initial intervals, as this lower bound grows very slowly (logarithmically) as $h \to 0$), the modified loss and the ODE both become independent of k, that is, the the later GD+momentum updates $\mathbf{x}^n$ is $O(h^2)$ close to $\widetilde{\mathbf{x}}(t_n)$ which is the solution to

$$\widetilde{\mathbf{x}}'(t) = -\frac{1}{1-\beta} \nabla \hat{E}(\widetilde{\mathbf{x}}(t)),$$

with $\hat{E}(\widetilde{\mathbf{x}}(t)) = E(\widetilde{\mathbf{x}}(t)) + \frac{(1+\beta)h}{4(1-\beta)^2} \|\nabla E(\widetilde{\mathbf{x}}(t))\|^2$.

Proof. Corollary (2.1) is a special case of Theorem 1.1, and the result straightforwardly follows by setting $E_n = E$ for all $1 \leq n \leq N$.

□

3 Proof of Remark 5.2

Theorem 3.1. *The expectation of the IGRM for (SGD+M) taken over the draw of random batches is* $\mathbb{E}(IGRM_s) = \frac{h(1+\beta)}{4(1-\beta)^3}\|\nabla E(\widetilde{\mathbf{x}}(t))\|^2 + \frac{h}{4(1-\beta)^2}\mathbb{E}(\|\nabla E_j(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2)$. *This expected IGR for (SGD+M) is greater than that of the expected IGR for (SGD) taken with respect to the draw of the batches.*

Proof. To avoid confusion of notation use, we use $\mathbb{E}$ as the symbol for expectation and E as the loss-function. Here the operator $\mathbb{E}$ denotes the expectation is with respect to the draw of all the random batches. Recalling from (1.7), the implicit regularizer while updating the n^{th} mini-batch was of the form:

$$IGRM_s = \frac{h}{4}\left(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2 + 2\sum_{k=0}^{n-1}\beta^{n-k}\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2\right),$$
$$\mathbb{E}(IGRM_s) = \frac{h}{4}\left(\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2) + 2\sum_{k=0}^{n-1}\beta^{n-k}\mathbb{E}(\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2)\right). \tag{3.1}$$

Here as the expectation $\mathbb{E}$ is taken over the draw of random batches, we first derive $\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}})\|^2)$ as follows

$$\begin{aligned}
&\mathbb{E}\left(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2\right)\\
&= \mathbb{E}\left(\Big(\sum_{k=0}^{n}\beta^{n-k}\nabla E_k(\widetilde{\mathbf{x}}(t))\Big)^T\Big(\sum_{r=0}^{n}\beta^{n-r}\nabla E_k(\widetilde{\mathbf{x}}(t))\Big)\right)\\
&= \sum_{i=0}^{n}\beta^{2n-2i}\mathbb{E}\left(\|\nabla E_i(\widetilde{\mathbf{x}}(t))\|^2\right) + \sum_{k=0}^{n}\sum_{r=0,k\neq r}^{n}\beta^{2n-r-k}\mathbb{E}\left(\nabla E_k(\widetilde{\mathbf{x}}(t))^T\nabla E_r(\widetilde{\mathbf{x}}(t))\right)\\
&= \sum_{i=0}^{n}\beta^{2n-2i}\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i))\|^2) + \underbrace{\sum_{k=0}^{n}\sum_{r=0,k\neq r}^{n}\beta^{2n-r-k}\mathbb{E}\left(\nabla E_k(\widetilde{\mathbf{x}}(t_k))^T\nabla E_r(\widetilde{\mathbf{x}}(t_r))\right)}_{III} + O(h).
\end{aligned} \tag{3.2}$$

We obtain the last step by replacing the variables $\widetilde{\mathbf{x}}(t)$ by $\widetilde{\mathbf{x}}(t_i)$, $\widetilde{\mathbf{x}}(t_k)$ and $\widetilde{\mathbf{x}}(t_r)$, respectively. We note that changing from $\widetilde{\mathbf{x}}(t)$ to $\widetilde{\mathbf{x}}(t_i)$ introduces an $O(h)$ error from Taylor series. This $O(h)$ error gets multiplied with the coefficient $\frac{h}{4}$ in front of the regularizer. This overall $O(h^2)$ error does not affect the regularizer because it is of $O(h)$.

