

Making Alchemy More Scientific (Part 10): Dismantling and Patching the Monotonicity Assumption

Su Jianlin

2026-09-14

In the previous post [Making Alchemy More Scientific \(Part 9\): Classic Adaptive Gradient Algorithms](#), we established a convergence framework for the general update rule $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{H}_t^{-1} \mathbf{g}_t$, and used it as a starting point to re-understand classic adaptive gradient algorithms such as AdaGrad. However, the complete conclusion there relied on two premises: first, that the learning rate η_t is constant, and second, that the preconditioning matrix $\mathbf{H}_t$ is non-decreasing in t with respect to the semidefinite order (in the positive semidefinite sense).

The former restricts the freedom of learning-rate strategies, while the latter excludes moving-average-type algorithms such as RMSProp and Adam from the theory. This post attempts to remove these two “scaffolds” one by one, extending the convergence theory to more general settings.

1 Remaining Questions

Let us first recall the core conclusion of the previous post. For the general update rule

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{H}_t^{-1} \mathbf{g}_t \quad (1)$$

where $\mathbf{g}_t = \mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)$ is the stochastic gradient computed at sample $\mathbf{x}_t$, and $\mathbf{H}_t$ is a (possibly data-dependent) positive definite preconditioning matrix, the convergence conclusion obtained in the previous post was

$$2 \sum_{t=1}^T \eta_t [L(\mathbf{x}_t, \boldsymbol{\theta}_t) - L(\mathbf{x}_t, \boldsymbol{\theta}^*)] \leq R^2 \text{tr}(\mathbf{H}_T) + \sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2 \quad (2)$$

Here $\boldsymbol{\theta}^*$ is the optimum of the loss function, and $R = \max(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|, \dots, \|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|)$ is the maximum distance from the optimization trajectory to the optimum; a more practical interpretation is: assume the optimization process always takes place within a ball of diameter no more than R , which if necessary can be guaranteed via a projection operation.

There is also a key premise for Equation (2) to hold: for any t , $\mathbf{H}_t - \mathbf{H}_{t-1}$ is positive semidefinite, denoted $\mathbf{H}_t \succeq \mathbf{H}_{t-1}$, i.e., $\mathbf{H}_t$ is non-decreasing under the semidefinite order. Under this premise, the previous post further assumed that η_t is a constant η , and obtained

$$\frac{1}{T} \sum_{t=1}^T L(\mathbf{x}_t, \boldsymbol{\theta}_t) - \frac{1}{T} \sum_{t=1}^T L(\mathbf{x}_t, \boldsymbol{\theta}^*) \leq \frac{R^2}{2T\eta} \text{tr}(\mathbf{H}_T) + \frac{\eta}{2T} \sum_{t=1}^T \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2 \quad (3)$$

Then, by minimizing the right-hand side with respect to $\mathbf{H}$, we obtained an understanding of the optimal preconditioning matrix—AdaGrad is precisely the product of this optimization. If we

replace the summation with a moving average, we can derive optimizers such as RMSProp and Adam, but this cannot guarantee $\mathbf{H}_t \succeq \mathbf{H}_{t-1}$, so the above convergence theory no longer holds automatically.

So, there are naturally two questions next: 1. If η_t is not constant, what should the conclusion become? 2. If the monotonicity $\mathbf{H}_t \succeq \mathbf{H}_{t-1}$ does not always hold, what should we do?

2 Average Convergence

Let us first consider a general dynamic learning rate η_t . Fortunately, a dynamic learning rate introduces no difficulty in the proof, and we can reuse almost all of the techniques from the previous posts. First, following the idea of [Making Alchemy More Scientific \(Part 2\): Extending the Conclusions to Unbounded Domains](#), we need to take expectations on both sides of Equation (2), with $\mathbb{E} \triangleq \mathbb{E}_{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T}$:

$$2 \sum_{t=1}^T \eta_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)] \leq \mathbb{E} \left[R^2 \text{tr}(\mathbf{H}_T) + \sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2 \right] \quad (4)$$

A few notes here: 1. We assume that η_t is independent of the data $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t$; 2. Apart from that, almost all variables are data-dependent, e.g., $\boldsymbol{\theta}_t$ depends on $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{t-1}$, and $\mathbf{H}_t$ may also depend on the data (e.g., AdaGrad); 3. $\mathbf{g}_t$ is shorthand for $\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t)$ and is naturally no exception; 4. If R is defined via max, then it is data-dependent, but under the bounded-ball interpretation it is independent, and we usually prefer the latter.

