

Making Alchemy More Scientific (VI): A Top-Down Elegant Construction

Su Jianlin

2026-01-16

In *Making Alchemy More Scientific (V): Gradient-Based Fine-Tuning of Learning Rates*, we entered a new chapter on scheduling learning rates based on gradients. However, as mentioned at the end of that article, we encountered difficulties in proving the optimality of the learning rate sequence that minimizes the final loss under dynamic gradients. Specifically, the optimal learning rate sequence "guessed" via the calculus of variations becomes very hard to validate by substitution and bounding. Thus, not only is optimality out of reach, we cannot even determine whether the proposed sequence is feasible.

In this article, we resolve this issue by obtaining a more precise result through an elegant construction. The proof process suggests that the precision of this result may be unimprovable. This breakthrough originates from the paper *Optimal Linear Decay Learning Rate Schedules and Further Refinements*.

1 Problem Recap

First, let us revisit the previous result. At the end of the last article, we obtained a general version of the conclusion from *Making Alchemy More Scientific (IV): A New Identity, A New Learning Rate*:

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

We aim to find a non-negative sequence $\eta_1, \eta_2, \dots, \eta_T \geq 0$ that minimizes the right-hand side of this inequality. Through continuous approximation and variational methods, we "guessed" the solution to be

$$\eta_t = \frac{RG_t^{-2}}{\sqrt{Q_T}}(1 - Q_t/Q_T) \quad (2)$$

where $Q_t = \sum_{k=1}^t G_k^{-2}$. However, we were unable to rigorously verify this guessproving its optimality would require additional assumptions. The main difficulty lies in the

term $\eta_{\min(t+1,T):T} = \eta_{t+1:T}$ (for $t < T$) appearing in the denominator of the sum, which makes it hard to bound ratios like $\eta_t/\eta_{t+1:T}$, complicating any bounding attempts. If we could improve the result so that the denominator becomes $\eta_{t:T}$, the proof would become "straightforward".

This article achieves the final proof by further increasing the precision of inequality (1). Rather than improving it explicitly, we carefully bound terms and construct the optimal learning rate sequence in a top-down manner, thus implicitly refining the precision.

2 Careful Bounding

Our starting point is the identity from *Making Alchemy More Scientific (IV): A New Identity, A New Learning Rate*:

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

Let $q_t = \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)]$, where $\mathbb{E}$ denotes expectation over all $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. Substituting gives

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] = \frac{1}{w_{1:T}} \sum_{t=1}^T w_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)] + \sum_{k=1}^{T-1} \left(\frac{1}{w_{k+1:T}} - \frac{1}{w_{k:T}} \right) \sum_{t=k+1}^T w_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}_k)] \quad (4)$$

From here on, we adhere strictly to the principle of "bounding only when necessary" to achieve maximum precision. Now, apply convexity for the first bounding step:

$$\mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)] = \mathbb{E}[L(\mathbf{x}_t, \boldsymbol{\theta}_t) - L(\mathbf{x}_t, \boldsymbol{\theta}^*)] \leq \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)] \quad (5)$$

$$\mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}_k)] = \mathbb{E}[L(\mathbf{x}_t, \boldsymbol{\theta}_t) - L(\mathbf{x}_t, \boldsymbol{\theta}_k)] \leq \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_k)] \quad (6)$$

Note that this requires each $\boldsymbol{\theta}_t$ to depend only on $\mathbf{x}_1, \dots, \mathbf{x}_{t-1}$, which holds in stochastic optimization. The second line also requires $t \geq k$, which is clearly satisfied. Substituting into (4) yields

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \underbrace{\frac{1}{w_{1:T}} \sum_{t=1}^T w_t \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)]}_{(A)} + \underbrace{\sum_{k=1}^{T-1} \left(\frac{1}{w_{k+1:T}} - \frac{1}{w_{k:T}} \right) \sum_{t=k+1}^T w_t \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_k)]}_{(B)} \quad (7)$$

In the article *Making Alchemy More Scientific (IV)*, the next step was to bound terms (A) and (B) separately and then add the results. This separate bounding amplifies the error and causes difficulties in the subsequent proof.