Note that the random selection of i^{th} mini-batch E_i is independent of $\widetilde{\mathbf{x}}(t_i)$, we calculate III as

$$III = \sum_{k=0}^{n}\sum_{r=0,k\neq r}^{n}\beta^{2n-r-k}\nabla E(\widetilde{\mathbf{x}}(t_k))^T\nabla E(\widetilde{\mathbf{x}}(t_r)).$$

Recall that the full-batch gradient loss is defined as $\nabla E(\widetilde{\mathbf{x}}) = \frac{1}{M}\sum_{j=1}^{M}\nabla E_{(j)}(\widetilde{\mathbf{x}})$, where $E_{(j)}$ is the jth mini-batch as in (7) of the main paper. The summands in the first term in (3.2) become

$$\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i))\|^2) = \|\nabla E(\widetilde{\mathbf{x}}(t_i))\|^2 + \mathbb{E}\|\nabla E_i(\widetilde{\mathbf{x}}(t_i)) - \nabla E(\widetilde{\mathbf{x}}(t_i))\|^2. \tag{3.3}$$

We can calculate $\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2)$ as follows

$$\begin{aligned}
&\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2)\\
&= \sum_{i=0}^{n} \beta^{2n-2i}\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i))\|^2) + \sum_{k=0}^{n}\sum_{r=0,k\neq r}^{n} \beta^{2n-r-k}\nabla E(\widetilde{\mathbf{x}}(t_k))^T\nabla E(\widetilde{\mathbf{x}}(t_r)) + O(h)\\
&= \sum_{i=0}^{n} \beta^{2n-2i}\left(\|\nabla E(\widetilde{\mathbf{x}}(t_i))\|^2 + \mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i)) - \nabla E(\widetilde{\mathbf{x}}(t_i))\|^2)\right)\\
&\quad + \sum_{k=0}^{n}\sum_{r=0,k\neq r}^{n} \beta^{2n-r-k}\nabla E(\widetilde{\mathbf{x}}(t_k))^T\nabla E(\widetilde{\mathbf{x}}(t_r)) + O(h)\\
&= \sum_{k=0}^{n}\sum_{r=0}^{n} \beta^{2n-r-k}\nabla E(\widetilde{\mathbf{x}}(t_k))^T\nabla E(\widetilde{\mathbf{x}}(t_r)) + \sum_{i=0}^{n} \beta^{2n-2i}\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i)) - \nabla E(\widetilde{\mathbf{x}}(t_i))\|^2) + O(h)\\
&= \left\|\sum_{k=0}^{n} \beta^{n-k}\nabla E(\widetilde{\mathbf{x}}(t))\right\|^2 + \sum_{i=0}^{n} \beta^{2n-2i}\mathbb{E}\left(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2\right) + O(h).\\
&= \left(\frac{1-\beta^{n+1}}{1-\beta}\right)^2 \|\nabla E(\widetilde{\mathbf{x}}(t))\|^2 + \frac{1-\beta^{2n+2}}{1-\beta^2}\mathbb{E}\left(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2\right) + O(h).
\end{aligned}$$

We write the second to last line similarly as before because changing $\widetilde{\mathbf{x}}(t_k)$ to $\widetilde{\mathbf{x}}(t)$ only introduces $O(h)$ error.

Similarly for any such k, we will have

$$\mathbb{E}(\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2) = \left(\frac{1-\beta^{k+1}}{1-\beta}\right)^2 \|\nabla E(\widetilde{\mathbf{x}}(t))\|^2 + \frac{1-\beta^{2k+2}}{1-\beta^2}\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2) + O(h).$$

