Stochastic Gradient Descent
with Hyperbolic-Tangent Decay

Bo Yang Hsueh
Department of Computer Science
National Chiao Tung University
byshiue@gmail.com

Wei Li
Department of Computer Science
National Chiao Tung University
fm.bigballon@gmail.com

I-Chen Wu
Department of Computer Science
National Chiao Tung University
icwu@cs.nctu.edu.tw

Abstract

Learning rate scheduler has been a critical issue in the deep neural network training. Several schedulers and methods have been proposed, including step decay scheduler, adaptive method, cosine scheduler and cyclical scheduler. This paper proposes a new scheduling method, named hyperbolic-tangent decay (HTD). We run experiments on several benchmarks such as: ResNet, Wide ResNet and DenseNet for CIFAR-10 and CIFAR-100 datasets, LSTM for PAMAP2 dataset, ResNet on ImageNet and Fashion-MNIST datasets. In our experiments, HTD outperforms step decay and cosine scheduler in nearly all cases, while requiring less hyperparameters than step decay, and more flexible than cosine scheduler.

1 Introduction

Deep Neural Networks (DNNs) are currently the best-performing method for many classification problems. The variants of DNN have significant performance on many areas. For example, Convolutional Neural Network (CNN) [15] is widely used in image classification, object localization and detection. Recurrent Neural Network (RNN) [8] is widely used in language translation and natural language processing.

Training DNN is usually considered as the non-convex optimization problem. Stochastic gradient descent (SGD) is one of the most used training algorithms for DNN. Although there are many different optimizers like Newton and Quasi-Newton methods [19] in tradition, these methods are hard to implement and need to handle the problem of large cost on computing and storage. Compared to them, SGD is simpler and has good performance. The update direction of SGD is determined by the gradient of loss function. The parameters θ_t (weights) at time t are updated by $\theta_t = \theta_{t-1} - \alpha_t \nabla_\theta L$, where L is a loss function and α_t is the learning rate at time t.

Unfortunately, it has been hard to tune the learning rate. A large learning rate makes the training diverge, while a small learning rate makes the training converge slowly. For a better performance, one usually needs to experiment with a variety of learning rates during the training process. A method of scheduling leaning rate, called a *learning rate scheduler* in this paper, is used to change the learning rate during the training progress.

There are many different schedulers used in the past, including step decay, adaptive learning rate methods [2, 13, 26, 29], SGDR [17], and so on. Step decay can get the ideal results in theory, but the process of tuning the learning rate is tedious and time-consuming. Adaptive methods can

adjust the step size of each parameter by themselves, but the final performance are usually worse than fine-tuned step decay. SGDR used the cosine function to perform cyclic learning rates. Pure cosine scheduler (without cyclic) is a special case of SGDR, but performs better than step decay and cyclic cosine scheduler in about half of their experiments. However, users are only able to adjust the maximum and minimum learning rate of cosine scheduler, which is less flexible than step decay.

This paper proposes a new learning rate scheduler, called as Hyperbolic-Tangent decay (HTD) scheduler. Compared to step decay scheduler, HTD has less hyperparameters to tune and performs better than step decay scheduler in all experiments. Compared to cosine scheduler, HTD requires slightly more hyperparameters to tune for higher performance, and outperforms cosine schedulers in nearly all experiments in this paper.

Section 2 reviews some optimizers and learning rate schedulers proposed in the past. Section 3 describes the proposed HTD scheduler. Section 4 shows experimental results of HTD against other learning rate schedulers on different architectures and datasets. Section 5 concludes the contributions of this paper and discusses the future work.

2   Related work

This section first reviews some optimizers and learning rate scheduler proposed in the past. Then, we review the traditional optimizers like stochastic gradient descent (SGD), SGD with momentum [21], and Nesterov momentum [18, 25]. Next, we review the adaptive learning optimizers including Adam, AdaDelta and so on [2, 13, 26, 29]. Finally, we review learning rate schedulers proposed last few years, including stochastic gradient descent with warm restarts (SGDR) [17], cyclical learning rate (CLR) [24] and exponential decay sine wave learning rate (ES-Learning) [1].