3 Identity Transformation

Therefore, in this section, we combine (A) and (B) through identity transformations before bounding, to achieve higher precision. First, assume the learning rate is independent of the data $\mathbf{x}_t$, so we can move the expectation $\mathbb{E}$ outside:

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \mathbb{E} \left[\underbrace{\frac{1}{w_{1:T}} \sum_{t=1}^T w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)}_{(A)} + \underbrace{\sum_{k=1}^{T-1} \left(\frac{1}{w_{k+1:T}} - \frac{1}{w_{k:T}} \right) \sum_{t=k+1}^T w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_k)}_{(B)} \right] \quad (8)$$

The second term is more complex. Using $\sum_{k=1}^{T-1} \sum_{t=k+1}^T = \sum_{t=1}^T \sum_{k=1}^{t-1}$ to swap summations:

$$\begin{aligned} (B) &= \sum_{t=1}^T \sum_{k=1}^{t-1} \left(\frac{1}{w_{k+1:T}} - \frac{1}{w_{k:T}} \right) w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_k) \\ &= \sum_{t=1}^T w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot \sum_{k=1}^{t-1} \left(\frac{1}{w_{k+1:T}} - \frac{1}{w_{k:T}} \right) (\boldsymbol{\theta}_t - \boldsymbol{\theta}_k) \\ &= \sum_{t=1}^T w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot \left(\left(\frac{1}{w_{t:T}} - \frac{1}{w_{1:T}} \right) \boldsymbol{\theta}_t - \sum_{k=1}^{t-1} \left(\frac{1}{w_{k+1:T}} - \frac{1}{w_{k:T}} \right) \boldsymbol{\theta}_k \right) \end{aligned} \quad (9)$$

Adding term (A), the term $\frac{1}{w_{1:T}} w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot \boldsymbol{\theta}_t$ cancels out. Collecting the remaining terms:

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \mathbb{E} \left[\frac{1}{w_{1:T}} \sum_{t=1}^T w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot \underbrace{\left(\frac{w_{1:T}}{w_{t:T}} \boldsymbol{\theta}_t - w_{1:T} \sum_{k=1}^{t-1} \left(\frac{1}{w_{k+1:T}} - \frac{1}{w_{k:T}} \right) \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \right)}_{\text{denote as } \boldsymbol{\psi}_t} \right] \quad (10)$$

Denoting the above expression as $\boldsymbol{\psi}_t$, the right-hand side takes the standard form of weighted average loss convergence:

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \mathbb{E} \left[\frac{1}{w_{1:T}} \sum_{t=1}^T w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\psi}_t - \boldsymbol{\theta}^*) \right] = \frac{1}{w_{1:T}} \sum_{t=1}^T w_t \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\psi}_t - \boldsymbol{\theta}^*)] \quad (11)$$

4 Update Rule

From the definition of $\boldsymbol{\psi}_t$, it follows directly that:

$$\boldsymbol{\psi}_{t+1} - \boldsymbol{\psi}_t = \frac{w_{1:T}}{w_{t+1:T}} (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) \quad (12)$$

Therefore, if we update ψ_t via $\psi_{t+1} = \psi_t - w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)$, then:

$$\psi_{t+1} = \psi_t - w_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \quad \Rightarrow \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{w_t w_{t+1:T}}{w_{1:T}} \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \quad (13)$$

For this ψ_t update rule, we have already proven in *Making Alchemy More Scientific (II): Extending Results to Unbounded Domains* and *Making Alchemy More Scientific (V): Gradient-Based Fine-Tuning of Learning Rates* that the following inequality holds:

$$\frac{1}{w_{1:T}} \sum_{t=1}^T w_t \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\psi_t - \boldsymbol{\theta}^*)] \leq \frac{1}{2w_{1:T}} \left(R^2 + \sum_{t=1}^T w_t^2 \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2] \right) \quad (14)$$

where $R = \|\psi_1 - \boldsymbol{\theta}^*\|$. If we initialize ψ_t and $\boldsymbol{\theta}_t$ at the same point, then $R = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|$.

