

Making Model Training More Scientific (5): Fine-Tuning Learning Rates Based on Gradients

Su Jianlin

2026-01-09

Introduction

In the previous four articles, we explored convergence results for SGD ranging from bounded domains to unbounded domains, and from average loss to final-point loss. Some readers might feel that SGD is an outdated topic perhaps something from the "ancient era"? That is not true at all! For instance, the core identity used in the fourth article, *Making Model Training More Scientific (4): A New Identity, A New Learning Rate*, dates back only to 2023; the conclusions in the third article, *Making Model Training More Scientific (3): SGD's Final-Point Loss Convergence*, are slightly earlier but still from 2020.

Moreover, as shown in the fourth article, we derived the commonly used practical learning rate strategy "linear decay," demonstrating that this series of theoretical derivations is not merely "armchair theorizing," but can effectively guide real-world practice. Next, we will discuss more refined learning rate strategies based on gradients. Understanding these strategies helps us grasp the principles behind learning rate scheduling and also lays the foundation for various adaptive learning rate optimizers.

1 The Initial Starting Point

If we revisit the earlier proof process carefully, we find that the entire series of conclusions begins with an unassuming identity:

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\varphi}\|^2 &= \|\boldsymbol{\theta}_t - \eta_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) - \boldsymbol{\varphi}\|^2 \\ &= \|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|^2 - 2\eta_t(\boldsymbol{\theta}_t - \boldsymbol{\varphi}) \cdot \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) + \eta_t^2 \|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2\end{aligned}\tag{1}$$

It is called "unassuming" because it is so trivial any competent high school student could understand and prove it. Yet, this simple identity supports a wide range of results in stochastic optimization, a testament to the profound simplicity and elegance of fundamental principles.

To understand this identity, we must note two aspects of "arbitrariness." First, φ is arbitrary. In most cases, we set it to the theoretical optimal solution θ^* to obtain the most meaningful results, but this does not change the inherent arbitrariness of φ . As we have seen in *Making Model Training More Scientific (3)*, *Making Model Training More Scientific (4)*, inserting $\varphi = \theta_{T-k}$ and $\varphi = \theta_k$ was key to deriving those results.

Second, $\mathbf{g}(\mathbf{x}_t, \theta_t)$ is arbitrary. This may seem surprising at first glance, since earlier articles assumed $\mathbf{g}(\mathbf{x}_t, \theta_t)$ to be the gradient of the loss function $\nabla_{\theta_t} L(\mathbf{x}_t, \theta_t)$. However, the sole purpose of this assumption is to combine it with the convexity of L and obtain:

$$L(\mathbf{x}_t, \theta_t) - L(\mathbf{x}_t, \varphi) \leq (\theta_t - \varphi) \cdot \mathbf{g}(\mathbf{x}_t, \theta_t) \quad (2)$$

which links the identity to the loss function. If we do not require this or have alternative ways to achieve it, the gradient assumption becomes unnecessary. This arbitrariness is crucialit allows us to construct more complex update rules and is key to generalizing results beyond SGD-based optimizers.

2 Classic Results

Slightly rearranging equation (1), we get:

$$2\eta_t(\theta_t - \varphi) \cdot \mathbf{g}(\mathbf{x}_t, \theta_t) = \|\theta_t - \varphi\|^2 - \|\theta_{t+1} - \varphi\|^2 + \eta_t^2 \|\mathbf{g}(\mathbf{x}_t, \theta_t)\|^2 \quad (3)$$

The "traditional" approach then proceeds by dividing both sides by $2\eta_t$ and summing over $t = 1$ to T . The advantage is that the left-hand side can leverage convexity to establish direct inequalities with the loss. The drawback is that one must assume a bounded domain and a non-increasing learning rate to easily bound the right-hand side. So far, only in the first article, *Making Model Training More Scientific (1): SGD's Average Loss Convergence*, have we used the "traditional" method, resulting in:

$$\frac{1}{T} \sum_{t=1}^T L(\mathbf{x}_t, \theta_t) - \frac{1}{T} \sum_{t=1}^T L(\mathbf{x}_t, \theta^*) \leq \frac{R^2}{2T\eta_T} + \frac{1}{2T} \sum_{t=1}^T \eta_t \|\mathbf{g}(\mathbf{x}_t, \theta_t)\|^2 \quad (4)$$