SGD is one of the most popular optimizers for training DNNs. The so-called *step decay learning rate scheduler* is usually used in SGD to change learning rates to specific values for different stages. For example, the step decay scheduler in [6] scheduled the learning rate as follows:

$$lr = \begin{cases} 0.1 & 0 < e \leq 81 \\ 0.01 & 81 < e \leq 122 \\ 0.001 & 122 < e \leq 200 \end{cases} \tag{1}$$

where e refers to the index of the current epoch. Ideally, the step decay scheduler can achieve a good performance by carefully changing the learning rate and stage periods, but trial and error is needed to find an acceptable step decay scheduler. Another traditional scheduler is the *exponential decay scheduler*. The exponential decay scheduler reduces the learning rate by a given factor at each iteration or epoch. The learning rate formula for the exponential decay scheduler is:

$$lr_t = lr_0 \times \lambda^t \tag{2}$$

where lr_t refers to the learning rate at time t, lr_0 is the initial learning rate, $\lambda \in [0, 1]$ is a *discount factor*. Figure 2 illustrates the learning rate curves of step decay scheduler and exponential decay scheduler.

Momentum [21] is designed to accelerate DNN training. The momentum algorithm records the gradients in the past iterations, and combines them with the current gradient to decide where to move in this iteration. We need to decide the hyperparameter $\beta \in [0, 1]$ in the momentum algorithm to determine the contributions of past gradients to the current update. Nesterov Momentum, a variant of the momentum algorithm, is proposed by Nesterov [18], and used to train DNN by Sutskever, Martens, Dahl and Hinton [25]. The gradient when using Nesterov Momentum is evaluated after applying the current velocity. It can add a correction factor to the standard momentum algorithm.

Adaptive learning rate methods usually adjust weights in a DNN based on a mechanism which requires users to determine some hyperparameters initially, but not for the overall design of the learning scheduler. There have been many different adaptive learning rate methods, including Ada-Grad [2], RMSProp [26], AdaDelta [29], Adam [13] and Adamax [13]. These adaptive learning rate methods do not require hyperparameters fine-tuning to obtain proper learning rate value, this comes at a significant computing cost, while the final performance tends to be inferior to fine-tuned step decay. Shirish and Richard proposed a method where the learning rate is defined with Adam from the start, and by SGD towards the end [12], combining the benefits of a fast convergent rate and with the superior performance of step decay. These kinds of methods reduce the complexity of setting the learning rate scheduler, but the performances are worse than step decay in most cases.
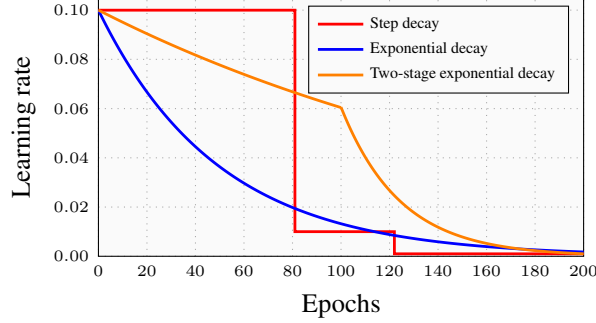
Figure 2: Comparison of exponential decay with $\lambda = 0.98$, step decay and two-stage exponential decay with $\lambda_1 = 0.995, \lambda_2 = 0.96$. It is obvious that the reduce speed of exponential decay is more higher than step decay initially.