Some readers might question this: in the above inequality, $\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)$ is the gradient at $\boldsymbol{\theta}_t$, not at ψ_t . Can we still use the previous result? As clarified in *Making Alchemy More Scientific (V)*, the purpose of using the gradient $\mathbf{g}(\mathbf{x}_t, \psi_t)$ at the current point ψ_t is to use convexity to relate $\mathbf{g}(\mathbf{x}_t, \psi_t) \cdot (\psi_t - \boldsymbol{\theta}^*)$ to the loss function. However, that relation is not needed here, so using $\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)$ is acceptable.

5 Strongest Result

Now substitute (14) into inequality (11) to perform the second bounding step:

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \frac{1}{2w_{1:T}} \left(R^2 + \sum_{t=1}^T w_t^2 G_t^2 \right) \quad (15)$$

where $G_t^2 = \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2]$. This means that if we perform SGD with learning rate $\eta_t = \frac{w_t w_{t+1:T}}{w_{1:T}}$, the final loss satisfies the above inequality. Due to more careful bounding, this upper bound is theoretically tighter than (1). However, it is somewhat implicit: we must first solve for w_t from η_t before evaluating the right-hand side, and solving w_t explicitly is non-trivial.

Instead, we can reverse the process: derive the optimal w_t first, then obtain the optimal η_t via $\eta_t = \frac{w_t w_{t+1:T}}{w_{1:T}}$. The minimum of the right-hand side of (15) was already computed in the previous article *Making Alchemy More Scientific (V)*. The solution is:

$$w_t = \frac{R G_t^{-2}}{\sqrt{Q_T}}, \quad Q_T = \sum_{k=1}^T G_k^{-2} \quad (16)$$

Hence, the optimal η_t is

$$\eta_t = \frac{w_t w_{t+1:T}}{w_{1:T}} = \frac{R G_t^{-2}}{\sqrt{Q_T}} (1 - Q_t/Q_T) \quad (17)$$

This matches exactly with (2)! Thus, we have completed the proof of the optimality of (2).

Regarding this result, we provided an initial interpretation in the previous article: the factor G_t^{-2} , inversely proportional to the squared gradient norm, explains the necessity of early warmup, since gradients are typically large initially; the factor $1 - Q_t/Q_T$ monotonically decays to zero, explaining the necessity of learning rate decay. In particular, if we assume the gradient norm is constant, we recover a linearly decaying learning rate commonly used in practice.

6 Post Hoc Adjustment

How can (17) guide practice? First, it is acausal and cannot be used directly. The original paper *Optimal Linear Decay Learning Rate Schedules and Further Refinements* provides a post hoc validation/adjustment strategy.

The idea is simple: pick any learning rate schedule and train once. Then, use the observed gradient information to compute the optimal learning rate (17) and compare its curve with the chosen schedule. If there is a large discrepancy, adjust the schedule and retrain. This process is called "Refinement" in the paper and is suitable for scenarios with preparatory experiments before full-scale training.

The paper provides several examples of refined learning rates, which mostly exhibit a "Warmup-Decay" pattern. Notably, in most experiments, the gradient norm in later stages is nearly constant, so the optimal decay shape during those stages closely resembles linear decay:

Finally, the above result applies to SGD, where $w_t \propto \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^{-2}]$. In practice, we approximate $\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|^2$; here, $\|\cdot\|$ denotes the L2 norm. For adaptive optimizers like Adam, the paper suggests using $w_t \propto \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)\|_1^{-1}]$, i.e., inversely proportional to the L1 norm. We will discuss adaptive optimizers in future articles.

7 Explicit Version

If we choose a specific learning rate schedule η_t and want to evaluate its optimality, the process is currently cumbersome because (15) is semi-implicit. To verify it, we must solve $\eta_t = \frac{w_t w_{t+1:T}}{w_{1:T}}$, which is not particularly hard but leads to non-compact solutions, making exact proofs difficult.

