Rethinking Learning Rate and Batch Size (Part 2): Mean Field

Jianlin Su

September 10, 2025

At the end of the previous article "Rethinking Learning Rate and Batch Size (Part 1): Status Quo", we mentioned that for cases like SignSGD and SoftSignSGD where $\tilde{\boldsymbol{\varphi}}_B$ depends non-linearly on $\tilde{\boldsymbol{g}}_B$, the cognitive load of the calculation process is quite heavy and faces difficulties in generalization. To this end, I have invested some effort in trying to simplify the derivations. Fortunately, there have been some gains, and the key idea is the theme of this article—Mean Field.

Mean field is a common approximate calculation method in physics. It does not have a fixed form, but the general idea is to move the expectation (averaging) inside the function. In fact, we already caught a glimpse of the charm of mean field in "Why is Adam's Update RMS 0.2?", and in this article, we will witness its miraculous effect in calculating the learning rate laws for SignSGD/SoftSignSGD.

1 Main Idea

Following the notation from the previous article, for SignSGD we have $\tilde{\boldsymbol{\varphi}}_B = \text{sign}(\tilde{\boldsymbol{g}}_B)$. We first need to calculate $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]$ and $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]$, from which we can calculate:

$$\eta^* \approx \frac{\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^\top \boldsymbol{g}}{\text{tr}(\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]\boldsymbol{H})} \tag{1}$$

where $\boldsymbol{g}$ is the gradient and $\boldsymbol{H}$ is the Hessian matrix. According to our assumptions, the random variable $\tilde{\boldsymbol{g}}_B$ has a mean of $\boldsymbol{g}$ and a covariance matrix of $\boldsymbol{\Sigma}/B$. We are primarily interested in the relationship between η^* and the Batch Size B. Since sign is an element-wise operation, we can start by experimenting with a single scalar. The mean field method originated from an approximate relationship I suddenly realized might hold:

$$\mathbb{E}[\text{sign}(\tilde{g}_B)] = \mathbb{E}\left[\frac{\tilde{g}_B}{\sqrt{\tilde{g}_B^2}}\right] \approx \frac{\mathbb{E}[\tilde{g}_B]}{\sqrt{\mathbb{E}[\tilde{g}_B^2]}} = \frac{g}{\sqrt{g^2 + \sigma^2/B}} \tag{2}$$

Readers who have seen "How Should the Learning Rate Change as Batch Size Increases?" might be surprised to find that this result, derived in just one line, differs from the result obtained through a long series of assumptions and approximations in the original text only by an insignificant constant $\pi/2$! This fact made me realize that the mean field approximation might be perfectly sufficient for the relationship between learning rate and batch size.

Derivations based on mean field have many advantages. First, there are fewer assumptions. The original derivation contained at least three: component independence, normal distribution, and approximating $\text{erf}(x)$ with $x/\sqrt{x^2 + c}$. However, the mean field approximation can remove the assumption of the distribution form, requiring only the assumption that the approximation itself is usable. Second, the calculation is simple. We completed the calculation in one line above, whereas the original derivation was much more complex even under many assumptions.

2 Calculation Process

In this section, we will use the mean field approximation to provide the complete calculation process for SignSGD. First is the mean $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]$. The calculation in the previous section was already nearly complete; here we only need to add a few details. Using component notation:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]_i = \mathbb{E}[\text{sign}((\tilde{g}_B)_i)] = \mathbb{E}\left[\frac{(\tilde{g}_B)_i}{\sqrt{(\tilde{g}_B)_i^2}}\right] \approx \frac{\mathbb{E}[(\tilde{g}_B)_i]}{\sqrt{\mathbb{E}[(\tilde{g}_B)_i^2]}} = \frac{g_i}{\sqrt{g_i^2 + \sigma_i^2/B}} = \frac{\text{sign}(g_i)}{\sqrt{1 + (\sigma_i^2/g_i^2)/B}} \tag{3}$$

where $\sigma_i^2 = \boldsymbol{\Sigma}_{i,i}$. Since we are ultimately interested in the relationship between η^* and B, both of which are scalars, we use the mean field approximation once more to separate the B-dependent denominator part in scalar form:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]_i \approx \frac{\text{sign}(g_i)}{\sqrt{1 + (\sigma_i^2/g_i^2)/B}} \approx \frac{\text{sign}(g_i)}{\sqrt{1 + \mathcal{B}_{\text{simple}}/B}} \triangleq \mu_i \tag{4}$$