SGDR [17], CLR [24] and ES-Learning [1] are similar. They all use warm restart mechanisms to reset the learning rate every some epochs (or iterations) and demonstrated that these mechanisms improved the performances. In [17], SGDR also proposed a *cosine learning rate scheduler*, which follows the cosine wave from the maximum to the minimum and makes the change of learning rate smooth. Cosine scheduler schedules the learning rate as follow:

$$lr_t^{\cos} = lr_{\min} + \frac{lr_{\max} - lr_{\min}}{2}\left(1 + \cos\left(\frac{\pi \times t}{T}\right)\right) \qquad (3)$$

where $lr_{\min}$ and $lr_{\max}$ are the minimum and maximum learning rate respectively, T is the total number of training epochs or iterations, and t is the index of the current epoch or iteration. Figure 4 illustrate the cosine with $lr_{\min} = 0, lr_{\max} = 0.1$. In [9, 10, 17, 20], cosine learning rate schedulers in outperformed step decay scheduler. However, the cosine learning rate scheduler is less flexible than step decay since only the maximum and minimum learning rates can be changed.

3 Hyperbolic-Tangent Decay

This section proposes a new scheduling method, named *hyperbolic-tangent decay* (*HTD*). Subsection 3.1 first analyzes the performance of step decay with different settings, and Subsection 3.2 analyzes the performance of exponential decay and compares it with step decay. The analyses of Subsection 3.1 and Subsection 3.2 motivate the design of HTD described in Subsection 3.3

3.1 Step decay

This subsection presents the performance analysis of step decay. We train Residual Network with 32 layers, denoted by ResNet-32, on CIFAR-10 with the following settings. The learning rate is 0.1 for the first S_1 epochs and 0.01 for the next S_2 epochs. For comparison, use different S_1 and S_2 such that $S_1 + S_2 = 200$. Then compare their averaged test error rates of 4 runs.

Figure 1: Test error rates of different ratios S_1/S_2 with $S_1 + S_2 = 200$. The performances are significantly different for different ratios.

The error rates of different ratios S_1/S_2, shown in Figure 1, indicate that the performances vary significantly as the ratio changes. In this experiment, the performance at the ratio 0.25 performs better than the one at 4 by reducing 1.38% error rate. This result implies that it is important for a scheduler to choose the ratio flexibly for training.
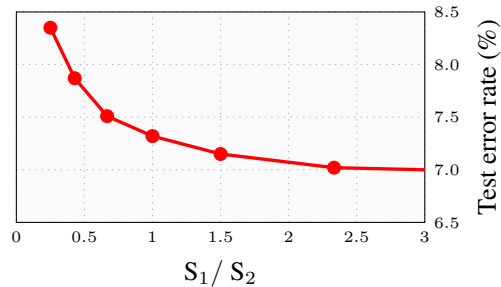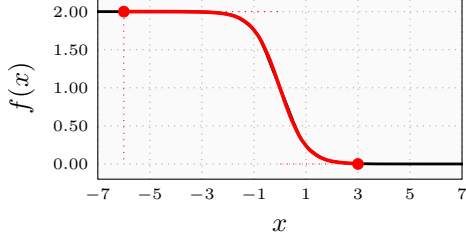
3

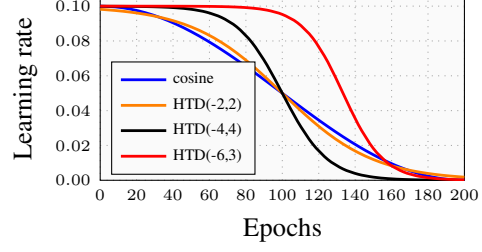Figure 3: Function of $f(x) = 1 - \tanh x$ with $L = -6$ and $U = 3$.

Figure 4: Learning rate curves of HTD with different L and U, and learning rate curve of cosine scheduler.

3.2 Exponential decay

This subsection analyzes the performance of exponential decay and compares it with step decay by training ResNet-32 on CIFAR-10 with 200 epochs and initial learning rate 0.1 as the previous subsection. For step decay, the learning rate is dropped by a factor of 0.1 at 81 and 122 epochs, like [6]. For exponential decay, the discount factor is 0.98, which makes the final learning rates of both exponential decay and step decay close. Each method runs 5 times and computes the averaged test error rates. Consequently, the averaged test error rates of exponential decay and step decay are 7.51% and 7.18% respectively in the experiments. Obviously, step decay performs better than exponential decay in this case.