Here, we further bound to obtain an explicit result in terms of η_t . The bounding technique is somewhat subtle and took the author considerable time, but the final expression is quite understandable:

$$\eta_t = \frac{w_t w_{t+1:T}}{w_{1:T}} \leq \frac{w_t (w_{t:T} + w_{t+1:T})}{2w_{1:T}} = \frac{w_{t:T}^2 - w_{t+1:T}^2}{2w_{1:T}} \quad (18)$$

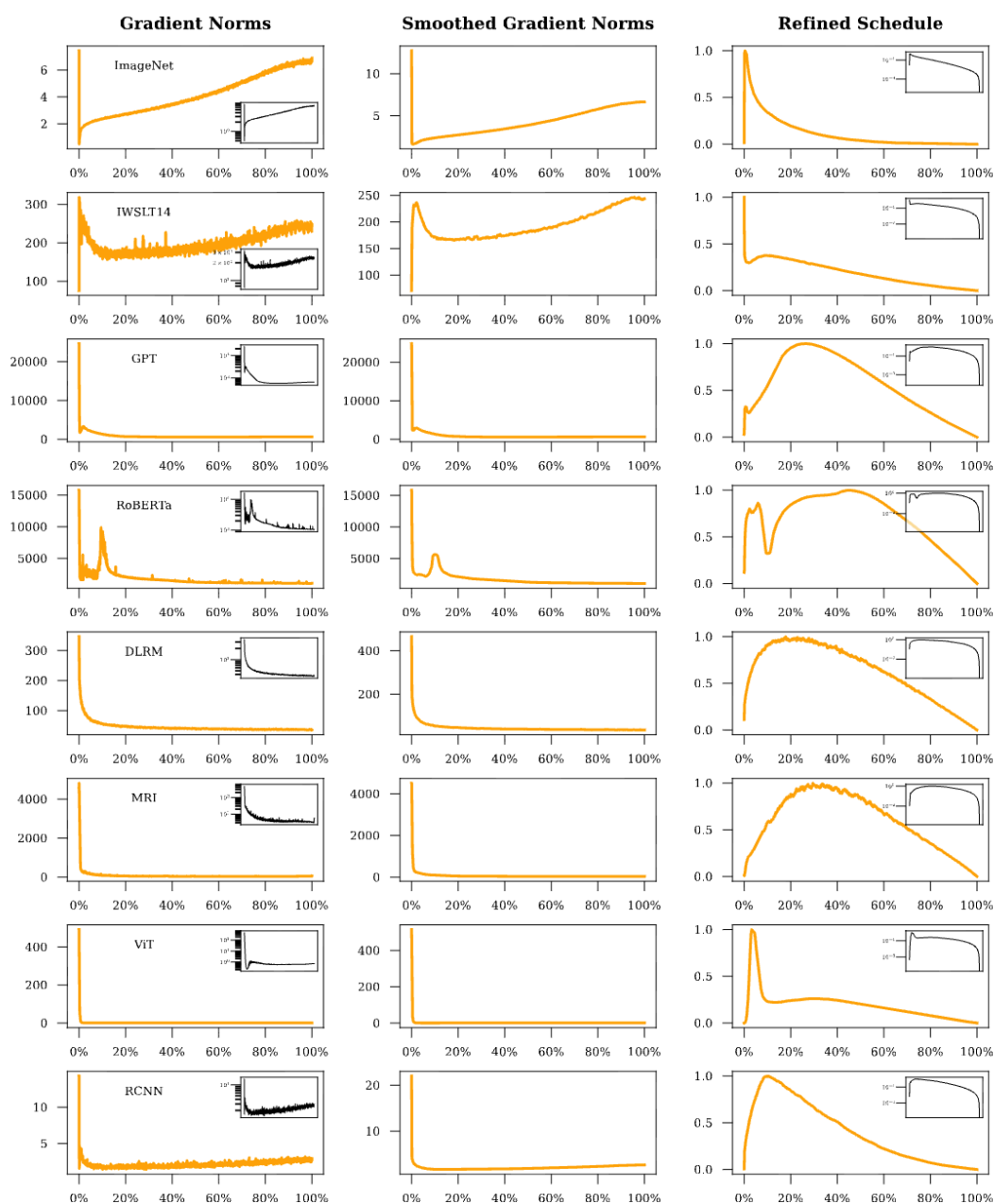


Figure 3: Gradient Norm sequences and the resulting Refined schedules, generated using an initial linear decay schedule with warmup for the initial run. Log scale views are inset for scale reference.

