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 $\mathbb{E}[f(w_{T^{eigen}+1}) - f(w_*)] \leq F(T^{eigen})$ with eigencurve's learning rate sequence $\{\eta_t^{eigen}\}_{t=0}^{T^{eigehn}}$, and there exists constant $C_{scale} \geq 1, C_{tail} \geq 0, C_{iter} \in (0,1]$, such that for any $0 \leq t \leq T^{eigen}$, cosine decay's learning rate sequence $\{\eta_t^{cos}\}_{t=0}^{T^{cos}}$ satisifies,*

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

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

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

$$\text{Loss}_j = \text{Bias}_j + \text{Variance}_j \leq \text{Bias}_j^{\text{upper}} + \text{Variance}_j^{\text{upper}} \tag{1.1}$$

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

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

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 $C_{iter} = 1$ and considering the scale constant and tail constant only. With a new scheduler that satisfying the first two constant constraint, i.e.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

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

Now we have the updated bias and variance term,

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

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

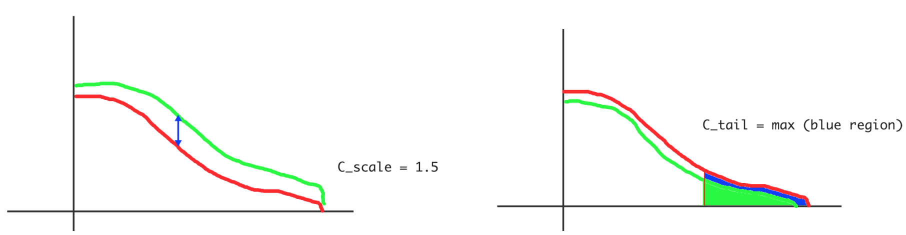

Figure 1: Intuitive explanation of C_{scale} and C_{tail}, where red curve means the eigencurve (the curve with theoretical guarantees), the green curve means the curve of cosine decay (the curve with similar shape but without theoretical guarantees)

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

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

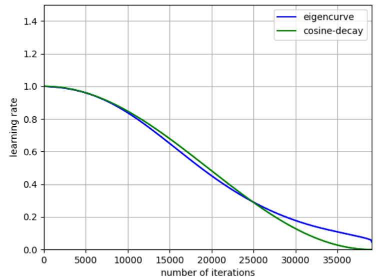

Figure 2: Comparison of original eigencurve and cosine decay's curve

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

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

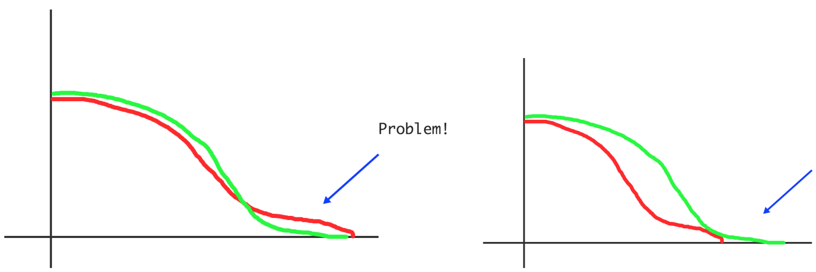

Figure 3: Intuitive explanation of C_{iter}, where red curve means the eigencurve (the curve with theoretical guarantees), the green curve means the curve of cosine decay (the curve with similar shape but without theoretical guarantees)

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

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

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

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

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

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

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

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

$\square$

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

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

The following table shows the empircal 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

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

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

$\square$