

Why Cosine Learning Rate Scheduler Works and How to Improve It?

Anonymous Author(s)

Affiliation

Address

email

1 An Additional Theorem for Similarity

Theorem 1. Let objective function $f(x)$ be quadratic and Assumption (1.7) hold. We choose SGD to optimize $f(x)$. If there is an upper bound for $\mathbb{E}[f(w_{T^{\text{eigen}}+1}) - f(w_*)] \leq F(T^{\text{eigen}})$ with eigencurve's learning rate sequence $\{\eta_t^{\text{eigen}}\}_{t=0}^{T^{\text{eigen}}}$, and there exists constant $C_{\text{scale}} \geq 1, C_{\text{tail}} \geq 0, C_{\text{iter}} \in (0, 1]$, such that for any $0 \leq t \leq T^{\text{eigen}}$, cosine decay's learning rate sequence $\{\eta_t^{\text{cos}}\}_{t=0}^{T^{\text{cos}}}$ satisfies,

1. Scale constant: $\eta_t^{\text{cos}} \leq C_{\text{scale}} \cdot \eta_t^{\text{eigen}}$
2. Tail constant: $\sum_{k=t}^{T^{\text{eigen}}} \eta_k^{\text{cos}} \geq (\sum_{k=t}^{T^{\text{eigen}}} \eta_k^{\text{eigen}}) - C_{\text{tail}}$
3. Iteration constant: $T^{\text{eigen}} = C_{\text{iter}} \cdot T^{\text{cos}}$, where $0 < C_{\text{iter}} \leq 1$,

then there is an upper bound for $\mathbb{E}[f(w_{T^{\text{cos}}+1}) - f(w_*)] \leq C \cdot F(T^{\text{cos}})$ with cosine decay's learning rate sequence η_t^{cos} for $0 \leq t \leq T^{\text{cos}}$, where the extra constant $C \leq C_{\text{scale}}^2 / C_{\text{iter}}^2 \cdot \exp(2C_{\text{tail}}L)$.

Proof. For dimension j , the key quantity of the quadratic loss is

$$\text{Loss}_j = \text{Bias}_j + \text{Variance}_j \leq \text{Bias}_j^{\text{upper}} + \text{Variance}_j^{\text{upper}} \quad (1.1)$$

$$\begin{aligned} \text{Bias}_j^{\text{upper}} &= \lambda_j \cdot (u_j^\top (w_0 - w_*))^2 \cdot \exp\left(\sum_{k=0}^T -2\eta_k \lambda_j\right) \\ &= c_j^{(1)} \cdot \exp\left(\sum_{k=0}^T -2\eta_k \lambda_j\right) \end{aligned} \quad (1.2)$$

$$\begin{aligned} \text{Variance}_j^{\text{upper}} &= \lambda_j^2 \sigma^2 \sum_{t=0}^T \eta_t^2 \exp\left(\sum_{k=t+1}^T -2\eta_k \lambda_j\right) \\ &= c_j^{(2)} \sum_{t=0}^T \eta_t^2 \exp\left(\sum_{k=t+1}^T -2\eta_k \lambda_j\right) \end{aligned} \quad (1.3)$$

Notice that the key terms here are η_t^2 and $\sum_{k=t+1}^T \eta_k$. We call the previous one as ‘‘Scale’’, and the later one as ‘‘Tail’’. The whole convergence guarantee of eigencurve is established based on those two terms.

To make the proof simpler, we first start with some simple case by setting $C_{\text{iter}} = 1$ and considering the scale constant and tail constant only. With a new scheduler that satisfying the first two constant constraint, i.e.

$$\begin{aligned}\eta'_t &\leq C_{scale} \cdot \eta_t \\ \sum_{k=t}^T \eta'_k &\geq \left(\sum_{k=t}^T \eta_k \right) - C_{tail}\end{aligned}$$

