Rethinking Learning Rate and Batch Size (Part IV): EMA

Jianlin Su

September 22, 2025

In Rethinking Learning Rate and Batch Size (Part II): Mean Field, we mentioned that one reason for focusing on SignSGD is that we typically use it as a theoretical approximation for Adam. This is a common simplification strategy used in the theoretical analysis of Adam. Besides analyzing learning rate scenarios, we have also used this simplification in posts such as "Can LoRA Gain More by Configuring Different Learning Rates?" and "A First Look at MuP: Hyperparameter Scaling Laws Across Model Scales".

However, is SignSGD truly a good approximation of Adam? One obvious difference is that the Update RMS of SignSGD is always 1, whereas this is not the case for Adam. I have found that the core reason for this discrepancy is momentum, which is ubiquitous in optimizers like Adam, Lion, and Muon. Therefore, in this article, we will examine the impact of momentum—or more broadly, EMA (Exponential Moving Average).

1 Problem Analysis

From the perspective of Adam, SignSGD corresponds to the special case where $\beta_1 = \beta_2 = 0$, or to the first update step of Adam (regardless of the values of β_1, β_2). Therefore, we believe it must share some commonalities with Adam and can capture certain general patterns.

However, there are also significant differences between them. A typical example is the difference in Update RMS: SignSGD is always 1, but Adam's is often significantly less than 1. Furthermore, Adam appears closer to SGD; it seems to be an intermediate version between SignSGD and SGD. Initially, I thought this difference was caused by the ϵ in Adam's denominator, so in "How Does Adam's Epsilon Affect the Learning Rate Scaling Law?", I specifically calculated the SoftSignSGD with ϵ.

Later, in "Why is Adam's Update RMS 0.2?", we estimated Adam's Update RMS from both simulation and theoretical perspectives. In fact, the estimate from the mean-field approximation is $\sqrt{\frac{1-\beta_1}{1+\beta_1}}$, and we verified that it aligns well with both simulation results and actual experiments. Since this result explicitly depends on β_1, it clearly points our thinking toward momentum.

This led to the following analysis. In summary, we can confirm that the role of ϵ is indeed secondary. The true protagonist is momentum—the "sliding average" of the gradient—which is precisely the subject of this article: "EMA (Exponential Moving Average)."

2 Gradient Descent

To analyze the variables introduced by EMA, we start with SGDM, which is SGD with momentum. In practice, we rarely use SGD without momentum:

$$\begin{aligned} \boldsymbol{m}_t &= \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t \\ \boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t \boldsymbol{m}_t \end{aligned} \tag{1}$$

In actual use, $\boldsymbol{g}_t$ is replaced by $\tilde{\boldsymbol{g}}_{B,t}$, which is a random variable with mean $\boldsymbol{g}_t$ and covariance matrix $\boldsymbol{\Sigma}_t/B$. These basic settings are the same as in Rethinking Learning Rate and Batch Size

(Part I): Current Status. The noise here is caused by random sampling of different batches, so we can reasonably assume that $\tilde{\boldsymbol{g}}_{B,t}$ are independent across different t.

Our task is to calculate:

$$\eta^* \approx \frac{\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^\top \boldsymbol{g}}{\mathrm{tr}(\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B \tilde{\boldsymbol{\varphi}}_B^\top]\boldsymbol{H})} \tag{2}$$

The relevant derivations have been given in previous articles and will not be repeated here. For SGDM, $\tilde{\boldsymbol{\varphi}}_B = \boldsymbol{m}_t$, which can be expanded as:

$$\boldsymbol{m}_t = (1-\beta_1)\sum_{s=1}^t \beta_1^{t-s}\tilde{\boldsymbol{g}}_{B,s} \tag{3}$$

3 Scaling the Batch Size

Now we can calculate:

$$\mathbb{E}[\boldsymbol{m}_t] = (1-\beta_1)\sum_{s=1}^t \beta_1^{t-s}\mathbb{E}[\tilde{\boldsymbol{g}}_{B,s}] = (1-\beta_1)\sum_{s=1}^t \beta_1^{t-s}\boldsymbol{g}_s \tag{4}$$

We further assume that once the model training is "on track," the gradient changes slowly. Thus, we can approximate $\boldsymbol{g}_s$ with the current gradient $\boldsymbol{g}_t$, yielding:

$$\mathbb{E}[\boldsymbol{m}_t] = (1-\beta_1)\sum_{s=1}^t \beta_1^{t-s}\boldsymbol{g}_t = (1-\beta_1^t)\boldsymbol{g}_t \approx \boldsymbol{g}_t \qquad (t\to\infty) \tag{5}$$