Now dividing both sides by $2 \sum_{t=1}^T \eta_t$, we obtain

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)]}{\sum_{t=1}^T \eta_t} \leq \mathbb{E} \left[\frac{R^2 \text{tr}(\mathbf{H}_T)}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2}{2 \sum_{t=1}^T \eta_t} \right] \quad (5)$$

This gives the weighted-average version of the convergence conclusion for a dynamic learning rate. On the other hand, a more fundamental prototype of inequality (2) is

$$2 \sum_{t=1}^T \eta_t \mathbf{g}_t \cdot (\boldsymbol{\theta}_t - \boldsymbol{\varphi}) \leq R^2 \text{tr}(\mathbf{H}_T) + \sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2 \quad (6)$$

That is, the above is obtained first, and then, assuming the convexity of L and taking $\boldsymbol{\varphi} = \boldsymbol{\theta}^*$, we get Equation (2). Repeating the proof of this section based on (6) yields

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}[\mathbf{g}_t \cdot (\boldsymbol{\theta}_t - \boldsymbol{\varphi})]}{\sum_{t=1}^T \eta_t} \leq \mathbb{E} \left[\frac{R^2 \text{tr}(\mathbf{H}_T)}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2}{2 \sum_{t=1}^T \eta_t} \right] \quad (7)$$

This is the inequality that will be used in the next section. Note that $\mathbf{g}_t$ here can in fact be replaced by an arbitrary vector, because the only role of $\mathbf{g}_t$ being the gradient at the current point is to relate it to the loss function via convexity, but here we are not involving the loss, so as far as this inequality alone is concerned, $\mathbf{g}_t$ is also arbitrary.

3 Last-Iterate Convergence

For practice, last-iterate convergence is often more informative than average convergence. To obtain a last-iterate convergence conclusion, we directly bring in an inequality proven in [Making Alchemy](#)

More Scientific (Part 6): An Ingenious Top-Down Construction:

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \frac{1}{w_{1:T}} \sum_{t=1}^T w_t \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\theta}_t) \cdot (\boldsymbol{\psi}_t - \boldsymbol{\theta}^*)] \quad (8)$$

where $\boldsymbol{\psi}_{t+1} - \boldsymbol{\psi}_t = \frac{w_{1:T}}{w_{t+1:T}}(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$. We assume the reader has followed this series of posts one by one, so most notations will not be reintroduced. The only condition for inequality (8) is the convexity of L ; that is, as long as we assume the convexity of L and the above relationship between $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$, inequality (8) holds, regardless of the specific optimizer.

The role of inequality (8) is very intuitive: it converts last-iterate convergence into average convergence, so that it can connect with the conclusion of the previous section. Now we set $\boldsymbol{\psi}_{t+1} = \boldsymbol{\psi}_t - w_t \mathbf{H}_t^{-1} \mathbf{g}_t$ and take $\boldsymbol{\psi}_1 = \boldsymbol{\theta}_1$; then from the relationship between the two we can derive

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{w_t w_{t+1:T}}{w_{1:T}} \mathbf{H}_t^{-1} \mathbf{g}_t \quad (9)$$

Applying Equation (7) to the right-hand side of inequality (8), we obtain

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \mathbb{E} \left[\frac{R^2 \text{tr}(\mathbf{H}_T)}{2 \sum_{t=1}^T w_t} + \frac{\sum_{t=1}^T w_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2}{2 \sum_{t=1}^T w_t} \right] \quad (10)$$