Figure 1: Examples of refined learning rates

Here, $w_{t:T}^2$ means $(w_{t:T})^2$. Summing both sides from t to T gives

$$\eta_{t:T} \leq \frac{w_{t:T}^2}{2w_{1:T}} \quad (19)$$

Substituting $t = 1$ gives $\frac{1}{w_{1:T}} \leq \frac{1}{2\eta_{1:T}}$. Replacing t with $t + 1$ gives $\frac{w_{1:T}}{w_{t+1:T}^2} \leq \frac{1}{2\eta_{t+1:T}}$.

Combining with the definition of η_t , we obtain $\frac{w_t^2}{w_{1:T}} = \eta_t^2 \frac{w_{1:T}}{w_{t+1:T}^2} \leq \frac{\eta_t^2}{2\eta_{t+1:T}}$. Finally, rewriting the right-hand side of (15) and substituting these bounds yields

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \frac{R^2}{2w_{1:T}} + \sum_{t=1}^T \frac{w_t^2}{2w_{1:T}} G_t^2 \leq \frac{R^2}{4\eta_{1:T}} + \sum_{t=1}^{T-2} \frac{\eta_t^2}{4\eta_{t+1:T}} G_t^2 + \frac{w_{T-1}^2}{2w_{1:T}} G_{T-1}^2 + \frac{w_T^2}{2w_{1:T}} G_T^2 \quad (20)$$

We do not bound the last two terms because, by definition, $\eta_T = 0$, which makes $\frac{\eta_t^2}{2\eta_{t+1:T}}$ infinite for $t = T-1$ and $t = T$. This also shows that $\eta_1, \dots, \eta_T$ have only $T-1$ free parameters, while the unknowns $w_1, \dots, w_T$ number T . With fewer equations than unknowns, we have an extra degree of freedom to adjust.

From the definition, $\eta_{T-1} = \frac{w_{T-1}w_T}{w_{1:T}}$. By the AM-GM inequality:

$$\frac{w_{T-1}^2}{2w_{1:T}} G_{T-1}^2 + \frac{w_T^2}{2w_{1:T}} G_T^2 \geq \frac{w_{T-1}w_T}{w_{1:T}} G_{T-1}G_T = \eta_{T-1} G_{T-1}G_T \quad (21)$$

Due to the existence of the "adjustable degree of freedom", we can choose w_{T-1}, w_T appropriately (specifically, set $w_{T-1}G_{T-1} = w_TG_T$) to achieve equality:

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \frac{R^2}{4\eta_{1:T}} + \sum_{t=1}^{T-2} \frac{\eta_t^2}{4\eta_{t+1:T}} G_t^2 + \eta_{T-1} G_{T-1}G_T \quad (22)$$

The author has not found this result in existing literature, so it is tentatively considered new. Its precision exceeds that of (1). Of course, (1) is not suitable for learning rate sequences that end at zero, so a direct comparison is not straightforward. However, judging from the leading coefficients of the first two main terms $1/2$ in (1) and $1/4$ here the average precision should be roughly doubled.

8 Summary

At the end of the previous article, we highlighted the difficulty in proving the optimality of the learning rate schedule minimizing the final loss. In this article, we resolved this issue by a top-down, careful construction and bounding process. We completed the proof and obtained a result of higher precision. We also discussed how this result provides insight into the "Warmup-Decay" mechanism of learning rates.

To reprint, please include this link: <https://kexue.fm/archives/11540>

For more details on reprints, see: *Science Space FAQ*