Putting the expression for $\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2)$ and $\mathbb{E}(\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2)$ into 3.1, we get:

$$\begin{aligned}
&\mathbb{E}(IGRM_s) && (3.4)\\
&\approx \frac{h}{4}\left(\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2) + 2\sum_{k=0}^{n-1} \beta^{n-k}\mathbb{E}(\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2)\right) && (3.5)\\
&= \frac{h}{4}\|\nabla E(\widetilde{\mathbf{x}}(t))\|^2\left(\left(\frac{1-\beta^{n+1}}{1-\beta}\right)^2 + 2\sum_{k=0}^{n-1} \beta^{n-k}\left(\frac{1-\beta^{k+1}}{1-\beta}\right)^2\right) && (3.6)\\
&\quad + \frac{h}{4}\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2)\left(\frac{1-\beta^{2n+2}}{1-\beta^2} + 2\sum_{k=0}^{n-1} \beta^{n-k}\frac{1-\beta^{2k+2}}{1-\beta^2}\right) + O(h^2) && (3.7)
\end{aligned}$$

For large number of iterations n, (3.7) reduces to

$$\mathbb{E}(IGRM_s) = \frac{h(1+\beta)}{4(1-\beta)^3}\|\nabla E(\widetilde{\mathbf{x}}(t))\|^2 + \frac{h}{4(1-\beta)^2}\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2).$$

□

4 Proof of Remark 5.3

Theorem 4.1. *Let* $\mathbf{C}$ *be the covariance matrix of the driving force of (SGD) at the* k^{th} *iteration ,i.e,* $cov(\nabla E_k(\mathbf{x})) = \mathbf{C} \in \mathbb{R}^{p\times p}$ *then the covariance matrix for the driving force for (SGD+M) (with adjusted learning rate) is* $cov((1-\beta)\nabla G_k(\mathbf{x})) = \frac{1-\beta}{1+\beta}\mathbf{C}$. *Here the random vectors* E_k *and* G_k *are evaluated at a fixed point* $\mathbf{x}$.

Proof. Assuming the stochastic gradient $\nabla E_k(\mathbf{x})$ is sampled from a distribution with i.i.d entries where the mean is the full-batch gradient $\nabla E(\mathbf{x})$ and the covariance matrix is $\mathbf{C}$. Then by definition,

$$\mathbb{E}((1-\beta)\nabla G_k(\mathbf{x})) = (1-\beta)\sum_{i=0}^{k}\beta^{k-i}\mathbb{E}(\nabla E_i(\mathbf{x})) = (1-\beta^{k+1})\nabla E(\mathbf{x}) \approx \nabla E(\mathbf{x}).$$

From definition of $\mathbf{C}$ we have

$$\mathbf{C} = \mathbb{E}((\nabla E_k(\mathbf{x}) - \nabla E(\mathbf{x}))(\nabla E_k(\mathbf{x}) - \nabla E(\mathbf{x}))^T) = \mathbb{E}(\nabla E_k(\mathbf{x})\nabla E_k(\mathbf{x})^T) - \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T. \tag{4.1}$$

Then the covariance matrix for (SGD+M) is :

$$\begin{aligned}
&cov((1-\beta)G_k(\mathbf{x})) \\
&\quad= \mathbb{E}(((1-\beta)\nabla G_k(\mathbf{x}) - \mathbb{E}((1-\beta)\nabla G_k(\mathbf{x})))((1-\beta)\nabla G_k(\mathbf{x}) - \mathbb{E}((1-\beta)\nabla G_k(\mathbf{x}))^T) \\
&\quad= \mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T)(1-\beta)^2 - \mathbb{E}((1-\beta)\nabla G_k(\mathbf{x}))\mathbb{E}((1-\beta)\nabla G_k(\mathbf{x}))^T \\
&\quad\approx \mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T)(1-\beta)^2 - \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T.
\end{aligned} \tag{4.2}$$