Strictly speaking, R here should be an upper bound on the distance from the $\boldsymbol{\psi}$ trajectory to $\boldsymbol{\theta}^*$; under the “bounded ball” interpretation, as long as both trajectories stay inside the same ball (guaranteed via projection if necessary), the same R can be used. Thus, as long as the η_t in $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{H}_t^{-1} \mathbf{g}_t$ can be expressed as $\frac{w_t w_{t+1:T}}{w_{1:T}}$, the above convergence conclusion holds, which extends the previous last-iterate convergence conclusion for scalar learning rates to matrix-valued learning rates.

An immediately applicable corollary consistent with practice can be drawn here: according to the experience in *Making Alchemy More Scientific (Part 5): Fine-Tuning the Learning Rate Based on Gradients*, if we assume that $\|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2$ is close to a constant (or is bounded by some constant, further relaxing the right-hand side), then the minimum of the right-hand side is attained when w_t is constant, and constant w_t corresponds to a linear decay of η_t . This affirms the universality of linearly decaying LR schedules for general adaptive-learning-rate optimizers.

Beyond the above implicit conclusion, following the approach of the “Explicit Version” section, we can also obtain an explicit conclusion:

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \mathbb{E} \left[\frac{R^2 \text{tr}(\mathbf{H}_T)}{4\eta_{1:T}} + \sum_{t=1}^{T-2} \frac{\eta_t^2}{4\eta_{t+1:T}} \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2 + \frac{\eta_{T-1}}{2} \left(\|\mathbf{g}_{T-1}\|_{\mathbf{H}_{T-1}^{-1}}^2 + \|\mathbf{g}_T\|_{\mathbf{H}_T^{-1}}^2 \right) \right] \quad (11)$$

4 Guaranteeing Monotonicity

Now let us look at the second question: when the monotonicity $\mathbf{H}_t \succeq \mathbf{H}_{t-1}$ does not always hold, how can we “remedy” it? A naive idea is to make the smallest possible modification to the original optimizer so that it satisfies the desired monotonicity. Specifically, given a positive-semidefinite matrix sequence $\mathbf{H}_1, \mathbf{H}_2, \dots$, we want to find a new sequence $\hat{\mathbf{H}}_1, \hat{\mathbf{H}}_2, \dots$ closest to it, with $\hat{\mathbf{H}}_t \succeq \hat{\mathbf{H}}_{t-1}$ always holding. The new sequence is constructed as

$$\hat{\mathbf{H}}_1 = \mathbf{H}_1, \quad \hat{\mathbf{H}}_t = \hat{\mathbf{H}}_{t-1} + [\mathbf{H}_t - \hat{\mathbf{H}}_{t-1}]_+ \quad (12)$$

where $[\mathbf{S}]_+$ denotes the new matrix obtained by applying $\max(\cdot, 0)$ to all eigenvalues of the real symmetric matrix $\mathbf{S}$. This construction is very intuitive; it is easy to see the inevitability of $\hat{\mathbf{H}}_t \succeq \hat{\mathbf{H}}_{t-1}$, and it can be proven to be optimal under the trace metric.

Proposition: Among all positive-semidefinite matrices $\mathbf{S}$ that simultaneously satisfy $\succeq \hat{\mathbf{H}}_{t-1}$ and $\succeq \mathbf{H}_t$, the above $\hat{\mathbf{H}}_t$ has the smallest trace.