Here $\mathcal{B}_{\text{simple}}$ is the same as in the previous article: $\mathcal{B}_{\text{simple}} = \text{tr}(\boldsymbol{\Sigma})/\boldsymbol{g}^\top\boldsymbol{g}$, which is also equal to $\mathbb{E}[\sigma_i^2]/\mathbb{E}[g_i^2]$ (where this $\mathbb{E}$ is the average over the index i). That is to say, it replaces the original σ_i^2/g_i^2 (which depends on index i) with a certain average value $\mathbb{E}[\sigma_i^2]/\mathbb{E}[g_i^2]$ that is independent of the index. After this approximation, the result is simplified but still retains the functional form with respect to B.

Next is the second moment $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]$. Here we re-introduce the component independence assumption to simplify the result. It is possible to calculate without this assumption, but the result would be more complex and would require other assumptions to simplify, so it is better to introduce the independence assumption directly. Under the independence assumption, $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]_{i,j}$ is calculated in two parts: $i \neq j$ and $i = j$. When $i \neq j$:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]_{i,j} = \mathbb{E}[(\tilde{\varphi}_B)_i(\tilde{\varphi}_B)_j] = \mathbb{E}[(\tilde{\varphi}_B)_i]\mathbb{E}[(\tilde{\varphi}_B)_j] \approx \mu_i\mu_j \tag{5}$$

The case $i = j$ is even simpler because the square of sign is necessarily 1, so its expectation is naturally 1. Therefore, the total result can be written as $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]_{i,j} \approx \mu_i\mu_j + \delta_{i,j}(1 - \mu_i\mu_j)$.

3 Anomalous Phenomena

Substituting the above calculation results into Equation (1), we get:

$$\eta^* \approx \frac{\sum_i |g_i|}{\frac{1}{\beta}\sum_i H_{i,i} + \beta\sum_{i\neq j} H_{i,j}\,\text{sign}(g_i g_j)} \tag{6}$$

where $\beta = (1 + \mathcal{B}_{\text{simple}}/B)^{-1/2}$. Note that β is monotonically increasing with respect to B, and $\beta \in (0, 1)$, so β can be seen as a standardized Batch Size. However, the expression is not always monotonic with respect to β. Thus, an anomalous behavior may occur where "as Batch Size increases, the learning rate should instead decrease," which the original paper calls the "Surge phenomenon."

Let's understand this step by step. When $B \ll \mathcal{B}_{\text{simple}}$, we have $\beta \approx \sqrt{B/\mathcal{B}_{\text{simple}}}$. Since $\beta \ll 1$, the $1/\beta$ term in the denominator of Equation (6) will dominate, leading to:

$$\eta^* \approx \frac{\sum_i |g_i|}{\sum_i H_{i,i}}\beta \approx \frac{\sum_i |g_i|}{\sum_i H_{i,i}}\sqrt{B/\mathcal{B}_{\text{simple}}} \propto \sqrt{B} \tag{7}$$

This indicates that the learning rate of SignSGD follows square root scaling at small batch sizes. Since we assume the positive definiteness of the Hessian matrix in our analysis, we must have $\sum_i H_{i,i} > 0$. Thus, when $\sum_{i\neq j} H_{i,j}\,\text{sign}(g_i g_j) \leq 0$, Equation (6) is always monotonically

increasing with respect to β, so η^* is also monotonically increasing with respect to B, and no anomalous behavior exists.