For further investigation, we try another experiment by proposing two-stage exponential decay as follows.

$$
lr = \begin{cases} lr_0 \times \lambda_1^e & e \le 100 \\ lr_0 \times \lambda_1^{100} \times \lambda_2^{e-100} & e > 100 \end{cases} \tag{4}
$$

where λ_1 is 0.995 and λ_2 is 0.96 are two discount factors in this experiment, such that the final learning rates of step decay, exponential decay and two-stage exponential decay are close. Figure 2 illustrates the learning rate curve of two-stage exponential decay. The figure shows that the learning rate curve of the first stage is quite flat while the decay speed for the second stage is faster. In this experiment, the averaged test error rates is 7.16%, interestingly even lower than step decay. This result implies that the initial decay speed should not be too high.

3.3 Hyperbolic-Tangent Decay (HTD)

From the empirical results in the previous subsections, this subsection proposes to use hyperbolic tangent functions for our learning rate scheduling as follow.

$$
\begin{aligned}
lr_t^{\text{HTD}} &= lr_{\min} + \frac{lr_{\max} - lr_{\min}}{2} \left(1 - \tanh \left(L + (U - L)\frac{t}{T} \right) \right) \\
&= lr_{\min} + \frac{lr_{\max} - lr_{\min}}{2} \left(1 - \tanh \left(L \left(1 - \frac{t}{T} \right) + U\frac{t}{T} \right) \right)
\end{aligned} \tag{5}
$$

where $lr_{\max}$ and $lr_{\min}$ are the maximum and minimum learning rates respectively, T is the total number of epochs (or iterations), $0 \le t \le T$ is the index of epoch (or iteration), L and U indicates the lower and upper bounds of the interval $[L, U]$ for the function $\tanh x$. Figure 3 illustrates the function $1 - \tanh x$ in the interval $[-6, 3]$. Let HTD($L, U, lr_{\max}, lr_{\min}$) denote the scheduler following the Formula 5. In this paper, $lr_{\min}$ is set to 0 and $lr_{\max}$ is set to the initial learning rate of step decay. For simplicity, let HTD(L, U) denote the scheduler. In this paper, we only consider $L \le 0$ and $U > 0$. Figure 4 illustrates the two learning rate curves of HTD(-4,4) and HTD(-6,3).

Close to two-stage exponential decay. Next, we show that HTD is close to the two-stage exponential decay. For simplicity, we use function $1 - \tanh x$ to approximate. Note that $x = 0$ is an inflection point of the function $1 - \tanh x$. We show that (a) the value of $1 - \tanh x$ decreases slowly before the inflection point, like the first stage of two-stage exponential decay, and (b) the value of $1 - \tanh x$ drops like exponential decay after the inflection point, like the second stage of two-stage
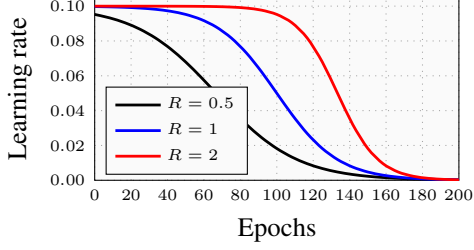
4

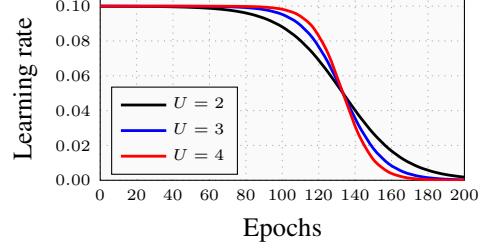Figure 5: Different R with $U = 3$. These curves have same final learning rates but different trends

Figure 6: Different U with $R = 2$. These curves have similar trends but different final learning rates.