As for $\mathbb{E}[\boldsymbol{m}_t\boldsymbol{m}_t^\top]$, we use the identity $\mathbb{E}[\boldsymbol{m}_t\boldsymbol{m}_t^\top] = \mathbb{E}[\boldsymbol{m}_t]\mathbb{E}[\boldsymbol{m}_t]^\top + \mathrm{Cov}[\boldsymbol{m}_t,\boldsymbol{m}_t]$, and then use the additivity of variance to get:

$$\mathrm{Cov}[\boldsymbol{m}_t,\boldsymbol{m}_t] = (1-\beta_1)^2\sum_{s=1}^t \beta_1^{2(t-s)}\boldsymbol{\Sigma}_s/B \tag{6}$$

Similarly, assuming the slow variation of the covariance matrix:

$$\mathrm{Cov}[\boldsymbol{m}_t] \approx (1-\beta_1)^2\sum_{s=1}^t \beta_1^{2(t-s)}\boldsymbol{\Sigma}_t/B = (1-\beta_1)^2\frac{1-\beta_1^{2t}}{1-\beta_1^2}\boldsymbol{\Sigma}_t/B = \frac{1-\beta_1}{1+\beta_1}\boldsymbol{\Sigma}_t/B \qquad (t\to\infty) \tag{7}$$

Substituting into Equation (2), we get:

$$\eta^* \approx \frac{\eta_{\max}}{1+\frac{1-\beta_1}{1+\beta_1}\mathcal{B}_{\text{noise}}/B}, \qquad \eta_{\max} = \frac{\boldsymbol{g}^\top\boldsymbol{g}}{\boldsymbol{g}^\top\boldsymbol{H}\boldsymbol{g}}, \quad \mathcal{B}_{\text{noise}} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{H})}{\boldsymbol{g}^\top\boldsymbol{H}\boldsymbol{g}} \tag{8}$$

From this result, we can see that the introduction of the momentum mechanism is equivalent to scaling the SGD batch size by a factor of $\frac{1+\beta_1}{1-\beta_1}$. According to my understanding, momentum eliminates gradient noise at a low cost by performing EMA on the gradients along the optimization trajectory, so this result is consistent with my interpretation of the significance of momentum.

4 Sign Momentum

Furthermore, we consider SignSGDM, which can be viewed as a special case of Lion. It is essentially SGDM with an added sign operation:

$$\begin{aligned}\boldsymbol{m}_t &= \beta_1\boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t \\ \boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t\,\mathrm{sign}(\boldsymbol{m}_t)\end{aligned} \tag{9}$$

In actual training, $\boldsymbol{g}_t$ is likewise replaced by $\tilde{\boldsymbol{g}}_{B,t}$. For SignSGDM, $\tilde{\boldsymbol{\varphi}}_B = \text{sign}(\boldsymbol{m}_t)$. According to the mean-field approximation:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B] = \mathbb{E}\left[\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{m}_t^2}}\right] \approx \frac{\mathbb{E}[\boldsymbol{m}_t]}{\sqrt{\mathbb{E}[\boldsymbol{m}_t^2]}} \tag{10}$$

where vector multiplication defaults to the Hadamard product. We have already calculated the numerator $\mathbb{E}[\boldsymbol{m}_t]$ in the previous section. The denominator $\mathbb{E}[\boldsymbol{m}_t^2]$ is actually equal to $\text{diag}(\mathbb{E}[\boldsymbol{m}_t\boldsymbol{m}_t^\top])$, so we can also substitute the results from the previous section to get:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B] \approx \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{g}_t^2 + \frac{1-\beta_1}{1+\beta_1}\boldsymbol{\sigma}_t^2/B}} = \frac{\text{sign}(\boldsymbol{g}_t)}{\sqrt{1 + \frac{1-\beta_1}{1+\beta_1}(\boldsymbol{\sigma}_t^2/\boldsymbol{g}_t^2)/B}} \approx \frac{\text{sign}(\boldsymbol{g}_t)}{\sqrt{1 + \frac{1-\beta_1}{1+\beta_1}\mathcal{B}_{\text{simple}}/B}} \tag{11}$$

where $\boldsymbol{\sigma}_t^2 = \text{diag}(\boldsymbol{\Sigma}_t)$ and $\mathcal{B}_{\text{simple}} = \text{tr}(\boldsymbol{\Sigma}_t)/\boldsymbol{g}_t^\top\boldsymbol{g}_t$. This formula is equivalent to SignSGD where B is replaced by $\frac{1+\beta_1}{1-\beta_1}B$. If we further calculate $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]$, we find the same conclusion. Thus, as with SGDM, momentum is equivalent to scaling the SignSGD batch size by a factor of $\frac{1+\beta_1}{1-\beta_1}$.