When $\sum_{i\neq j} H_{i,j}\operatorname{sign}(g_i g_j) > 0$, according to the AM-GM inequality, we can conclude that the denominator of Equation (6) has a minimum point at:

$$\beta^* = \sqrt{\frac{\sum_i H_{i,i}}{\sum_{i\neq j} H_{i,j}\operatorname{sign}(g_i g_j)}} \tag{8}$$

Note that $\beta \in (0,1)$, so there is an additional condition $\beta^* \in (0,1)$. In this case, η^* is no longer monotonically increasing with respect to B, but rather increases first and then decreases. There exists a critical Batch Size, beyond which the learning rate should instead be reduced. This is the "Surge phenomenon."

4 Reflection on Causes

Why does the anomalous behavior of the Surge phenomenon occur? In fact, this is a manifestation of the incompatibility between the optimizer's own assumptions and our analysis method. Specifically, to estimate the optimal learning rate, we expanded the increment of the Loss to a second-order approximation and assumed the positive definiteness of the Hessian matrix. Under these settings, the optimal update should be Newton's method, i.e., $\boldsymbol{H}^{-1}\boldsymbol{g}$.

From the perspective of Newton's method, different optimizers are actually different assumptions about the Hessian matrix. For example, SGD corresponds to the assumption $\boldsymbol{H} = \eta_{\max}^{-1}\boldsymbol{I}$, while SignSGD corresponds to the assumption $\boldsymbol{H} = \eta_{\max}^{-1}\operatorname{diag}(|\boldsymbol{g}|)$ (though in actual training we can only replace $\boldsymbol{g}$ with $\tilde{\boldsymbol{g}}_B$). The Surge phenomenon actually reflects that as $B\to\infty$, the deviation between the Hessian matrix assumed by SignSGD and the actual Hessian matrix increases.

We know that the parameters of today's LLM models start at hundreds of millions. Calculating either the full Hessian matrix or the covariance matrix is nearly impossible. This is one of the reasons we introduced the independence assumption when calculating the second moment $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]$; in this case, the covariance matrix is just a diagonal matrix, making estimation feasible. The Hessian matrix is similar; we can often only perform calculations for specific Hessian structures.

For example, substituting $\boldsymbol{H} = \eta_{\max}^{-1}\operatorname{diag}(|\boldsymbol{g}|)$ into Equation (6) yields $\eta^* \approx \eta_{\max}\beta = \eta_{\max}/\sqrt{1+\mathcal{B}_{\text{simple}}/B}$. This form is very concise and has no anomalous behavior. Does this mean the Surge phenomenon will not appear? No, the Surge phenomenon exists objectively. The point here is more that when we observe the Surge phenomenon in experiments, perhaps the first thing to do is not to correct the variation law of η^*, but rather to consider changing the optimizer.

5 Loss Change

With $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]$ and $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]$, we can also calculate $\overline{\Delta\mathcal{L}}$ as in the previous article. Interestingly, it has the same format as the SGD result:

$$\overline{\Delta\mathcal{L}} = \mathcal{L}(\boldsymbol{w}) - \mathbb{E}[\mathcal{L}(\boldsymbol{w}-\eta^*\tilde{\boldsymbol{g}}_B)] \approx \frac{(\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^\top\boldsymbol{g})^2}{2\operatorname{tr}(\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]\boldsymbol{H})} \approx \frac{\Delta\mathcal{L}_{\max}}{1+\mathcal{B}_{\text{noise}}/B} \tag{9}$$

where

$$\Delta\mathcal{L}_{\max} = \frac{\frac{1}{2}(\sum_i|g_i|)^2}{\sum_i H_{i,i} + \sum_{i\neq j} H_{i,j}\operatorname{sign}(g_i g_j)},\quad \mathcal{B}_{\text{noise}} = \frac{\mathcal{B}_{\text{simple}}\sum_i H_{i,i}}{\sum_i H_{i,i} + \sum_{i\neq j} H_{i,j}\operatorname{sign}(g_i g_j)} \tag{10}$$