exponential decay. Given displacement $\delta > 0$, we define the decreasing ratio of $1 - \tanh x$ as:

$$r(x, \delta) := \frac{1 - \tanh(x + \delta)}{1 - \tanh x} = \frac{2e^{-x-\delta}}{e^{x+\delta} + e^{-x-\delta}} \frac{e^x + e^{-x}}{2e^{-x}}$$
$$= \frac{e^{-\delta}(e^x + e^{-x})}{e^{x+\delta} + e^{-x-\delta}} = \frac{e^{x-\delta} + e^{-x-\delta}}{e^{x+\delta} + e^{-x-\delta}} = \frac{e^{2x} + 1}{e^{2x+2\delta} + 1} \quad (6)$$

for any position $x \in \mathbb{R}$. We observe the decreasing ratio for different position $x \in \mathbb{R}$. For (a), since $e^{2x+2\delta} \approx 0$ and $e^{2x} \approx 0$ when $x < 0$ is sufficiently small, the decreasing ratio $r(x, \delta) \approx 1$ and hence decreasing speed is very slow, like the first stage of two-stage exponential decay. For (b), since $r(x, \delta) \approx \frac{1}{e^{2\delta}}$ when $x > 0$ is sufficiently large, the decreasing speed is close to exponential decay, like the second stage of two-stage exponential decay.

Importance of *L, U* and their ratio. In HTD(L, U), the hyperparameters U affects the final learning rate. For example, when $U = 3$, the final learning rate is $lr_{\max} \cdot (1 - \tanh 3) \approx lr_{\max} \cdot 0.005$. Besides, the ratio $R = |L|/|U|$ indicates the ratio of training times before and after the inflection point. Like Subsection 3.1, when the ratio R is larger, the training time before the inflection point is longer. Figure 5 illustrates the learning rate curves with same U but different R. Figure 6 illustrates the learning rate curves with same R but different U.

Compared to cosine. The hyperparameters of cosine scheduler include $lr_{\max} and lr_{\min}$. while the hyperparameters of HTD include $lr_{\max}, lr_{\min}, L$ and U. The additional hyperparameters of HTD are L and U, which can determine the ratio of training time before and after the inflection point. These additional hyperparameters make HTD more flexible than cosine scheduler. We also find that HTD(-2,2) is close to the cosine scheduler, which is illustrated in Figure 4.

A variation of HTD without *U* and *T*. Let us convert Formula 5 to another form:

$$lr_t^{\text{HTD}} = lr_{\min} + \frac{lr_{\max} - lr_{\min}}{2} \left(1 - \tanh\left(L + \frac{t}{C}\right)\right) \quad (7)$$

where C is a hyperparameter to determine the speed of decay. In this variation, HTD does not need to fix the total number of epoch, i.e., T, and therefore U is not required neither. Hence, users only need to determine L and the decreasing speed C, like the discount factor λ in Formula 2. This variation is useful when users do not know how many T is in advance, e.g., stop training when the loss is lower than a threshold.

4 Experiments

We empirically demonstrate HTD's effectiveness on several benchmark datasets and networks by comparing HTD with step decay scheduler and cosine scheduler. The code and the details of experiments will be available on the Internet.

4.1 Datasets

CIFAR. Two CIFAR datasets [14], CIFAR-10 and CIFAR-100, consist of 60,000 color images with 32×32 pixels, 50,000 for training and 10,000 for testing. CIFAR-10 images are classified into

10 classes with labels and CIFAR-100 images are 100 classes. For image preprocessing, we normalize the input data using the means and standard deviations based on [28]. For data augmentation, we perform random crops from the image padded by 4 pixels on each side, filling missing pixels with reflections of the original image and horizontal flips with 50% probability.

Fashion-MNIST. Fashion-MNIST [27] is a MNIST-like [16] dataset, which consists of 70,000 gray-scale images with 28×28 pixels, classified into 10 categories each with 7,000 images. There are 60,000 and 10,000 images for training and testing respectively. For image preprocessing, we normalize the input data using the mean and standard deviation based on [28]. For data augmentation, we also perform random crops from the image padded by 3 pixels on each side, filling missing pixels with reflections of the original image and horizontal flips with 50% probability.