Proof: Let the eigendecomposition of $\mathbf{H}_t - \hat{\mathbf{H}}_{t-1}$ be $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, and denote $\mathbf{N} = \mathbf{V}^\top(\mathbf{S} - \hat{\mathbf{H}}_{t-1})\mathbf{V}$. Then the two conditions give $\mathbf{N} \succeq \mathbf{0}$ and $\mathbf{N} \succeq \mathbf{\Lambda}$ respectively, which in turn imply $N_{i,i} \geq 0$ and $N_{i,i} \geq \Lambda_{i,i}$, or equivalently $N_{i,i} \geq \max(\Lambda_{i,i}, 0)$. Hence

$$\text{tr}(\mathbf{S}) - \text{tr}(\hat{\mathbf{H}}_{t-1}) = \text{tr}(\mathbf{S} - \hat{\mathbf{H}}_{t-1}) = \text{tr}(\mathbf{N}) \geq \text{tr}([\mathbf{H}_t - \hat{\mathbf{H}}_{t-1}]_+) \quad (13)$$

Taking $\mathbf{S} = \hat{\mathbf{H}}_t = \hat{\mathbf{H}}_{t-1} + [\mathbf{H}_t - \hat{\mathbf{H}}_{t-1}]_+$ makes the equality hold exactly. Q.E.D.

Corollary: If $\mathbf{H}_t \succeq \mathbf{H}_{t-1}$ already always holds, then $\hat{\mathbf{H}}_t = \mathbf{H}_t$.

With $\hat{\mathbf{H}}_t$ in hand, we replace the original $\mathbf{H}_t$ with $\hat{\mathbf{H}}_t$, i.e., we change the update rule to

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \hat{\mathbf{H}}_t^{-1} \mathbf{g}_t \quad (14)$$

This guarantees monotonicity, allowing us to reuse the earlier convergence conclusions by simply replacing $\mathbf{H}$ with $\hat{\mathbf{H}}$, while minimizing the modification to the original optimizer. In particular, when $\mathbf{H}_t$ is a diagonal matrix, the update rule for $\hat{\mathbf{H}}_t$ simplifies to the more intuitive

$$\hat{\mathbf{H}}_t = \max(\hat{\mathbf{H}}_{t-1}, \mathbf{H}_t) \quad (15)$$

where $\max$ is taken element-wise. If we use this to replace the second moment of Adam, we obtain the classic [AMSGrad](#) optimizer, which has better convergence properties at the cost of an extra set of state variables. Meanwhile, like AdaGrad, optimizers that guarantee the monotonicity of $\mathbf{H}_t$ often raise concerns of being overly conservative in practice, so AMSGrad in fact never became popular either.

5 Back to the Source

If we insist on not modifying the optimizer, then we can only start from the conclusions. In this section we go back to the source and re-examine the proof process of the previous post, to see where exactly the monotonicity of $\mathbf{H}_t$ was used. The answer is: we used the positive semidefiniteness of $\mathbf{H}_t - \mathbf{H}_{t-1}$ to derive the relaxation $\mathbf{H}_t - \mathbf{H}_{t-1} \preceq \text{tr}(\mathbf{H}_t - \mathbf{H}_{t-1})\mathbf{I}$, and then

$$\|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|_{\mathbf{H}_t - \mathbf{H}_{t-1}}^2 \leq \|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|^2 \text{tr}(\mathbf{H}_t - \mathbf{H}_{t-1}) \leq R^2 \text{tr}(\mathbf{H}_t - \mathbf{H}_{t-1}) \quad (16)$$

Subsequently,

$$\begin{aligned} \sum_{t=1}^T (\|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|_{\mathbf{H}_t}^2 - \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\varphi}\|_{\mathbf{H}_t}^2) &= \|\boldsymbol{\theta}_1 - \boldsymbol{\varphi}\|_{\mathbf{H}_1}^2 - \|\boldsymbol{\theta}_{T+1} - \boldsymbol{\varphi}\|_{\mathbf{H}_T}^2 + \sum_{t=2}^T \|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|_{\mathbf{H}_t - \mathbf{H}_{t-1}}^2 \\ &\leq R^2 \text{tr}(\mathbf{H}_1) - \|\boldsymbol{\theta}_{T+1} - \boldsymbol{\varphi}\|_{\mathbf{H}_T}^2 + \sum_{t=2}^T R^2 \text{tr}(\mathbf{H}_t - \mathbf{H}_{t-1}) \\ &= R^2 \text{tr}(\mathbf{H}_T) \end{aligned} \quad (17)$$