Here we explicitly retain $\|\mathbf{g}(\mathbf{x}_t, \theta_t)\|$ instead of assuming $\|\mathbf{g}(\mathbf{x}_t, \theta_t)\| \leq G$ and simplifying further. With the assumption of a non-increasing learning rate, we have:

$$\frac{R^2}{2T\eta_T} + \frac{1}{2T} \sum_{t=1}^T \eta_t \|\mathbf{g}(\mathbf{x}_t, \theta_t)\|^2 \geq \frac{R^2}{2T\eta_T} + \frac{\eta_T}{2T} \sum_{t=1}^T \|\mathbf{g}(\mathbf{x}_t, \theta_t)\|^2 \geq \frac{R}{T} \sqrt{\sum_{t=1}^T \|\mathbf{g}(\mathbf{x}_t, \theta_t)\|^2} \quad (5)$$

Let $V_t = \sum_{k=1}^t \|\mathbf{g}(\mathbf{x}_k, \theta_k)\|^2$. The equality condition holds when $\eta_1 = \eta_2 = \dots = \eta_T = R/\sqrt{V_T}$. That is, the right-hand side of inequality (4) achieves its minimum with the constant learning rate $R/\sqrt{V_T}$, which gives the fastest convergence. However, this result

violates causality: at step t , we cannot "know" the gradient magnitudes at future steps. One causality-compliant modification is:

$$\eta_t = \frac{R}{\sqrt{V_t}} = \frac{R}{\sqrt{\sum_{k=1}^t \|\mathbf{g}(\mathbf{x}_k, \boldsymbol{\theta}_k)\|^2}} \quad (6)$$

With this modification, we need to re-prove:

$$\sum_{t=1}^T \eta_t \|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2 = R \sum_{t=1}^T \frac{V_t - V_{t-1}}{\sqrt{V_t}} \leq 2R \sum_{t=1}^T \frac{V_t - V_{t-1}}{\sqrt{V_t} + \sqrt{V_{t-1}}} = 2R \sum_{t=1}^T (\sqrt{V_t} - \sqrt{V_{t-1}}) = 2R\sqrt{V_T} \quad (7)$$

Substituting back into inequality (4) gives:

$$\frac{1}{T} \sum_{t=1}^T L(\mathbf{x}_t, \boldsymbol{\theta}_t) - \frac{1}{T} \sum_{t=1}^T L(\mathbf{x}_t, \boldsymbol{\theta}^*) \leq \frac{3R}{2T} \sqrt{\sum_{t=1}^T \|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2} \quad (8)$$

This result is only 50% worse than the ideal (5), which is quite acceptable more importantly, it respects causality and provides a practical learning rate strategy. It is also the prototype of the AdaGrad optimizer. Applying (6) element-wise per parameter component yields the standard AdaGrad algorithm, which we will return to later.

3 Be Careful with Expectations

Starting from the second article, *Making Model Training More Scientific (2): Extending Conclusions to Unbounded Domains*, we adopted a "modern" technique directly summing both sides of (3):

$$\begin{aligned} \sum_{t=1}^T 2\eta_t (\boldsymbol{\theta}_t - \boldsymbol{\varphi}) \cdot \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) &= \|\boldsymbol{\theta}_1 - \boldsymbol{\varphi}\|^2 - \|\boldsymbol{\theta}_{T+1} - \boldsymbol{\varphi}\|^2 + \sum_{t=1}^T \eta_t^2 \|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2 \\ &\leq \|\boldsymbol{\theta}_1 - \boldsymbol{\varphi}\|^2 + \sum_{t=1}^T \eta_t^2 \|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2 \end{aligned} \quad (9)$$

Clearly, the "modern" approach avoids assuming monotonic learning rates or bounded domains; the terms $\|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|^2$ naturally cancel out. The trade-off is that the left-hand sum is weighted by η_t . This weight is manageable for SGD but nearly fatal for adaptive learning rate optimizers. Hence, despite its elegance, the modern method rarely generalizes to adaptive optimizers, and we must often fall back on the "traditional" method.