Now let's evaluate $\mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T)$ as follows:

$$\begin{aligned}
&\mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T) \\
&\quad= \mathbb{E}((\sum_{k=0}^{n}\beta^{n-k}\nabla E_k(\mathbf{x}))(\sum_{k=0}^{n}\beta^{n-k}\nabla E_k(\mathbf{x}))^T) \\
&\quad= \sum_{p=0}^{n}\beta^{2n-2p}\underbrace{\mathbb{E}(\nabla E_p(\mathbf{x})\nabla E_p(\mathbf{x})^T)}_{\mathbf{C}+\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T \text{from 4.1}} + \sum_{i=0}^{n}\sum_{j=0,j\neq i}^{n}\beta^{2n-i-j}\underbrace{\mathbb{E}(\nabla E_i(\mathbf{x})\nabla E_j(\mathbf{x})^T)}_{\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T} \\
&\quad= (\mathbf{C}+\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T)\frac{(1-\beta^{2k+2})}{(1-\beta^2)} + \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T[(\frac{1-\beta^{k+1}}{1-\beta})^2 - \frac{1-\beta^{2k+2}}{1-\beta^2}] \\
&\quad= \frac{1-\beta^{2k+2}}{1-\beta^2}\mathbf{C} + (\frac{1-\beta^{k+1}}{1-\beta})^2\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T.
\end{aligned} \tag{4.3}$$

Putting 4.3 into 4.2, we have:

$$\begin{aligned}
&cov((1-\beta)G_k(\mathbf{x})) \\
&\quad\approx \mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T)(1-\beta)^2 - \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T \\
&\quad= \left(\frac{1-\beta^{2k+2}}{1-\beta^2}\mathbf{C} + (\frac{1-\beta^{k+1}}{1-\beta})^2\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T\right)(1-\beta)^2 - \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T \\
&\quad= \frac{(1-\beta^{2k+2})(1-\beta)^2}{1-\beta^2}\mathbf{C} - \beta^{k+1}(2-\beta^{k+1})\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T.
\end{aligned}$$

For a high enough iteration k, it reduces to:

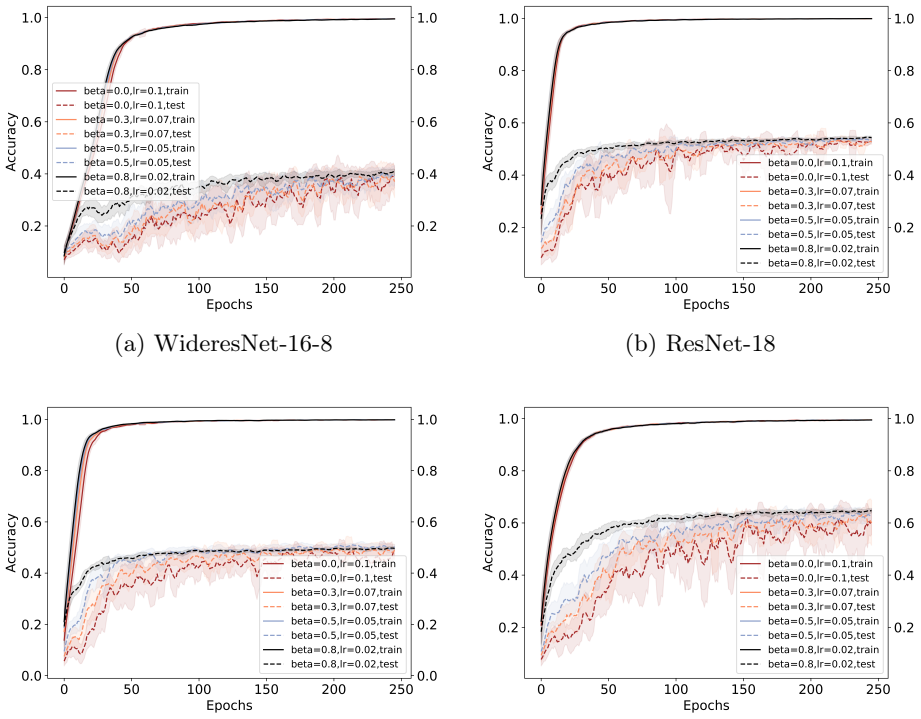
$$cov((1-\beta)G_k(\mathbf{x})) = \frac{(1-\beta)^2}{1-\beta^2}\mathbf{C} = \frac{1-\beta}{1+\beta}\mathbf{C}.$$

□

5 Additional Experiments

We delay the result for CIFAR-100 classification here in the appendix. The final test accuracy is reported in Table-1 in the manuscript.