PAMAP2. PAMAP2 [22] is a physical activity monitoring dataset, and consists of raw input recorded by sensor. It has 9 subjects, 8 males and 1 females with ages between 23 and 31. We only choose 12 (lie, sit, stand, iron, vacuum, ascend stairs, descend stairs, normal walk, nordic walk, cycle, run, and rope jump) from all 18 activities. For data preprocessing, we down sample data from 100Hz to 20Hz and choose all of data without data missing. 70% of data are used to train, and 30% of data are used to test.

ImageNet. ILSVRC dataset [23] consists of 1000 classes and is split into three sets: 1.28 million training images, 50,000 validation images and 50,000 testing images. Each image is a 224×224 color image. Image preprocessing and data augmentation follow the settings of [6]. We adopt 1-crop and evaluate both top-1 and top-5 error rates in testing.

4.2 Implementation

Architectures. We run experiments on several benchmarks including Residual Network (ResNet) [6, 7], Wide Residual Network (Wide ResNet) [28] and Densely Network (DenseNet) [11]. Let ResNet-d denote the ResNet with depth d, WRN-d-k denote the Wide ResNet with depth d and width k, and DenseNet-BC-L-k denote the DenseNet with depth L and growth rate k. For DenseNet, BC represents the use of bottleneck mechanism with 50% reduced rate. For PAMAP2, we train with a Bidirectional-LSTM (BLSTM) [3] based on [4].

Optimizer. All the networks are trained by SGD, using a Nesterov momentum [25] of 0.9 and adopt the weight initialization introduced by [5]. DenseNet is trained with a mini-batch size of 64, ResNet and Wide ResNet are trained with a mini-batch size of 128 except ImageNet, which is trained with a mini-batch size of 256. For ResNet, we use a weight decay of 0.0001, and start with a learning rate of 0.1, reduced by a factor of 0.1 at 81 and 122 epochs in step decay. For Wide ResNet, we use a weight decay of 0.0005 according to [28], and start with a learning rate of 0.1, reduced by a factor of 0.2 at 60, 120 and 160 epochs in step decay. For DenseNet, we use weight decay of 0.0001, and start with a initial learning rate 0.01, reduced by a factor of 0.1 at 50% and 75% of the total number of training epochs in step decay. ResNet and Wide ResNet are trained with 200 epochs, while DenseNet is trained with 300 epochs.

Methods. We compares three types of schedulers in following experiments: step decay, cosine scheduler (proposed by [17]) and HTD. Both cosine and HTD set the minimal learning rate to 0, and use the initial learning rate of step decay for the maximum learning rate. For HTD, we use the two versions HTD(-4,4) and HTD(-6,3) only.

4.3 Experiments result

4.3.1 Different Networks for CIFAR

For CIFAR-10 and CIFAR-100, this paper tests them on ResNet-110, WRN-28-10, DenseNet-BC-100-12 and DenseNet-BC-250-24. For DenseNet, we directly use the source code given in [11, 20], so we do not test step decay again, that is, we use the results in [11] directly. For each setting, run 5 times, and then compare the averaged test error rates for all settings. Table 1 shows that HTD(-6,3) outperforms all the others except for DenseNet-BC-250-24 on CIFAR-10, where the error rate of HTD(-6,3) is slightly higher than cosine scheduler by 0.03%. Table 1 also shows that HTD(-6,3)

Table 1: The error rates(%) of CIFAR-10 and CIFAR-100. Best results are written in **blue**. The character * indicates results are directly obtained from the original paper.

Network	Method	runs	**CIFAR-10**	**CIFAR-100**
ResNet-110	step decay	med. of 5	6.03	27.49
	cosine	med. of 5	5.91	27.00
	HTD(-4,4)	med. of 5	5.86	26.84
	HTD(-6,3)	med. of 5	**5.68**	**26.78**
DenseNet-BC-100-12	step decay	[11]	4.51*	22.27*
	cosine	med. of 5	4.51	22.59
	HTD(-4,4)	med. of 5	4.45	22.47
	HTD(-6,3)	med. of 5	**4.43**	**22.17**
WRN-28-10	step decay	med. of 5	4.32	20.43
	cosine	med. of 5	4.31	20.54
	HTD(-4,4)	med. of 5	4.31	19.74
	HTD(-6,3)	med. of 5	**4.22**	**19.73**
DenseNet-BC-250-24	step decay	[11]	3.62*	17.60*
	cosine	med. of 2	**3.39**	17.44
	HTD(-4,4)	med. of 2	3.40	17.29
	HTD(-6,3)	med. of 2	3.42	**17.02**

Table 2: The error rates (%, 1-crop testing) on ImageNet. Best results are written in **blue**. The character * indicates results are directly obtained from the original paper.