Lets return to SGD analysis. Using convexity on the left-hand side:

$$\sum_{t=1}^T \eta_t [L(\mathbf{x}_t, \boldsymbol{\theta}_t) - L(\mathbf{x}_t, \boldsymbol{\varphi})] \leq \frac{\|\boldsymbol{\theta}_1 - \boldsymbol{\varphi}\|^2}{2} + \frac{1}{2} \sum_{t=1}^T \eta_t^2 \|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2 \quad (10)$$

In the first three articles, the next step was to take the mathematical expectation $\mathbb{E}$ on both sides. But we must be very careful here! The expectation is taken over all $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. In those articles, we assumed the learning rate η_t was independent of the data, so $\mathbb{E}[\eta_t[L(\mathbf{x}_t, \boldsymbol{\theta}_t) - L(\mathbf{x}_t, \boldsymbol{\varphi})]] = \eta_t \mathbb{E}[L(\mathbf{x}_t, \boldsymbol{\theta}_t) - L(\mathbf{x}_t, \boldsymbol{\varphi})] = \eta_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\varphi})]$, linking the left-hand side to the target loss. However, in this article we consider gradient-dependent learning rates, where η_t cannot be simply factored out.

A simple fix is to assume η_t depends only on $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{t-1}$, so the expectation over $\mathbf{x}_t$ can be computed separately, e.g., $\mathbb{E}[\eta_t L(\mathbf{x}_t, \boldsymbol{\theta}_t)] = \mathbb{E}[\eta_t \mathbb{E}_{\mathbf{x}_t}[L(\mathbf{x}_t, \boldsymbol{\theta}_t)]] = \mathbb{E}[\eta_t L(\boldsymbol{\theta}_t)]$. But this is restrictive. Therefore, we will instead assume η_t is data-independent. Does this prevent gradient-based learning rate adjustment? Not entirely, it just means we can only adjust based on the expected gradient magnitude $G_t = \sqrt{\mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2]}$, leading to:

$$\sum_{t=1}^T \eta_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\varphi})] \leq \frac{\mathbb{E}[\|\boldsymbol{\theta}_1 - \boldsymbol{\varphi}\|^2]}{2} + \frac{1}{2} \sum_{t=1}^T \underbrace{\eta_t^2 \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2]}_{G_t^2} \quad (11)$$

4 Still the Old Way

With inequality (11), the rest of the derivation becomes routine. We will extend the results from the second, third, and fourth articles to see what new insights they offer for learning rate tuning. The second article, *Making Model Training More Scientific (2)*, has two results. The first assumes non-increasing learning rates, replaces η_t on the left with η_T , and sets $\boldsymbol{\varphi} = \boldsymbol{\theta}^*$, yielding:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)] \leq \frac{R^2}{2T\eta_T} + \frac{1}{2T\eta_T} \sum_{t=1}^T \eta_t^2 G_t^2 \quad (12)$$

Here $R = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|$. Since the learning rate is non-increasing, the right-hand side is minimized with a constant learning rate:

$$\eta_t = \frac{R}{\sqrt{\sum_{k=1}^T G_k}} \quad (13)$$

Again, this violates causality, so we modify it to:

$$\eta_t = \frac{R}{\sqrt{\sum_{k=1}^t G_k^2}} = \frac{R}{\sqrt{\sum_{k=1}^t \mathbb{E}[\|\mathbf{g}(\mathbf{x}_k, \boldsymbol{\theta}_k)\|^2]}} \quad (14)$$

But the $\mathbb{E}$ in the denominator is unrealistic; it requires averaging over many repeated runs. We simply drop the $\mathbb{E}$ as an approximation, or assume that $\|\mathbf{g}(\mathbf{x}_k, \boldsymbol{\theta}_k)\|^2$ has very low variance, so a single sample suffices. This is a common default operation in what follows. Thus, after removing $\mathbb{E}$, this result is essentially the same as (6), offering no new insights yet.

5 Inverse to Gradient

The second result in *Making Model Training More Scientific (2)* is a weighted average form:

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)]}{\sum_{t=1}^T \eta_t} \leq \frac{R^2}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 G_t^2}{2 \sum_{t=1}^T \eta_t} \quad (15)$$

Using the Cauchy-Schwarz inequality:

$$\sum_{t=1}^T \eta_t^2 G_t^2 = \frac{\sum_{t=1}^T (\eta_t G_t)^2 \sum_{t=1}^T G_t^{-2}}{\sum_{t=1}^T G_t^{-2}} \geq \frac{\left(\sum_{t=1}^T \eta_t G_t G_t^{-1}\right)^2}{\sum_{t=1}^T G_t^{-2}} = \frac{\left(\sum_{t=1}^T \eta_t\right)^2}{\sum_{t=1}^T G_t^{-2}} \quad (16)$$