5.1 CIFAR-100 classification results



(a) WideresNet-16-8 (b) ResNet-18

(c) ResNet-50 (d) DenseNet-121

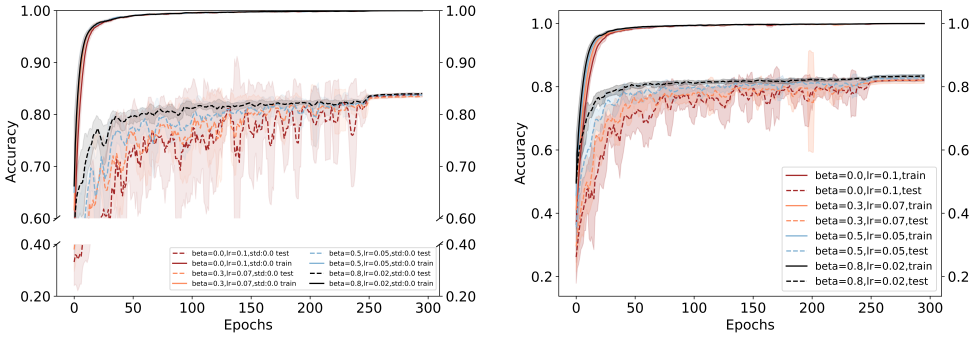
Figure 1: Classification results for CIFAR-100 dataset with various network architectures with combinations of (h, β) chosen such that the effective learning rate $\frac{h}{(1-\beta)}$ remains same. In all of the experiments, external regularization like weight-decay, l.r scheduler, dropout,label-smoothing are kept off (except Batch-normalization). The results have been averaged over 3 random seeds having different initializations

.

5.2 Effect of learning-rate scheduler

Learning rate schedulers are a common practice in training classification networks hence exploring the effect of IGR and IGR-M in schedulers is important. In the experiment, we train a Resnet-18 and a Resnet-50 network to classify CIFAR-10 dataset to compare the performance of (SGD) and (SGD+M) under the effect of learning rate scheduler.

We observe that just like our previous experiments comparing (SGD) and (SGD+M), the test accuracy is higher with increasing β. We attributed this effect due to the stronger implicit regularization for momentum than plain SGD. However, after the effect of scheduler, the learning rate is decreased by a factor of 10. This diminishes the effect of the implicit regularizer for both SGD and SGD+M as $IGR \propto h$. However, from empirical observations (Fig-2) the difference in test accuracy of (SGD) and (SGD+M) (near convergence) still exists but may not be in a pronounced way as the initial iterations. We believe this is because during the earlier iterations, the significantly stronger IGR for (SGD+M) guides it's trajectory through flatter sub-manifolds than that of (SGD). The effect is prominent enough that even after scheduler is activated (also



(a) Resnet-18 (b) Resnet-50

Figure 2: Classification result of CIFAR-10 with step scheduler ($\frac{1}{10}$) activated at epoch =250 with various β but the same effective learning rate $\frac{h}{(1-\beta)}$.

near convergence), (SGD+M) still has a slightly higher test accuracy than (SGD).

6 In the stable regime, convergence-rate of (GD+M) is $\frac{1}{(1-\beta)}$ larger than (GD) using classical convergence analysis

In this section, we show that when (GD) with a learning-rate h and (GD+M) with an effective learning rate $\frac{h}{(1-\beta)}$ both fall inside the stable regime of (GD),then the convergence-rate of (GD+M) is $\frac{1}{(1-\beta)}$ larger than (GD).

Classical convergence of (GD) and (GD+M) is considered in a locally quadratic surface. On a standard quadratic, the minimization is $\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - b^T\mathbf{x} + c$, where $\mathbf{A}$ is positive semi-definite matrix with eigen-values in $[\mu, L]$. A simple change of variable would mean doing a minimization of the form $\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}$, where Σ contains the eigenvalues of A on the diagonal. Hence $\nabla f(\mathbf{x}) = \Sigma\mathbf{x}$ and $\nabla^2 f(\mathbf{x}) = \Sigma$. Furthermore, the condition number of the objective function is denoted as $\kappa = \frac{L}{\mu}$.