In Rethinking Learning Rate and Batch Size (Part III): Muon, we calculated the learning rate laws for Muon and found them consistent with SignSGD. Therefore, we can assert that the role of momentum in Muon is the same as in SignSGDM, roughly equivalent to scaling the batch size by $\frac{1+\beta_1}{1-\beta_1}$.

5 Double Smoothing

Finally, let's look at Adam:

$$\begin{aligned}
\boldsymbol{m}_t &= \beta_1\boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t \\
\boldsymbol{v}_t &= \beta_2\boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2 \\
\hat{\boldsymbol{m}}_t &= \boldsymbol{m}_t/(1-\beta_1^t) \\
\hat{\boldsymbol{v}}_t &= \boldsymbol{v}_t/(1-\beta_2^t) \\
\boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} - \eta_t\hat{\boldsymbol{m}}_t/(\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon)
\end{aligned} \tag{12}$$

In actual training, $\boldsymbol{g}_t$ is replaced by $\tilde{\boldsymbol{g}}_{B,t}$. We consider the state where training is already "on track," i.e., $t \to \infty$, so we do not distinguish between $\boldsymbol{m}_t$ and $\hat{\boldsymbol{m}}_t$, or $\boldsymbol{v}_t$ and $\hat{\boldsymbol{v}}_t$. At the same time, we focus on the role of EMA, so we set $\epsilon = 0$. For Adam, $\tilde{\boldsymbol{\varphi}}_B = \boldsymbol{m}_t/\sqrt{\boldsymbol{v}_t}$. The difference from SignSGDM is that the denominator $\boldsymbol{m}_t^2$ is replaced by another EMA statistic $\boldsymbol{v}_t$.

From the mean-field approximation:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B] = \mathbb{E}\left[\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\right] \approx \frac{\mathbb{E}[\boldsymbol{m}_t]}{\sqrt{\mathbb{E}[\boldsymbol{v}_t]}} \tag{13}$$

We have already calculated $\mathbb{E}[\boldsymbol{m}_t]$, so we only need to calculate $\mathbb{E}[\boldsymbol{v}_t]$:

$$\mathbb{E}[\boldsymbol{v}_t] = (1-\beta_2)\sum_{s=1}^t \beta_2^{t-s}\mathbb{E}[\tilde{\boldsymbol{g}}_{B,s}^2] = (1-\beta_2)\sum_{s=1}^t \beta_2^{t-s}(\boldsymbol{g}_s^2 + \boldsymbol{\sigma}_s^2/B) \approx \boldsymbol{g}_t^2 + \boldsymbol{\sigma}_t^2/B \tag{14}$$

As before, the last approximation assumes slow variation of the gradient and variance, and $t \to \infty$. Thus, we have:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B] \approx \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{g}_t^2 + \boldsymbol{\sigma}_t^2/B}} \approx \frac{\text{sign}(\boldsymbol{g}_t)}{\sqrt{1 + \mathcal{B}_{\text{simple}}/B}} \tag{15}$$

This result is the same as for SignSGD. Therefore, from the perspective of the first moment alone, it is reasonable to use SignSGD as an approximation for Adam. However, we also have

the second moment $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]$. Under the assumption of independent components, we only need to calculate $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B^2]$:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B^2] = \mathbb{E}\left[\frac{\boldsymbol{m}_t^2}{\boldsymbol{v}_t}\right] \approx \frac{\mathbb{E}[\boldsymbol{m}_t^2]}{\mathbb{E}[\boldsymbol{v}_t]} \approx \frac{\boldsymbol{g}_t^2 + \frac{1-\beta_1}{1+\beta_1}\boldsymbol{\sigma}_t^2/B}{\boldsymbol{g}_t^2 + \boldsymbol{\sigma}_t^2/B} \tag{16}$$

6 Two Special Cases

We observe two special cases. First, when $\beta_1 = 0$, the numerator and denominator are the same, and $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B^2]$ is a vector of all ones, consistent with SignSGD. Thus, SignSGD is a good approximation for Adam with $\beta_1 = 0$ (which is RMSProp). As β_1 increases, the approximation worsens.

When $\beta_1 = 1$, we have:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B^2] \approx \frac{\boldsymbol{g}_t^2}{\boldsymbol{g}_t^2 + \boldsymbol{\sigma}_t^2/B} \approx \mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^2 \tag{17}$$

From this, we get $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top] \approx \mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^\top$. Substituting this into Equation (2), we get:

$$\eta^* \approx \frac{\|\boldsymbol{g}\|_1\sqrt{1 + \mathcal{B}_{\text{simple}}/B}}{\text{sign}(\boldsymbol{g})^\top \boldsymbol{H}\,\text{sign}(\boldsymbol{g})} \tag{18}$$

Note that this is a monotonically decreasing function of B, meaning the optimal learning rate should decrease as the batch size increases. From this, we can infer that an increase in Adam's β_1 will accelerate the appearance of the "Surge phenomenon".

