Making Model Training More Scientific (7): Step Size Scheduling and Weight Averaging

Su Jianlin

2026-07-06

Students proficient in model training know that step size scheduling, or learning rate scheduling (LR Schedule), is crucial to the final performance of the model. In previous articles, we have derived that even if only SGD is considered, the optimal learning rate function for terminal convergence also takes the Warmup-Decay form, which is commonly used in practice. However, recently there have also been some works, such as Schedule-Free, attempting to replace learning rate scheduling with some form of model weight averaging, and have also made some progress.

How can we theoretically analyze the connection between learning rate scheduling and weight averaging? To what extent can they replace each other? In this article, we will try to answer these questions.

Weight Averaging

For weight averaging, in fact, we already gave a basic result in *Making Model Training More Scientific (3): Terminal Loss Convergence of SGD*:

$$\mathbb{E}[L(\bar{\boldsymbol{\theta}}_T) - L(\boldsymbol{\theta}^*)] \leq \frac{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|^2}{2T\eta_T} + \frac{G^2}{2T}\sum_{t=1}^T \frac{\eta_t^2}{\eta_T} \tag{1}$$

where $\bar{\boldsymbol{\theta}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}_t$. This result means that as long as we choose a constant learning rate $\eta \propto 1/\sqrt{T}$, the right-hand side can achieve the ideal convergence rate of $\mathcal{O}(1/\sqrt{T})$, and the left-hand side implies that $\bar{\boldsymbol{\theta}}_T$ converges to $\boldsymbol{\theta}^*$. Thus, we design the following optimizer:

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \eta\, \boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{\theta}_t)\\ \boldsymbol{\mu}_{t+1} &= (1 - c_{t+1})\boldsymbol{\mu}_t + c_{t+1}\boldsymbol{\theta}_{t+1}\end{aligned} \tag{2}$$

where $c_t = 1/t$. That is to say, a new iteration step is added that does not interfere with the training trajectory at all, averaging the training trajectory with equal weights. The model weights used for prediction in the end are not the terminal value $\boldsymbol{\theta}_T$, but the average

value $\boldsymbol{\mu}_T$. However, this approach does not perform well in practice, which is one of the classic examples of discrepancy between theory and practice.

However, if we change c_t to some appropriate constant independent of t, such as 10^{-3}, it often performs quite well in practice—this is exactly the Exponential Moving Average (EMA) trick that the author frequently used in competitions before. Compared to $c_t = 1/t$, this moving average approach is more inclined to average the recent optimization trajectory rather than the entire historical trajectory. Its excellent performance also indicates that we do not need to pay attention to results from the too-distant past.

How to theoretically explain this moving average trick, or more generally, explain the feasibility of arbitrary weight averaging, is the core task next.

Generalization Again

In the article *Making Model Training More Scientific (4): New Identity, New Learning Rate*, to generalize the result of average loss to terminal loss, we introduced the identity

$$q_T = \frac{1}{w_{1:T}}\sum_{t=1}^{T} w_t q_t + \sum_{k=1}^{T-1}\left(\frac{1}{w_{k+1:T}} - \frac{1}{w_{k:T}}\right)\sum_{t=k+1}^{T} w_t(q_t - q_k) \tag{3}$$

where $w_{k:T} \triangleq \sum_{t=k}^{T} w_t$. In this section, we generalize it again to

$$\frac{1}{v_{1:T}}\sum_{t=1}^{T} v_t q_t = \frac{1}{w_{1:T}}\sum_{t=1}^{T} w_t q_t + \frac{1}{v_{1:T}}\sum_{k=1}^{T-1}\left(\frac{v_{k+1:T}}{w_{k+1:T}} - \frac{v_{k:T}}{w_{k:T}}\right)\sum_{t=k+1}^{T} w_t(q_t - q_k) \tag{4}$$