18 Now we have the updated bias and variance term,

$$\begin{aligned}\text{Bias}_j^{\text{upper}} &= c_j^{(1)} \cdot \exp \left(\sum_{k=0}^T -2\eta'_k \lambda_j \right) \\ &= c_j^{(1)} \cdot \exp \left(-2\lambda_j \cdot \sum_{k=0}^T \eta'_k \right) \\ &\leq c_j^{(1)} \cdot \exp \left(-2\lambda_j \cdot \sum_{k=0}^T \eta_k + 2C_{tail} \lambda_j \right) \\ &= \text{Bias}_j^{\text{upper}} \cdot \exp(2C_{tail} \lambda_j) \\ &= \text{Bias}_j^{\text{upper}} \cdot \exp(2C_{tail} L) \\ \text{Variance}_j^{\text{upper}} &= c_j^{(2)} \sum_{t=0}^T \eta_t^2 \exp \left(\sum_{k=t+1}^T -2\eta'_k \lambda_j \right) \\ &\leq c_j^{(2)} \sum_{t=0}^T (C_{scale} \eta_t)^2 \exp \left(\sum_{k=t+1}^T -2\eta_k \lambda_j \right) \cdot \exp(2C_{tail} \lambda_j) \\ &= \text{Variance}_j^{\text{upper}} \cdot C_{scale}^2 \cdot \exp(2C_{tail} \lambda_j) \\ &\leq \text{Variance}_j^{\text{upper}} \cdot C_{scale}^2 \cdot \exp(2C_{tail} L)\end{aligned}$$

19 Intuitively, smaller C_{scale} means the newly generated variance is smaller. Smaller C_{tail} means the
20 power of reducing variance is stronger. In fact the whole theoretical analysis is the tradeoff between
21 those two terms.

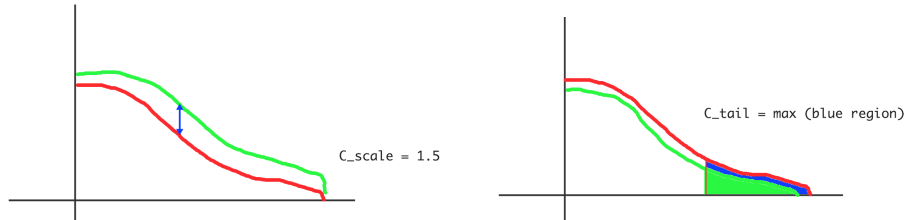


Figure 1: Intuitive explanation of C_{scale} and C_{tail} , where red curve means the eigencurve (the curve with theoretical guarantees), the green curve means the curve of cosine decay (the curve with similar shape but without theoretical guarantees)

22 However, those two constant are not sufficient to yield meaningful constant that is small enough to
23 bound cosine decay. In practice, we observed that the last part of eigencurve is obviously “higher”
24 than cosine decay. That is the reason why we introduce the iteration constant C_{iter} , to reduce the
25 effect of this last part.

26 The motivation is that, all previous analysis are conducted in y-axis. We may now “compress” the
27 eigencurve in x-axis. For example, by setting $T' = 0.9T$, eigencurve can achieve a loss that is only
28 $1/0.9^2$ from the original upper bound, since eigencurve’s convergence rate is at least $1/T^2$.

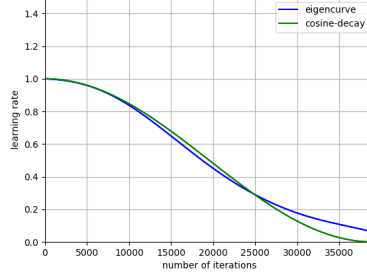


Figure 2: Comparison of original eigencurve and cosine decay's curve

$$\mathbb{E}[f(w_{T+1}) - f(w_*)] \leq (f(w_0) - f(w_*)) \cdot \frac{\kappa^2 \cdot \left(\sum_{i=0}^{I_{\max}-1} \sqrt{s_i}\right)^2}{s_0 T^2} + \frac{15 \left(\sum_{i=0}^{I_{\max}-1} \sqrt{s_i}\right)^2}{T} \cdot \sigma^2. \quad (1.4)$$

29 Then this high part of eigencurve can be moved to compare with the forefront of cosine decay.

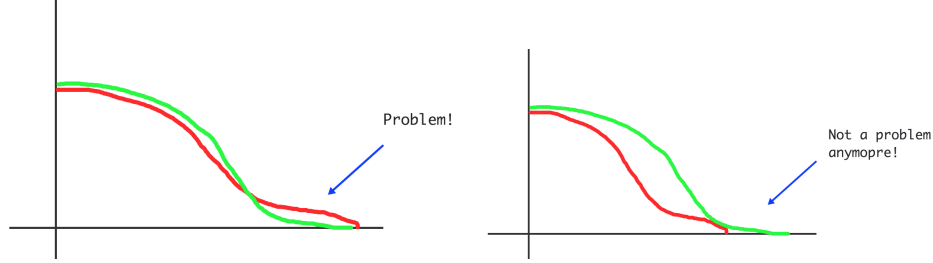


Figure 3: Intuitive explanation of C_{iter} , where red curve means the eigencurve (the curve with theoretical guarantees), the green curve means the curve of cosine decay (the curve with similar shape but without theoretical guarantees)

30 In fact, a new scheduler is generated based eigencurve, which takes $T' = C_{iter}T$ iterations, and
 31 leaves the remaining iterations with $\eta_t = 0$.