For Heavy-Ball method, the iterates follow:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\nabla f(\mathbf{x}^k) + \beta(\mathbf{x}^k - \mathbf{x}^{k-1}) \tag{6.1}$$

On a locally quadratic, the iterates roughly follow

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\Sigma\mathbf{x} + \beta(\mathbf{x}^k - \mathbf{x}^{k-1}) = ((1+\beta)\mathbf{I} - h\Sigma)\mathbf{x}^k - \beta\mathbf{x}^{k-1} \tag{6.2}$$

With slight rearrangement, which could be written as :

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} (1+\beta)\mathbf{I} - h\Sigma & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix} \tag{6.3}$$

Denoting $\mathbf{y}^k = \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix}$ and $\mathbf{T} = \begin{bmatrix} (1+\beta)\mathbf{I} - h\Sigma & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$, the norm of $\|\mathbf{y}^k\|_2$ is derived as follows:

$$\|\mathbf{y}^k\| = \|\mathbf{T}\mathbf{y}^{k-1}\| = \|\mathbf{T}^k\mathbf{y}^0\| \leq \|\mathbf{T}^k\|_2\|\mathbf{y}^0\| \leq (\rho(\mathbf{T}))^k\kappa(V)\|\mathbf{y}^0\| \tag{6.4}$$

where $\rho(\mathbf{T})$ is the spectral radius of $\mathbf{T}$ and $\mathbf{T}$ has an eigen-decomposition $\mathbf{T} = VDV^{-1}$, $\kappa(V)$ being the condition number of V. $\mathbf{T}$ is permutation-similar to the block-diagonal matrix $\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 & . & . & \mathbf{0} \\ . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & . & . & \mathbf{T}_n \end{bmatrix}$, where $\mathbf{T}_j = \begin{bmatrix} 1+\beta-\alpha\lambda_j & -\beta \\ 1 & 0 \end{bmatrix}$ is a 2×2 matrix for $j = 1, 2..n$. Letting r_j denote the eigen-values for each block matrix $\mathbf{T}_j$ and would satisfy

$r_j = \begin{cases} \frac{1}{2}((1+\beta-\alpha\lambda_j) \pm \sqrt{(1+\beta-h\lambda_j)^2 - 4\beta}), & \text{if } (1+\beta-h\lambda_j)^2 - 4\beta = \Delta_j > 0 \\ \frac{1}{2}((1+\beta-\alpha\lambda_j) \pm i\sqrt{|\Delta_j|}, & \text{otherwise} \end{cases}$ where $i = \sqrt{-1}$. Due to the block-matrix structure of $\mathbf{T}$, the convergence factor $\rho(\mathbf{T})$ is determined by the largest vectors among all the block matrices $\mathbf{T}_j$, i.e, $\rho(\mathbf{T}) = \max_j r_j = \max r_1, r_n$.

Now depending upon the 4 conditions $\Delta_j \leq 0 \equiv \beta \geq (1-\sqrt{h\lambda_j})$, $\Delta_j > 0 \equiv \beta \leq (1-\sqrt{h\lambda_j})$, $|1-\sqrt{h\mu}| < |1-\sqrt{hL}|$ and $|1-\sqrt{h\mu}| > |1-\sqrt{hL}|$, we have four sub-cases to determine $\rho(\mathbf{T})$:

1. If $0 < h \leq (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ and $\beta \geq (1-\sqrt{h\mu})^2$
2. If $0 < h \leq (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ and $\beta < (1-\sqrt{h\mu})^2$
3. $h > (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ and $\beta \geq (\sqrt{hL}-1)^2$
4. $h > (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ and $\beta < (\sqrt{hL}-1)^2$