Note that the full Hessian matrix is retained here, so the result is quite interesting—although the learning rate η^* may exhibit the Surge phenomenon, the average increment of the loss function does not. It is always monotonically increasing with respect to B and maintains the same form as SGD. This means we can derive the same "training data amount - training steps" relationship:

$$\left(\frac{S}{S_{\min}}-1\right)\left(\frac{E}{E_{\min}}-1\right)=1 \tag{11}$$

A question worth pondering is why the update amounts of SGD and SignSGD are completely different, including significant differences in the behavior of the learning rate η^*, yet the relationship of $\overline{\Delta\mathcal{L}}$ with respect to B has the same form. Is this purely a coincidence, or is there a deeper principle supporting it?

6 General Laws

Starting again from the mean field approximation, I obtained an answer leaning towards the latter. Whether for η^* or $\overline{\Delta\mathcal{L}}$, the core difficulty lies in calculating $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]$ and $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]$. Thus, our goal is to explore a unified calculation law for both.

We generally set $\tilde{\boldsymbol{\varphi}}_B=\tilde{\boldsymbol{H}}_B^{-1}\tilde{\boldsymbol{g}}_B$, where $\tilde{\boldsymbol{H}}_B$ is some semi-definite matrix. Then we can write:

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]=\mathbb{E}[\tilde{\boldsymbol{H}}_B^{-1}\tilde{\boldsymbol{g}}_B]\approx\underbrace{\mathbb{E}[\tilde{\boldsymbol{H}}_B]^{-1}}_{\text{denoted as }\hat{\boldsymbol{H}}^{-1}}\mathbb{E}[\tilde{\boldsymbol{g}}_B]=\hat{\boldsymbol{H}}^{-1}\boldsymbol{g} \tag{12}$$

and

$$\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]=\mathbb{E}[\tilde{\boldsymbol{H}}_B^{-1}\tilde{\boldsymbol{g}}_B\tilde{\boldsymbol{g}}_B^\top\tilde{\boldsymbol{H}}_B^{-1}]\approx\mathbb{E}[\tilde{\boldsymbol{H}}_B]^{-1}\mathbb{E}[\tilde{\boldsymbol{g}}_B\tilde{\boldsymbol{g}}_B^\top]\mathbb{E}[\tilde{\boldsymbol{H}}_B]^{-1}=\hat{\boldsymbol{H}}^{-1}(\boldsymbol{g}\boldsymbol{g}^\top+\boldsymbol{\Sigma}/B)\hat{\boldsymbol{H}}^{-1} \tag{13}$$

Substituting into the expression for $\overline{\Delta\mathcal{L}}$, we get:

$$\overline{\Delta\mathcal{L}}\approx\frac{1}{2}\frac{(\boldsymbol{g}^\top\hat{\boldsymbol{H}}^{-1}\boldsymbol{g})^2}{\boldsymbol{g}^\top\hat{\boldsymbol{H}}^{-1}\boldsymbol{H}\hat{\boldsymbol{H}}^{-1}\boldsymbol{g}+\operatorname{tr}(\boldsymbol{\Sigma}\hat{\boldsymbol{H}}^{-1}\boldsymbol{H}\hat{\boldsymbol{H}}^{-1})/B} \tag{14}$$

Note that the above expression is homogeneous with respect to $\hat{\boldsymbol{H}}$. If we assume that the relationship between $\hat{\boldsymbol{H}}$ and B can be separated into a scalar form such as $\hat{\boldsymbol{H}}\approx f(B)\boldsymbol{G}$, where $f(B)$ is a scalar function of B and $\boldsymbol{G}$ is not significantly related to B, then $f(B)$ can be canceled out from both the numerator and denominator. The final relationship with respect to B can be organized into the following form:

$$\overline{\Delta\mathcal{L}}\approx\frac{\Delta\mathcal{L}_{\max}}{1+\mathcal{B}_{\text{noise}}/B} \tag{15}$$

This proves that $\overline{\Delta\mathcal{L}}$ has the same asymptotic law with respect to B, the core of which is the homogeneity with respect to $\hat{\boldsymbol{H}}$. In contrast, η^* does not have such a unified result because it is not homogeneous with respect to $\hat{\boldsymbol{H}}$.