32 By comparing cosine decay with this new scheduler in $t \in [0, T']$, cosine decay can achieve the
 33 same theoretical guarantee as this new scheduler at the point of T' , with an extra constant of
 34 $C_{scale}^2 \cdot \exp(2C_{tail}L)$.

35 This new scheduler is only $1/C_{iter}^2$ away from the original eigencurve scheduler, this yield our final
 36 constant

$$C \leq C_{scale}^2 / C_{iter}^2 \cdot \exp(2C_{tail}L). \quad (1.5)$$

37 Last problem in theory. This is only the theoretical guarantee for cosine decay at iteration $t = T' =$
 38 $C_{iter}T$, but not the final iterate. So further analysis is required.

39 First, the bias term is decreasing with more iterations. As for the variance term, according to
 40 Lemma 10 in the Appendix of our current version of paper, it is either decreasing, or at most $\eta'_{T'}/\lambda_j$.
 41 By taking the scale constant into account, when comparing with eigencurve, this term is at most
 42 $\eta'_{T'}/\lambda_j \leq C_{scale} \cdot \eta_{T'}/\lambda_j$. This term is negligible, since according to the proof of Lemma 12 in the
 43 Appendix, the dominant term in eigencurve is $\eta_{t_{i+1}+1}/\lambda_j \geq \eta_{T'}/\lambda_j$.

44 Now take the similarity constant $C \leq C_{scale}^2 / C_{iter}^2 \cdot \exp(2C_{tail}L)$ into account, the dominant term
 45 becomes at least $C\eta_{T'}/\lambda_j \geq C_{scale}^2 \eta_{T'}/\lambda_j \geq C_{scale} \eta_{T'} \lambda_j$.

46 This means that the last part of cosine decay either improves the loss, or makes the loss worse but is
 47 still bounded by eigencurve’s convergence rate, with an extra constant at most C .

48 □

49 **Corollary 2.** Let objective function $f(x)$ be quadratic and Assumption (1.7) hold. We choose SGD
 50 to optimize $f(x)$. If the Hessian eigenvalue distribution is the same as our estimated Resnet18
 51 distribution on CIFAR-10, $19550 \leq T \leq 78200$, i.e. 50-200 epochs with batch size 128, and
 52 $\eta_0 = 1/L, \eta_{min} = 0$ holds for cosine learning rate scheduler, then cosine decay enjoys the same
 53 convergence rate as eigencurve, with an extra constant factor of $C < 10$.

54 *Proof.* This proof is done by empirically comparing the learning rate curve of eigencurve and
 55 cosine decay with proper choice of $C_{scale}, C_{tail}, C_{iter}$. Since we notice that the constant $C \leq$
 56 $C_{scale}^2/C_{iter}^2 \cdot \exp(2C_{tail}L)$ ’s dependence on C_{tail} is exponential, normally we set this constant to
 57 0. In addition, with fixed C_{iter} , the minimum C_{scale} can be computed directly. So the only tunable
 58 parameter is C_{iter} .

59 The following table shows the empirical constant we measures at certain points.

$T \in$	C_{iter}	$C_{scale} \in$	$C \in$
[19550, 30000]	0.77	(1.75, 2.42)	(5.22, 9.86)
[30000, 45000]	0.81	(1.77, 2.52)	(4.79, 9.63)
[45000, 65000]	0.84	(1.83, 2.56)	(4.76, 9.28)
[65000, 78200]	0.87	(1.82, 2.10)	(4.41, 5.83)

Table 1: Constant for different T

60 Notice that here for every fixed C_{iter} , we only need to measure the left/right end point for each
 61 iteraval and ensure $C < 10$, since the shape eigencurve is continuous changing (precisely speaking,
 62 the value of $\eta_{p,T}$ for fixed p is monotonically decreasing with increasing value of T), so the value of
 63 C_{scale} is also continuous changing inside those ranges of T .

64 In all the above ranges, we have $C < 10$, this proves the corollary.

65 □