For a small h and fixed β, satisfies condition-2 and the effective learning rate lies in the stability regime of GD. Under this particular condition (2), we have $\Delta_1 > 0$, hence the spectral radius $\rho(\mathbf{T})$ becomes (by taking the larger r_j) :

$$\rho^{(GD+M)} = \frac{1}{2}(1+\beta-h\mu+\sqrt{(1+\beta-h\mu)^2-4\beta}) \quad \text{[considering the larger term]} \tag{6.5}$$

$$= \frac{1}{2}(1+\beta-h\mu+\sqrt{(1-\beta)^2-2h\mu(1+\beta)+h^2\mu^2}) \tag{6.6}$$

$$= \frac{1}{2}(1+\beta-h\mu+(1-\beta)(\underbrace{\sqrt{1-\frac{2h\mu(1+\beta)+h^2\mu^2}{(1-\beta)^2}}}_{1-\frac{1}{2}\frac{2h\mu(1+\beta)}{(1-\beta)^2}+O(h^2)}-1)+(1-\beta)) \tag{6.7}$$

$$\approx \frac{1}{2}(1+\beta-h\mu-\frac{h\mu(1+\beta)}{(1-\beta)}+(1-\beta)) \quad \text{[small } h \text{ approximation]} \tag{6.8}$$

$$= 1-\frac{h\mu}{(1-\beta)} \tag{6.9}$$

Similarly, for (GD) with learning-rate $\tilde{h}$ minimizing a locally quadratic function, using the classical convergence approach, we have $\|\mathbf{x}^k\| \leq \rho_{\tilde{h}}^k \|\mathbf{x}^0\|$ where $\rho_{\tilde{h}} = \max(|1-\tilde{h}\mu|, |1-\tilde{h}L|)$. Hence for a small enough h i.e,($0 < \tilde{h} \leq \frac{2}{L+\mu}$), we have for the convergence rate for GD to be :

$$\rho^{GD} = 1-\tilde{h}\mu \tag{6.10}$$

Putting $\tilde{h} = \frac{h}{(1-\beta)}$, we see that $\rho^{(GD+M)} \approx \rho^{(GD)}$. Which means if we use a learning rate $\frac{1}{(1-\beta)}$ times larger for GD, it will match the convergence rate of (GD+M).

Equivalently under the same learning rate for (GD) and (GD+M) (say h), the convergence rate of (GD+M) is $\frac{1}{(1-\beta)}$ times larger than that of (GD),i.e, $\rho^{(GD+M)} \approx \frac{1}{(1-\beta)}\rho^{(GD)}$.

7 Role of variance in mini-batch gradients in finding better minima

Losses of deep neural network are usually highly non-convex containing a lot of local minima. A good optimizer should have the ability of escaping local and bad (i.e., sharp) minimizers to settle for a good/flat minimum. In SGD, the mini-batch gradient can be thought of as a noisy version of the full-batch gradient: $\nabla E_i(x) = \nabla E(x) + \eta_i$. So, when an optimizer is stuck in a valley having a bad local minima, the randomness in the noisy gradient $\nabla E_i(x)$ provides a possibility of **escaping** the valley (having a bad local minima). Very recently, this intuition has been mathematically formalized by Ibayashi & Imaizumi (2021). In their Theorem 2, the authors showed that the escape efficiency (reciprocal of mean exit time) of SGD is $\propto \exp(-\frac{B}{h}\Delta E \lambda_{max}^{-\frac{1}{2}})$, where B, h, ΔE and λ_{max} denote batch-size, learning rate, depth of minima and the largest eigenvalue of the Hessian, respectively. In short, a smaller batch-size (B) and a larger learning rate are crucial to escaping bad local minima.

8 IGR-M in 2D model with non-linear (sigmoid) activation

Beyond the linear case in Section-4.1 of the manuscript, now we consider a 2D nonlinear model that has a Sigmoid activation function to explore the effect of IGR-M. The loss function E is minimized using two learnable parameters (w_1, w_2) but with a sigmoid layer in-between. Here the optimization problem is as follows:

$$(\hat{w_1}, \hat{w_2}) = \arg\min_{w_1,w_2} \frac{1}{2}(y - w_1\sigma(w_2 x))^2 \equiv \arg\min_{w_1,w_2} \frac{1}{2}\left(y - \frac{w_1}{1+e^{-w_2 x}}\right)^2 := E(w_1, w_2)$$

where σ is the Sigmoid activation function. The norm of the gradient has the following expression in this case:

$$\|\nabla E\|^2 = \left|\frac{\partial E}{\partial w_1}\right|^2 + \left|\frac{\partial E}{\partial w_2}\right|^2 = \left(\frac{1}{(1+e^{-w_2 x})^2} + \frac{w_1^2 x^2 e^{-2w_2 x}}{(1+e^{-w_2 x})^4}\right)\left(y - \frac{w_1}{1+e^{-w_2 x}}\right)^2.$$