This conclusion might seem a bit confusing, but it is easier to understand from another perspective. The "Surge phenomenon" refers to the situation where the optimal learning rate decreases as the batch size increases beyond a certain threshold. The results for SGDM and SignSGDM both indicate that the introduction of momentum is roughly equivalent to scaling the batch size by $\frac{1+\beta_1}{1-\beta_1} > 1$, which naturally increases the likelihood of exceeding the threshold.

In other words, the conclusion that "as β_1 increases, the Surge phenomenon will be more likely to occur" holds even for SignSGDM. While Adam has some new characteristics compared to SignSGDM, the fact that "the momentum mechanism is roughly equivalent to scaling the batch size" remains true, so it is not difficult to understand why the same conclusion arises.

7 General Analysis

Let's rewrite Equation (16):

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B^2] \approx \frac{\boldsymbol{g}_t^2 + \frac{1-\beta_1}{1+\beta_1}\boldsymbol{\sigma}_t^2/B}{\boldsymbol{g}_t^2 + \boldsymbol{\sigma}_t^2/B} = \frac{2\beta_1}{1+\beta_1}\frac{\boldsymbol{g}_t^2}{\boldsymbol{g}_t^2 + \boldsymbol{\sigma}_t^2/B} + \frac{1-\beta_1}{1+\beta_1} \approx \frac{2\beta_1}{1+\beta_1}\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^2 + \frac{1-\beta_1}{1+\beta_1} \tag{19}$$

From this, we can write:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top] \approx \mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^\top + \frac{1-\beta_1}{1+\beta_1}\,\text{diag}\left(1 - \mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^2\right) \tag{20}$$

Then:

$$\eta^* \approx \frac{\sum_i |g_i|}{\frac{1}{\beta}\frac{1-\beta_1}{1+\beta_1}\sum_i H_{i,i} + \beta\left(\sum_{i,j} H_{i,j}\,\text{sign}(g_i g_j) - \frac{1-\beta_1}{1+\beta_1}\sum_i H_{i,i}\right)} \tag{21}$$

Here, the β without a subscript is equal to $(1 + \mathcal{B}_{\text{simple}}/B)^{-1/2}$. I apologize if this is confused with β_1, β_2; I have followed the notation from the previous two articles. Unlike SignSGD, which does not exhibit the Surge phenomenon if the Hessian matrix is assumed to be diagonal, the

above formula shows that the Surge phenomenon still occurs even under the diagonal Hessian assumption:

$$\eta^* \approx \frac{\sum_i |g_i|}{\left(\frac{1}{\beta}\frac{1-\beta_1}{1+\beta_1} + \beta\frac{2\beta_1}{1+\beta_1}\right)\sum_i H_{i,i}} \tag{22}$$

By the AM-GM inequality, the above expression reaches its maximum at $\beta^* = \sqrt{\frac{1-\beta_1}{2\beta_1}}$. However, note that by definition $\beta \in (0,1)$, so we must check if $\beta^* \in (0,1)$, which requires $\beta_1 > 1/3$. If this condition is not met, the maximum is still reached at $\beta = 1$, and there is no Surge phenomenon. Conversely, when $\beta_1 > 1/3$ and $\beta > \beta^*$ (i.e., $B > \frac{1-\beta_1}{3\beta_1-1}\mathcal{B}_{\text{simple}}$), the learning rate should decrease as the batch size increases.

This conclusion can preliminarily explain why Muon can support larger batch sizes. As seen in Rethinking Learning Rate and Batch Size (Part III): Muon, Muon's behavior is similar to SignSGDM. Under certain Hessian structure assumptions, it does not exhibit the Surge phenomenon, meaning that increasing the batch size can always improve learning efficiency, although the relative gains become smaller and smaller.

In contrast, Adam, under common settings (such as $\beta_1 = 0.9$), will exhibit the Surge phenomenon even if the Hessian is assumed to be diagonal. This means that once the batch size exceeds a certain value, learning efficiency decreases.

8 Summary

This article provides a preliminary analysis of the impact of the optimizer's EMA mechanism on the scaling laws of learning rate and batch size. It confirms that the introduction of EMA, particularly the momentum mechanism, slightly alters the scaling laws. Optimizers like Adam, which involve double EMA operations, exhibit some new characteristics that differ from SignSGD.