Intuitively, it provides an idea for transforming one weighted sum into another. When $v_T = 1$ and the remaining v_t equal 0, it degenerates into the terminal form. The proof is actually a generalization of the previous proof. Let $S_k = \frac{1}{w_{k:T}}\sum_{t=k}^{T} w_t q_t$, we can write

$$\begin{aligned}
v_{k:T}S_k - v_{k+1:T}S_{k+1} &= v_{k:T}\left(\frac{w_{k+1:T}S_{k+1} + w_k q_k}{w_{k:T}}\right) - v_{k+1:T}S_{k+1} \\
&= v_k q_k + \left(\frac{v_{k:T}}{w_{k:T}}w_k - v_k\right)q_k + \left(\frac{v_{k:T}}{w_{k:T}}w_{k+1:T} - v_{k+1:T}\right)S_{k+1} \\
&= v_k q_k + \left(\frac{v_{k:T}}{w_{k:T}}w_{k+1:T} - v_{k+1:T}\right)(S_{k+1} - q_k) \\
&= v_k q_k + \left(\frac{v_{k:T}}{w_{k:T}} - \frac{v_{k+1:T}}{w_{k+1:T}}\right)(w_{k+1:T}S_{k+1} - w_{k+1:T}q_k) \\
&= v_k q_k + \left(\frac{v_{k:T}}{w_{k:T}} - \frac{v_{k+1:T}}{w_{k+1:T}}\right)\sum_{t=k+1}^{T} w_t(q_t - q_k)
\end{aligned} \tag{5}$$

The third equality holds because directly adding the two parentheses in the second equality gives 0, so these two parentheses are opposites of each other. Now sum both left and right

sides over $k=1\sim T-1$. The left side equals $v_{1:T}S_1 - v_T q_T$. Moving $v_T q_T$ to the right side, dividing both sides by $v_{1:T}$, and rearranging slightly, we obtain the identity to be proved (4).

Scaling Transformation

Next, following the idea of the previous article *Making Model Training More Scientific (6): Top-Down Exquisite Construction*, we derive a more general convergence conclusion. Let $q_t = \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)]$, substituting it into Equation (4) gives

$$\frac{1}{v_{1:T}}\sum_{t=1}^{T} v_t \mathbb{E}[L(\boldsymbol{\theta}_t)-L(\boldsymbol{\theta}^*)] = \frac{1}{w_{1:T}}\sum_{t=1}^{T} w_t \mathbb{E}[L(\boldsymbol{\theta}_t)-L(\boldsymbol{\theta}^*)] + \frac{1}{v_{1:T}}\sum_{k=1}^{T-1}\left(\frac{v_{k+1:T}}{w_{k+1:T}} - \frac{v_{k:T}}{w_{k:T}}\right)\sum_{t=k+1}^{T} w_t \mathbb{E}[L(\boldsymbol{\theta}_t)-L(\boldsymbol{\theta}_k)] \tag{6}$$

Using the convexity of L to scale $\mathbb{E}[L(\boldsymbol{\theta}_t)-L(\boldsymbol{\theta}^*)]$ and $\mathbb{E}[L(\boldsymbol{\theta}_t)-L(\boldsymbol{\theta}_k)]$, and then carefully organizing it like the "Identity Transformation" section in the previous article, we get

$$\frac{1}{v_{1:T}}\sum_{t=1}^{T} v_t\, \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)] \leq \frac{1}{w_{1:T}}\sum_{t=1}^{T} w_t\, \mathbb{E}[\boldsymbol{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_t)\cdot(\boldsymbol{\psi}_t - \boldsymbol{\theta}^*)] \tag{7}$$

where

$$\boldsymbol{\psi}_t = \frac{w_{1:T}}{v_{1:T}}\left[\frac{v_{t:T}}{w_{t:T}}\boldsymbol{\theta}_t - \sum_{k=1}^{t-1}\left(\frac{v_{k+1:T}}{w_{k+1:T}} - \frac{v_{k:T}}{w_{k:T}}\right)\boldsymbol{\theta}_k\right] \tag{8}$$

and it can be directly verified that

$$\boldsymbol{\psi}_{t+1} - \boldsymbol{\psi}_t = \frac{v_{t+1:T}}{v_{1:T}}\frac{w_{1:T}}{w_{t+1:T}}(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) \tag{9}$$

Therefore, if we let $\boldsymbol{\psi}_t$ update according to $\boldsymbol{\psi}_{t+1} = \boldsymbol{\psi}_t - w_t \boldsymbol{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_t)$, then the update rule for the corresponding $\boldsymbol{\theta}_t$ is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \underbrace{\frac{v_{1:T}}{v_{t+1:T}}\frac{w_t w_{t+1:T}}{w_{1:T}}}_{\eta_t}\boldsymbol{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_t) \tag{10}$$