However, the relaxation $\mathbf{H}_t - \mathbf{H}_{t-1} \preceq \text{tr}(\mathbf{H}_t - \mathbf{H}_{t-1})\mathbf{I}$ is actually quite loose and depends on the positive semidefiniteness of $\mathbf{H}_t - \mathbf{H}_{t-1}$. The reason we used it is that tr is a linear function, which allows the final summation to cancel out, leaving the clean $R^2 \text{tr}(\mathbf{H}_T)$, convenient for analysis. In fact, a tighter relaxation that does not require assuming $\mathbf{H}_t \succeq \mathbf{H}_{t-1}$ is

$$\|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|_{\mathbf{H}_t - \mathbf{H}_{t-1}}^2 \leq \|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|^2 \lambda_{\max}(\mathbf{H}_t - \mathbf{H}_{t-1}) \leq R^2 [\lambda_{\max}(\mathbf{H}_t - \mathbf{H}_{t-1})]_+ \quad (18)$$

where $[x]_+ = \max(x, 0)$. This is equivalent to saying $\mathbf{H}_t - \mathbf{H}_{t-1} \preceq [\lambda_{\max}(\mathbf{H}_t - \mathbf{H}_{t-1})]_+ \mathbf{I}$, with $\lambda_{\max}$ denoting the largest eigenvalue. Using this result, we obtain

$$\sum_{t=1}^T (\|\boldsymbol{\theta}_t - \boldsymbol{\varphi}\|_{\mathbf{H}_t}^2 - \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\varphi}\|_{\mathbf{H}_t}^2) \leq R^2 \left(\lambda_{\max}(\mathbf{H}_1) + \sum_{t=2}^T [\lambda_{\max}(\mathbf{H}_t - \mathbf{H}_{t-1})]_+ \right) \quad (19)$$

Note that $\lambda_{\max}$ is not a linear function, so we can no longer cancel further. However, although this result is not as clean as $R^2 \text{tr}(\mathbf{H}_T)$, when monotonicity holds it is in fact never worse than (and usually strictly better than) $R^2 \text{tr}(\mathbf{H}_T)$, so this is actually a stronger and more general result.

6 New Conclusions

Reviewing the entire derivation, the only place where the monotonicity $\mathbf{H}_t \succeq \mathbf{H}_{t-1}$ was used is the relaxation in the previous section. Therefore, after replacing it with the $\lambda_{\max}$ version, all of the previous convergence conclusions can be extended verbatim to the general case; we only need to uniformly replace $R^2 \text{tr}(\mathbf{H}_T)$ by $R^2 \Lambda_T$, where

$$\Lambda_T \triangleq \lambda_{\max}(\mathbf{H}_1) + \sum_{t=2}^T [\lambda_{\max}(\mathbf{H}_t - \mathbf{H}_{t-1})]_+ \quad (20)$$

Specifically, the general version of Equation (7) is:

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}[\mathbf{g}_t \cdot (\boldsymbol{\theta}_t - \boldsymbol{\varphi})]}{\sum_{t=1}^T \eta_t} \leq \mathbb{E} \left[\frac{R^2 \Lambda_T}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2}{2 \sum_{t=1}^T \eta_t} \right] \quad (21)$$

The weighted-average convergence conclusion is

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)]}{\sum_{t=1}^T \eta_t} \leq \mathbb{E} \left[\frac{R^2 \Lambda_T}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2}{2 \sum_{t=1}^T \eta_t} \right] \quad (22)$$