Equality holds when $\eta_t G_t \propto G_t^{-1}$, i.e., $\eta_t \propto G_t^{-2}$. Substituting into the right-hand side of (15):

$$\frac{R^2}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 G_t^2}{2 \sum_{t=1}^T \eta_t} \geq \frac{R^2}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t}{2 \sum_{t=1}^T G_t^{-2}} \geq \frac{R}{\sqrt{\sum_{t=1}^T G_t^{-2}}} \quad (17)$$

The overall equality condition is $\eta_t = R G_t^{-2} / \sqrt{Q_T}$, where $Q_t = \sum_{k=1}^t G_k^{-2}$. Most strikingly, this implies the learning rate should be inversely proportional to the square of the gradient magnitude. This helps explain the necessity of warmup: early in training, gradients are large, but they decrease and stabilize over time. The causality-compliant form is:

$$\eta_t = \frac{R G_t^{-2}}{\sqrt{Q_t}} = \frac{R G_t^{-2}}{\sqrt{\sum_{k=1}^t G_k^{-2}}} \quad (18)$$

We must re-prove it after modification:

$$\sum_{t=1}^T \eta_t = \sum_{t=1}^T \frac{R(Q_t - Q_{t-1})}{\sqrt{Q_t}} \geq \sum_{t=1}^T \frac{R(Q_t - Q_{t-1})}{\sqrt{Q_t} + \sqrt{Q_{t-1}}} = \sum_{t=1}^T R(\sqrt{Q_t} - \sqrt{Q_{t-1}}) = R\sqrt{Q_T} \quad (19)$$

$$\sum_{t=1}^T \eta_t^2 G_t^2 = \sum_{t=1}^T \frac{R^2(Q_t - Q_{t-1})}{Q_t} = R^2 + \sum_{t=2}^T \frac{R^2(Q_t - Q_{t-1})}{Q_t} \leq R^2 + R^2 \sum_{t=2}^T \ln \frac{Q_t}{Q_{t-1}} = R^2 + R^2 \ln \frac{Q_T}{Q_1} \quad (20)$$

Substituting into (15):

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)]}{\sum_{t=1}^T \eta_t} \leq \frac{R}{\sqrt{Q_T}} \left(1 + \frac{1}{2} \ln \frac{Q_T}{Q_1}\right) \quad (21)$$

Compared to the optimal, the key difference is the logarithmic factor $\ln(Q_T/Q_1)$, a common phenomenon when shifting from static to dynamic learning rates.

6 Focus on the Present

From the third article, *Making Model Training More Scientific (3): SGD's Final-Point Loss Convergence*, our results began focusing on final-point loss convergence. Generalizing its core conclusion to dynamic gradients gives:

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \frac{R^2}{2T\eta_T} + \frac{1}{2\eta_T} \sum_{t=1}^T \frac{\eta_t^2 G_t^2}{\max(1, T-t)} \quad (22)$$

This result also requires non-increasing η_t , and the right-hand side is minimized at $R\sqrt{V_T/T}$ with constant learning rate:

$$\eta_t = \frac{R}{\sqrt{TV_T}}, \quad V_t = \sum_{k=1}^t \frac{G_k^2}{\max(1, t-k)} \quad (23)$$

This learning rate is particularly interesting it shares the denominator form with the "mini" AdaGrad rate (6), where V_t sums gradient squares. The difference is that at time k , the gradient is weighted by $1/\max(1, t-k)$, emphasizing more recent gradients. This is a new feature arising from the shift from average loss to final-point loss, resembling RMSProp and Adam, which use EMA to update second moments.

Intuitively, the causal version might be $\eta_t = R/\sqrt{tV_t}$, but this is imprecise. The correct version is $\eta_t = R/\sqrt{TV_t}$. Substituting into (22):

$$\begin{aligned} \frac{R^2}{2T\eta_T} + \frac{1}{2\eta_T} \sum_{t=1}^T \frac{\eta_t^2 G_t^2}{\max(1, T-t)} &= \frac{R}{2} \sqrt{\frac{V_T}{T}} \left(1 + \sum_{t=1}^T \frac{G_t^2/V_t}{\max(1, T-t)} \right) \\ &\leq \frac{R}{2} \sqrt{\frac{V_T}{T}} \left(1 + \sum_{t=1}^T \frac{1}{\max(1, T-t)} \right) \\ &\leq \frac{R}{2} \sqrt{\frac{V_T}{T}} (3 + \ln T) \end{aligned} \quad (24)$$

where we used $G_t^2 \leq V_t$. Compared to the optimal, we again have a logarithmic factor $\ln T$.