General Conclusion

The next operation is rather straightforward. Applying the conclusions from *Making Model Training More Scientific (5): Fine-tuning Learning Rate Based on Gradient* and other articles to the right-hand side of Equation (7), we get

$$\frac{1}{w_{1:T}}\sum_{t=1}^{T} w_t\, \mathbb{E}[\boldsymbol{g}(\boldsymbol{x}_t,\boldsymbol{\theta}_t)\cdot(\boldsymbol{\psi}_t - \boldsymbol{\theta}^*)] \leq \frac{1}{2w_{1:T}}\left(R^2 + \sum_{t=1}^{T} w_t^2 G_t^2\right) \tag{11}$$

where $R = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|, G_t^2 = \mathbb{E}[\|\boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{\theta}_t)\|^2]$, and we agree that $\boldsymbol{\psi}_1 = \boldsymbol{\theta}_1$. As for the left-hand side, we apply convexity to get

$$\mathbb{E}\left[L\left(\frac{1}{v_{1:T}}\sum_{t=1}^{T} v_t\boldsymbol{\theta}_t\right) - L(\boldsymbol{\theta}^*)\right] \leq \frac{1}{v_{1:T}}\sum_{t=1}^{T} v_t\,\mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)] \tag{12}$$

Combining them, we have

$$\mathbb{E}\left[L\left(\frac{1}{v_{1:T}}\sum_{t=1}^{T} v_t\boldsymbol{\theta}_t\right) - L(\boldsymbol{\theta}^*)\right] \leq \frac{1}{2w_{1:T}}\left(R^2 + \sum_{t=1}^{T} w_t^2 G_t^2\right) \tag{13}$$

The minimum value on the right side is achieved at

$$w_t = \frac{RG_t^{-2}}{\sqrt{Q_T}}, \qquad Q_T = \sum_{k=1}^{T} G_k^{-2} \tag{14}$$

which implies that the optimal learning rate takes the form

$$\eta_t = \frac{v_{1:T}}{v_{t+1:T}}\frac{w_t w_{t+1:T}}{w_{1:T}} = \frac{v_{1:T}}{v_{t+1:T}}\frac{RG_t^{-2}}{\sqrt{Q_T}}\left(1 - \frac{Q_t}{Q_T}\right) \tag{15}$$

Interpretation of Results

The terminal convergence result given in the previous article *Making Model Training More Scientific (6): Top-Down Exquisite Construction* is

$$\mathbb{E}\left[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)\right] \leq \frac{1}{2w_{1:T}}\left(R^2 + \sum_{t=1}^{T} w_t^2 G_t^2\right) \tag{16}$$

The right-hand side is consistent with Equation (13), but the learning rate becomes $\eta_t = \frac{w_t w_{t+1:T}}{w_{1:T}}$. From this, it can be seen that if the averaged weights want to achieve a similar effect to the terminal weights, the learning rate needs to be multiplied by an extra factor of $\frac{v_{1:T}}{v_{t+1:T}}$, or conversely, under the same learning rate, the effect of weight averaging is equivalent to the terminal effect after the learning rate is multiplied by $\frac{v_{t+1:T}}{v_{1:T}}$. It is obvious that $\frac{v_{t+1:T}}{v_{1:T}}$ is less than or equal to 1 and monotonically decreasing. Multiplying by this term plays a role similar to LR Decay.

Let us consider some specific examples. First is the simplest $v_t \equiv 1$, corresponding to Equation (2) at the beginning. In this case, $\frac{v_{t+1:T}}{v_{1:T}} = 1 - t/T$, which is exactly a linear decay. This means that "constant learning rate + equal weight averaging" theoretically has a similar effect to a linear decay learning rate. However, the measured effect is often that directly applying linear decay to the learning rate yields good terminal effects, but the effect of averaged weights, i.e., Equation (2), is suboptimal.

Reviewing the entire derivation process, we can guess that the problem probably lies in the assumption of a "convex function". This series of articles actually all relies on the assumption that "the objective function is convex", but the actual situation is often non-convex, and the conclusions of convex optimization cannot be precisely reproduced. A compromise is to assume a local approximate convexity, so equal weight averaging can be performed over a finite interval, or one can directly consider its smoothed version—Exponential Moving Average (EMA). In this case, $v_t = \gamma^{-t}$ (where $0 < \gamma < 1$), then

$$\frac{v_{t+1:T}}{v_{1:T}} = \frac{1-\gamma^{T-t}}{1-\gamma^T} \tag{17}$$

It maintains an averaging interval of $\mathcal{O}((1-\gamma)^{-1})$. Due to the memoryless property of exponential functions, if we run EMA while training with a constant learning rate, it is equivalent to simultaneously obtaining the weights after $\mathcal{O}((1-\gamma)^{-1})$ steps of linear decay at each step, allowing us to preview the effect after linear decay in advance.