The dashed black curve plots global minima given by the equation $w_2 = -\frac{\log(\frac{w_1}{y}-1)}{x}$. Unlike the linear case, (where the IGR was proportional to the norm of the weights w_1 and w_2), here the IGR $\|\nabla E\|^2$ has a more complicated level set (Figure 3. So, to help understand the effect of IGR-M, we plot two reference curves, one is the dark blue curve that represents the gradient flow for the original loss function, given as

$$\mathbf{x}'(t) = -\nabla E(\mathbf{x}(t)).$$

where $\mathbf{x} = [w_1, w_2]^T$. The other is the solid black curve that shows the gradient flow for implicit regularizer $\|\nabla E\|^2$ given as:

$$\mathbf{x}'(t) = -\nabla \|\nabla E(\mathbf{x}(t))\|_2^2.$$

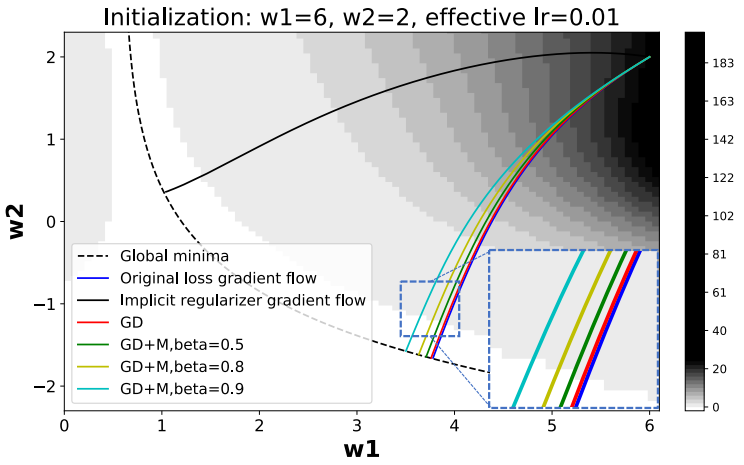


Figure 3: Trajectories for (GD) and (GD+M) for various β but with the same effective learning rate $\frac{h}{(1-\beta)}$. With increasing β, the trajectory becomes closer to the gradient flow of the implicit regularizer (solid back line), hence supporting our theory. The background color denotes the magnitude $\|\nabla E\|_2^2$

A method with a stronger IGR would have a trajectory closer to the solid black curve. So, we plot the trajectories of (GD) ($\beta = 0$) and (GD+M) with $\beta = 0.5, 0.8$, and 0.9, with the same initialization $(w_1 = 6, w_2 = 2)$. The effective learning-rate is kept the same

for all the trajectories which equals the learning rate for GD, i.e., $\frac{h}{1-\beta} = 0.01$. We see how the trajectory for (GD+M) is closer to the gradient flow for implicit regularizer (the solid black curve) than that of (GD). More explicitly, we observe that all the trajectories converge to the curve of global minima. However, with a larger β, the trajectory becomes closer to the gradient flow minimizing $\|\nabla E\|^2$ (the solid black curve). This observation agrees with our theorem which states the modified loss is a weighted combination of the original loss E and implicit regularizer $\|\nabla E\|^2$, and larger β leads to a larger weight for the regularizer $\|\nabla E\|^2$, hence making it closer to the solid black curve.

9 Future directions

IGR, although a great tool for examining generalization, relies on low-order Taylor approximations that works well under small learning rates. In addition, the current analysis is based on fixed values of β while letting $h \to 0$. In practice, to better guide the choice of hyper-parameters, a bound that is asymptotic in both h and β ($h \to 0, \beta \to 1$) might be more helpful. We leave it as future work.

References

Hikaru Ibayashi and Masaaki Imaizumi. Exponential escape efficiency of sgd from sharp minima in non-stationary regime, 2021. URL `https://arxiv.org/abs/2111.04004`.