7 The Culmination

In *Making Model Training More Scientific (4)*, we derived the strongest final-point loss convergence result to date. Extending it to dynamic gradients yields the "culmination" of (15) and (22):

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \frac{R^2}{2\eta_{1:T}} + \frac{1}{2} \sum_{t=1}^T \frac{\eta_t^2 G_t^2}{\eta_{\min(t+1, T):T}} \quad (25)$$

This result looks simple but is actually very complex at least, we cannot intuitively see under which learning rate pattern the right-hand side is minimized. In the previous article, the author suggested a continuous approximation with variational calculus. We can try the same here: let $S_t = \eta_{\min(t+1, T):T}$. Then for $t < T-1$, $\eta_t = S_{t-1} - S_t \approx -\dot{S}_t$. Approximating η_t by $-\dot{S}_t$ and sums by integrals, the right-hand side becomes approximately (substitute $S_t = W_t^2$):

$$\frac{R^2}{2S_0} + \frac{1}{2} \int_0^T \frac{\dot{S}_t^2 G_t^2}{S_t} dt = \frac{R^2}{2W_0^2} + 2 \int_0^T \dot{W}_t^2 G_t^2 dt \quad (26)$$

By definition $W_T = 0$. Fix W_0 . The integral becomes a boundary-fixed variational problem. The EulerLagrange equation gives $\frac{d}{dt}(\dot{W}_t G_t^2) = 0$, so $\dot{W}_t \propto G_t^{-2}$. Integrating and using $W_T = 0$ yields $W_t = W_0 \int_t^T G_s^{-2} ds / \int_0^T G_s^{-2} ds$. Substituting into (26) gives $R^2/2W_0^2 + 2W_0^2 / \int_0^T G_s^{-2} ds$, minimized when $2W_0^2 = R(\int_0^T G_s^{-2} ds)^{1/2}$. Finally, using $\eta_t \approx -\dot{S}_t = -2W_t \dot{W}_t$, we get:

$$\eta_t \approx \frac{R G_t^{-2} \int_t^T G_s^{-2} ds}{(\int_0^T G_s^{-2} ds)^{3/2}} \quad (27)$$

Discretizing, we guess the optimal learning rate is roughly:

$$\eta_t = \frac{R G_t^{-2} (Q_T - Q_t)}{Q_T^{3/2}} = \frac{R G_t^{-2}}{\sqrt{Q_T}} (1 - Q_t/Q_T) \quad (28)$$

where $Q_t = \sum_{k=1}^t G_k^{-2}$. Remarkably, this "guess" is actually the correct answer! However, verifying it by substitution into (25) is extremely difficult: because the denominator is $\eta_{t+1:T}$, not $\eta_{t:T}$, we cannot guarantee $\eta_t/\eta_{t+1:T}$ is bounded, making the bounding process daunting. The author spent a week without success. We will not pursue this further; instead, in the next article, a more elegant construction will be used to prove it.

Now, let's analyze (28). Note that $R G_t^{-2} / \sqrt{Q_T}$ is exactly the optimal rate from (15), proportional to G_t^{-2} , which explains warmup. Equation (28) multiplies this by $1 - Q_t/Q_T$, which is monotonically decreasing and, under mild assumptions, linearly decaying explaining decay. Thus, (28) corresponds precisely to the classic "warmup-decay" learning rate schedule.

Even more interesting: the optimal rate in (22) tells us to focus more on recent gradients, but (28) is more extreme it depends only on the current and future gradients, completely ignoring past gradients. This violates causality. How can we make it causal? Replace $\sqrt{Q_T}$ with $\sqrt{Q_t}$, and write Q_t/Q_T as $(Q_T/T) \times T$, then replace Q_T/T with Q_t/t , resulting in:

$$\eta_t = \frac{R G_t^{-2}}{\sqrt{Q_t}} (1 - t/T) \quad (29)$$

This seems reasonable, but proving its validity by substituting back into (25) remains nontrivial.

8 Summary

In this article, we began considering gradient-based learning rate scheduling. These analyses help us understand the principles behind strategies like warmup and decay and provide valuable insights for various adaptive learning rate optimizers.

*Please include this link when reposting: <https://kevue.fm/archives/11530>
For detailed reposting guidelines, consult: Scientific Space FAQ*