Related Work

Regarding the suboptimal performance of Equation (2), the paper *The Road Less Scheduled* proposes "Schedule-Free", which changes the point for calculating the gradient to an interpolation of $\boldsymbol{\theta}_t$ and $\boldsymbol{\mu}_t$:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_t &= (1-\beta)\boldsymbol{\theta}_t + \beta\boldsymbol{\mu}_t \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \eta\, \boldsymbol{g}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_t) \\ \boldsymbol{\mu}_{t+1} &= (1-c_{t+1})\boldsymbol{\mu}_t + c_{t+1}\boldsymbol{\theta}_{t+1}\end{aligned} \tag{18}$$

Intuitively, the original averaging operation was completely independent of the training process, and the model never perceived whether $\boldsymbol{\mu}_t$ was good or not. Now we change the gradient calculation point to an interpolation of $\boldsymbol{\theta}_t$ and $\boldsymbol{\mu}_t$, allowing the model to perceive the quality of $\boldsymbol{\mu}_t$ and thus make adjustments. This modification is theoretically guaranteed and can indeed significantly improve the effect of the averaged weight $\boldsymbol{\mu}_T$.

However, as the training scale expands, Schedule-Free has gradually shown a decline. The authors recently proposed ScheduleFree+, adding quite a few "patches" to the original Schedule-Free, making it an SOTA again. However, the Plus version adds quite a lot of extra operations, making the whole algorithm complicated and increasing the difficulty of understanding, so we will not expand on it in detail here. We will come back to discuss it later.

It is worth mentioning that the authors of Schedule-Free and its Plus version are exactly the authors of the classic paper *Optimal Linear Decay Learning Rate Schedules and Further Refinements* that we introduced in previous articles. These works have all successfully applied classical convex optimization theory to practical training, rather than staying at "armchair theorizing" on toy models. They deserve our respect!

Extended Thinking

What is truly worth pondering is, can the premise of Schedule-Free really hold? The so-called Schedule-Free has two goals: first, it hopes to train from beginning to end with a constant learning rate; second, it hopes that various hyperparameters do not depend on the total number of training steps T. If these two conditions are met, the training process can be stopped at any time, and also resumed at any time. And if it stops at any random step, it is the optimal solution up to the current step. It is indeed very ideal.

However, these two goals themselves are open to question. First, for actual training, there is no essential difference between a constant learning rate and dynamic Cosine Decay, Linear Decay, etc. It is all about tuning and comparing to choose the best, so pursuing a constant learning rate has little practical significance. Moreover, even Schedule-Free and its Plus version have not gotten rid of the dependence on Warmup, so in reality, they still somewhat rely on a manually customized LR Schedule.

Secondly, hoping that various hyperparameters do not depend on the total number of training steps T is indeed a good goal, but looking at it only from a theoretical perspective, it is very difficult to hold. Taking Equation (2) as an example again, even if all assumptions are satisfied and Equation (2) can perform well, its optimal learning rate η still depends on T. Similarly, Schedule-Free cannot escape this fact; it obtains the result that η does not depend on T by introducing new assumptions.

We can also observe this from Equation (15). For equal weight averaging, we have $\frac{v_{1:T}}{v_{t+1:T}} = 1/(1-t/T)$. Further assuming G_t is a constant G, then $1 - Q_t/Q_T = 1 - t/T$, which exactly cancels out. Therefore, the optimal learning rate is $\frac{RG^{-2}}{\sqrt{Q_T}}$, where Q_T depends on T. On the other hand, G_t is usually large first and then small, so according to the characteristic in Equation (15) that it is inversely proportional to G_t^2, the early learning rate should be smaller. This can actually explain the phenomenon that Schedule-Free still cannot completely get rid of Schedule and still needs Warmup.

For this reason, ScheduleFree+ and some concurrent works have gradually shifted from "Schedule-Free" to "LR-Free". But overall, this area is still in a "budding" stage, awaiting more essential exploration.

Article Summary

This article introduced a new identity and continued to generalize the results of the previous article, which allows us to quantitatively reveal the connection between weight averaging and learning rate decay in theory. In addition, we also discussed issues such as why equal weight averaging performs worse than moving average, and why Schedule-Free still cannot completely get rid of Schedule.

For more detailed reposting matters, please refer to: *Scientific Space FAQ*