The two last-iterate convergence conclusions are respectively

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \mathbb{E} \left[\frac{R^2 \Lambda_T}{2 \sum_{t=1}^T w_t} + \frac{\sum_{t=1}^T w_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2}{2 \sum_{t=1}^T w_t} \right] \quad (23)$$

$$\mathbb{E}[L(\boldsymbol{\theta}_T) - L(\boldsymbol{\theta}^*)] \leq \mathbb{E} \left[\frac{R^2 \Lambda_T}{4 \eta_{1:T}} + \sum_{t=1}^{T-2} \frac{\eta_t^2}{4 \eta_{t+1:T}} \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2 + \frac{\eta_{T-1}}{2} (\|\mathbf{g}_{T-1}\|_{\mathbf{H}_{T-1}^{-1}}^2 + \|\mathbf{g}_T\|_{\mathbf{H}_T^{-1}}^2) \right] \quad (24)$$

Note that Λ_T has two basic properties: first, clearly Λ_T is always non-negative, so the new bound carries no risk of degeneration; second, when the monotonicity of $\mathbf{H}_t$ holds, we have $[\lambda_{\max}(\mathbf{H}_t - \mathbf{H}_{t-1})]_+ \leq \text{tr}(\mathbf{H}_t - \mathbf{H}_{t-1})$, hence $\Lambda_T \leq \text{tr}(\mathbf{H}_T)$, i.e., the new conclusion indeed subsumes the old one.

7 Further Reflections

At this point, the second scaffold has also been dismantled. Seen this way, the non-monotonicity of $\mathbf{H}_t$ does not substantively impede the convergence proof. Does this mean that optimizers such as RMSProp and Adam have no convergence problems?

Not so. Being able to write down a convergence conclusion does not mean convergence is guaranteed. Taking Equation (22) as an example, if we regard R and $\|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2$ as constants independent of t and T , then the minimum of the right-hand side is attained when η_t is constant; in that case $\eta_{1:T}$ grows linearly with T , and if Λ_T also grows linearly or superlinearly with T , the right-hand side cannot tend to zero as $T \rightarrow \infty$, i.e., convergence cannot be guaranteed.

From this argument we can see that it is actually not directly related to whether $\mathbf{H}_t$ is monotonic; as long as the growth rate of Λ_T with respect to T is at least linear, non-convergence phenomena may appear. For AdaGrad, it can be shown that under the assumption of bounded gradients, $\Lambda_T \leq \text{tr}(\mathbf{H}_T) \sim \sqrt{T}$, i.e., Λ_T grows at most at an $\mathcal{O}(\sqrt{T})$ rate, below linear, so convergence is guaranteed.

But for moving-average-type preconditioners of optimizers like RMSProp and Adam, although $\mathbf{H}_t$ itself is smoother, the eigenvalues of $\mathbf{H}_t - \mathbf{H}_{t-1}$ may repeatedly alternate in sign, while Λ_T accumulates the $[\lambda_{\max}]_+$ of each step’s change. In extreme cases, if the changes contain a persistent, stable positive direction, then it grows at a constant rate, yielding a linear Λ_T , leading to failure of convergence.

However, thinking in reverse, this to some extent also justifies why RMSProp, Adam, etc. are effective in practice, because generally the growth rate of Λ_T is not that bad; on the contrary, due to the smoothness of $\mathbf{H}_t$, their Λ_T may grow even more gently than AdaGrad’s, resulting in better practical performance. And most importantly, we are only discussing the convergence of the upper bound on the loss; non-convergence of the upper bound in principle does not allow us to conclude that the loss does not converge.

8 Summary

Building on the convergence framework of the previous post, this post removed two “scaffolds”: first, allowing the learning rate η_t to change dynamically, deriving the corresponding average and last-iterate convergence conclusions, and preliminarily explaining the universality of LR decay for optimizers; second, for non-monotonic $\mathbf{H}_t$, discussing a correction scheme based on the idea of minimal modification, and deriving more general convergence conclusions that require no monotonicity assumption.

Overall, this post is dense with information, and we skipped quite a few proof details, because the key proof techniques were already given in detail in the earlier series of posts on SGD, so this post mostly presents the key steps and results directly.

When reposting, please include the URL of this post: <https://kexue.fm/archives/11885>

For more detailed reposting matters, please refer to: [Scientific Spaces FAQ](#)