Network	Method	top-1 err.	top-5 err.
ResNet-18	step decay	30.43*	10.76*
	HTD(-6,3)	**29.85**	**10.43**
ResNet-34	step decay	26.73*	8.74*
	HTD(-6,3)	**26.25**	**8.43**

performs better than HTD(-4,4), except for DenseNet-BC-250-24 on CIFAR-10. However, HTD(-4,4) still outperforms step decay and cosine except for WRN-28-10 and DenseNet-BC-250-24 on CIFAR-10.

From above, HTD(-6,3) improves the performance over step decay by lowering the error rates significantly, e.g., improving 0.35% for ResNet-110 on CIFAR-10, 0.2% for DenseNet-BC-250-24 on CIFAR-10, 0.7% for ResNet-110 and WRN-28-10 on CIFAR-100, and 0.58% for DenseNet-BC-250-24 on CIFAR-100. In addition, HTD(-6,3) outperforms cosine scheduler, by 0.42% for DenseNet-BC-100-12 and DenseNet-BC-250-24 on CIFAR-100, 0.81% for WRN-28-10 on CIFAR-100.

4.3.2 Other Data Sets

This subsection demonstrates the effectiveness of HTD for different datasets.

ImageNet. We only evaluate the ResNet-18 and ResNet-34 by using the source code of [6] since it takes a large amount of time to train ImageNet. For the same reason, we run HTD(-6,3) with 1 run, and directly use the results of step decay [1] for the network. Table 2 shows that HTD(-6,3) improve the performance over step decay significantly, e.g., improving 0.58% and 0.48 % top-1 error for ResNet-18 and ResNet-34 respectively, and 0.33% and 0.31% top-5 error for ResNet-18 and ResNet-34 respectively.

Fashion-MNIST and PAMAP2. For Fashion-MNIST, this paper uses ResNet-110 like CIFAR datasets. For the dataset PAMAP2, this paper uses BLSTM to train. For each setting, use the aver-

---

[1] https://github.com/facebook/fb.resnet.torch/blob/master/TRAINING.md

Table 3: The error rates (%) on PAMAP2 and Fashion-MNIST. Best results are written in **blue**.

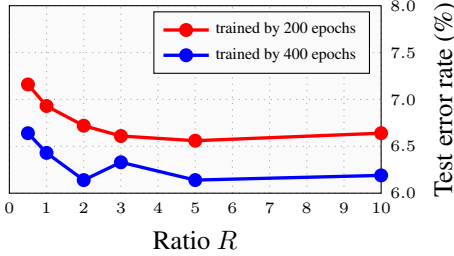| datasets/Network	Method	runs	error rate
PAMAP2/BLSTM	step decay	med. of 5	11.88
	cosine	med. of 5	11.94
	HTD(-4,4)	med. of 5	11.79
	HTD(-6,3)	med. of 5	**11.72**
Fashion-MNIST/ResNet-110	step decay	med. of 5	4.99
	cosine	med. of 5	4.90
	HTD(-4,4)	med. of 5	4.92
	HTD(-6,3)	med. of 5	**4.84**

Figure 7: Different ratio R with same upper bound $U = 3$ on CIFAR-10.
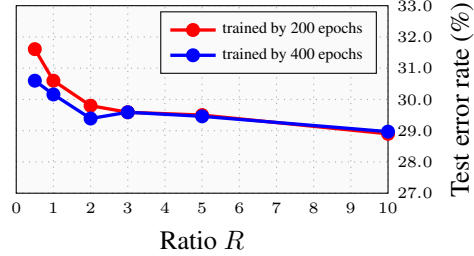

Figure 8: Different ratio R with same upper bound $U = 3$ on CIFAR-100.

aged test error rate of 5 runs. Table 3 shows that HTD(-6,3) outperforms others on both PAMAP2 and Fashion-MNIST datasets. Compared to step decay, HTD(-6,3) improves 0.16% and 0.15% for PAMAP2 and Fashion-MNIST respectively. Compared to cosine scheduler, HTD(-6,3) improves 0.22% and 0.06% for PAMAP2 and Fashion-MNIST respectively.