7 Validity Analysis

By now, everyone should have an understanding of the mean field method. Its main characteristic is simplicity of calculation, or more essentially, mean field chooses to calculate in directions that are simple and calculable, which leads to its great flexibility. Flexibility is also a disadvantage in many cases; it means it is difficult to grasp the law for the next step.

As for explaining why this approach is effective, that is even harder. It can only be analyzed on a case-by-case basis, and it is even possible that specific problems are difficult to analyze

further. My feeling is that the mean field method is **three parts calculation**, **three parts luck**, **three parts intuition**, plus **one part mysticism**. Of course, there is no harm in trying. Let's take the previous SignSGD calculation as an example and try to perform an analysis.

Obviously, the core calculation of SignSGD is $\mathbb{E}[\text{sign}(x)]$. We denote $\mathbb{E}[x] = \mu, \mathbb{E}[x^2] = \mu^2 + \sigma^2$, and write:

$$\text{sign}(x) = \frac{x}{\sqrt{x^2}} = \frac{x}{\sqrt{\mu^2 + \sigma^2 + (x^2 - \mu^2 - \sigma^2)}} \tag{16}$$

Assuming $x^2 - \mu^2 - \sigma^2$ is a small quantity, we perform a Taylor expansion:

$$\text{sign}(x) = \frac{x}{\sqrt{\mu^2 + \sigma^2}} - \frac{1}{2}\frac{x(x^2 - \mu^2 - \sigma^2)}{(\mu^2 + \sigma^2)^{3/2}} + \frac{3}{8}\frac{x(x^2 - \mu^2 - \sigma^2)^2}{(\mu^2 + \sigma^2)^{5/2}} - \cdots \tag{17}$$

Now the denominators are independent of x, and the numerators are polynomials in x. Taking the expectation on both sides, the first term is the result of the mean field approximation $\mu/\sqrt{\mu^2 + \sigma^2}$. To observe the rationality of the mean field approximation, we calculate the second term:

$$\frac{1}{2}\frac{\mathbb{E}[x(x^2 - \mu^2 - \sigma^2)]}{(\mu^2 + \sigma^2)^{3/2}} = \frac{1}{2}\frac{\mathbb{E}[x^3] - (\mu^3 + \mu\sigma^2)}{(\mu^2 + \sigma^2)^{3/2}} \tag{18}$$

This involves $\mathbb{E}[x^3]$, which is a new statistic and a key factor in the mean field error. We can use the normal distribution $\mathcal{N}(x; \mu, \sigma^2)$ to get a sense of it. In this case, $\mathbb{E}[x^3] = \mu^3 + 3\mu\sigma^2$. Substituting this into the above:

$$\frac{\mu\sigma^2}{(\mu^2 + \sigma^2)^{3/2}} = \frac{\sigma^2/\mu^2}{(1 + \sigma^2/\mu^2)^{3/2}} \tag{19}$$

The right side is a bounded expression, reaching its maximum at $\sigma^2/\mu^2 = 2$, with a result of $2/3^{3/2} = 0.3849\cdots$. This indicates that the error of the mean field approximation is likely finite, and the error term tends to 0 as $\sigma \to 0$ and $\sigma \to \infty$. All of these reflect the usability of the mean field approximation to some extent.