4.3.3 The effect of ratio in HTD

To demonstrate the importance of adjusting the ratio R, we design experiments similar to Subsection 3.1. We train ResNet-32 on CIFAR-10 and CIFAR-100 with 200 epochs and 400 epochs and fix $U = 3$. For each setting, use the averaged test error rate of 4 runs. Figure 7 shows the performance on CIFAR-10, while Figure 8 shows the performance on CIFAR-100.

Interestingly, the performance of CIFAR-10 improves significantly when the total epoch number doubles, i.e., $T = 400$, while the performance of CIFAR-100 does not improve for $T = 400$ when the ratio R larger than 3. This implies that $T = 200$ is not sufficient for CIFAR-10, but sufficient for CIFAR-100. For CIFAR-10, the best performance is $R = 5$ for 200 epochs and $R = 2$ for 400 epochs. For CIFAR-100, the best performance is $R = 10$ for both 200 and 400 epochs. This implies that the best ratios are different for different datasets and settings of hyperparameters.

5 Conclusion

This paper proposes a new learning rate scheduler, HTD, on SGD. Our experiments show HTD is superior to the step decay and cosine scheduler in the following aspects. Compared to step decay, HTD outperforms it in all experiments, and has less hyperparameters to tune. Compared to cosine scheduler, HTD has better performance in nearly all cases, and is more flexible to achieve better performance.

More investigations on HTD are to be researched in the future. First, it is still an open problem to determine the best hyperparameters, such as L and U, for many different datasets or networks. For example, Subsubsection 4.3.3 shows that $L = -15, U = 3$ ($R = 5$) and $L = -30, U = 3$ ($R = 10$) performs best on CIFAR-10 and CIFAR-100 respectively. This implies that the best L and U are different for different datasets. Hence, it is still likely to explore better results. Second, this paper has not yet tried some hybrid mechanisms for HTD. For example, incorporate HTD into the restart

mechanisms like [1, 17, 24]. With restart mechanism, HTD is also able to use the snapshot like [9] to improve the performance and avoid over-fitting.

References

[1] Wangpeng An, Haoqian Wang, Yulun Zhang, and Qionghai Dai. Exponential decay sine wave learning rate for fast deep neural network training. In *Visual Communications and Image Processing (VCIP), 2017 IEEE*, pages 1–4. IEEE, 2017.

[2] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[3] Alex Graves. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, pages 602–610, 2005.

[4] Nils Y Hammerla, Shane Halloran, and Thomas Ploetz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[9] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

[10] Gao Huang, Shichen Liu, Laurens van der Maaten, and Kilian Q Weinberger. Condensenet: An efficient densenet using learned group convolutions. *arXiv preprint arXiv:1711.09224*, 2017.

[11] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3, 2017.

[12] Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[15] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] Ilya Loshchilov and Frank Hutter. Sgdr: stochastic gradient descent with restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[18] Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[19] Jorge Nocedal and Stephen J Wright. *Sequential quadratic programming*. Springer, 2006.

[20] Geoff Pleiss, Danlu Chen, Gao Huang, Tongcheng Li, Laurens van der Maaten, and Kilian Q Weinberger. Memory-efficient implementation of densenets. *arXiv preprint arXiv:1707.06990*, 2017.

[21] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[22] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 108–109. IEEE, 2012.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[24] Leslie N Smith. Cyclical learning rates for training neural networks. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 464–472. IEEE, 2017.

[25] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[26] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[27] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[28] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[29] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.