8 Generalized Approximation

The reason for choosing to analyze SignSGD is partly that we usually use it as a theoretical approximation for Adam. In "How Adam's epsilon Affects the Learning Rate Scaling Law?", we calculated a theoretically better approximation, SoftSignSGD, which considers the influence of ϵ:

$$\text{sign}(x) = \frac{x}{\sqrt{x^2}} \quad \to \quad \text{softsign}(x) = \frac{x}{\sqrt{x^2 + \epsilon^2}} \tag{20}$$

In this case, $\tilde{\boldsymbol{\varphi}}_B = \text{softsign}(\tilde{\boldsymbol{g}}_B)$. Let's get straight to the point:

$$\begin{aligned}\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]_i &= \mathbb{E}[\text{softsign}((\tilde{g}_B)_i)] = \mathbb{E}\left[\frac{(\tilde{g}_B)_i}{\sqrt{(\tilde{g}_B)_i^2 + \epsilon^2}}\right] \approx \frac{\mathbb{E}[(\tilde{g}_B)_i]}{\sqrt{\mathbb{E}[(\tilde{g}_B)_i^2] + \epsilon^2}} \\ &= \frac{g_i}{\sqrt{g_i^2 + \sigma_i^2/B + \epsilon^2}} = \frac{\text{softsign}(g_i)}{\sqrt{1 + \sigma_i^2/(g_i^2 + \epsilon^2)/B}} \approx \frac{\text{softsign}(g_i)}{\sqrt{1 + \mathcal{B}_{\text{simple}}/B}} \triangleq \nu_i\beta\end{aligned} \tag{21}$$

Here $\mathcal{B}_{\text{simple}}$ is slightly different; it is $\text{tr}(\boldsymbol{\Sigma})/(\boldsymbol{g}^\top\boldsymbol{g} + N\epsilon^2)$, where N is the total number of model parameters, i.e., $\boldsymbol{g} \in \mathbb{R}^N$. As for the final terms, $\nu_i = \text{softsign}(g_i)$ and $\beta = (1 + \mathcal{B}_{\text{simple}}/B)^{-1/2}$. Next, we calculate $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]$. Under the independence assumption, when $i \neq j$, we can still

take the means separately, so $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]_{i,j} = \nu_i\nu_j\beta^2$. Thus, we only need to calculate the case $i=j$:

$$\begin{aligned}\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]_{i,i} &= \mathbb{E}[\text{softsign}((\tilde{g}_B)_i)^2] = \mathbb{E}\left[\frac{(\tilde{g}_B)_i^2}{(\tilde{g}_B)_i^2+\epsilon^2}\right] \approx \frac{\mathbb{E}[(\tilde{g}_B)_i^2]}{\mathbb{E}[(\tilde{g}_B)_i^2]+\epsilon^2} \\ &= \frac{g_i^2+\sigma_i^2/B}{g_i^2+\sigma_i^2/B+\epsilon^2} = 1 - \frac{1-\text{softsign}(g_i)^2}{1+\sigma_i^2/(g_i^2+\epsilon^2)/B} \approx 1 - \frac{1-\text{softsign}(g_i)^2}{1+\mathcal{B}_{\text{simple}}/B}\end{aligned} \tag{22}$$

This can be written uniformly as $\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]_{i,j} \approx \nu_i\nu_j\beta^2 + \delta_{i,j}(1-\beta^2)$, so:

$$\eta^* \approx \frac{\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B]^\top\boldsymbol{g}}{\text{tr}(\mathbb{E}[\tilde{\boldsymbol{\varphi}}_B\tilde{\boldsymbol{\varphi}}_B^\top]\boldsymbol{H})} \approx \frac{\beta\sum_i \nu_i g_i}{\sum_i H_{i,i} + \beta^2(\sum_{i,j}\nu_i\nu_j H_{i,j} - \sum_i H_{i,i})} \tag{23}$$

In the above equation, except for β, all other parts are independent of B. Therefore, we have obtained the explicit relationship of η^* with respect to B, which is similar to that of SignSGD. The remaining analysis can refer to "How Adam's epsilon Affects the Learning Rate Scaling Law?" or follow the previous content.

9 Summary

In this article, we used the mean field approximation to recalculate the conclusions for SignSGD and SoftSignSGD, greatly simplifying the relevant calculation processes and preliminarily considering the general laws of these calculations.

Original address: https://kexue.fm/archives/11280