Investigating the Role of Weight Decay in Enhancing Nonconvex SGD

Tao Sun*
National University of Defense Technology

Yuhao Huang
University of Utah

Li Shen
Sun Yat-sen University

Kele Xu
National University of Defense Technology

Bao Wang
University of Utah

Abstract

Weight decay is a widely used technique in training machine learning models, known to empirically enhance the generalization of Stochastic Gradient Descent (SGD). While intuitively weight decay allows SGD to train a regularized model rather than the original one, there is limited theoretical understanding of why SGD with weight decay (SGDW) yields results consistent with the unregularized model, or how weight decay improves generalization. This paper establishes a convergence theory for SGDW in the context of the unregularized model, under weaker assumptions than previous analyses of weight decay. Our theory demonstrates that weight decay does not accelerate the convergence of SGD. For generalization, we provide the first theoretical proof of weight decay's benefit in nonconvex optimization. Additionally, we extend our results to sign-based stochastic gradient algorithms, such as SignSGD. Numerical experiments on classical benchmarks validate our theoretical findings.

1. Introduction

Empirical Risk Minimization (ERM) is a foundational task in machine learning, which can be formulated as follows:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} f_S(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w};\xi_i), \tag{1}$$

where $f(\boldsymbol{w};\xi_i)$ is the loss function associated with the training data ξ_i and the parameter $\boldsymbol{w}$ and $\{\xi_1,\xi_2,\dots,\xi_n\}\subset\Xi$[1] represents the training data. The algorithms of choice for ERM are Stochastic Gradient Descent (SGD) and its variants [8, 29] due to their simplicity, scalability, and generalizability.

*Corresponding Author, suntao.saltfish@outlook.com.

[1] Ξ denotes the distribution of training data.

1.1. SGD with Weight Decay

SGD with Weight decay (SGDW) is more used than without weight decay in training Deep Neural Networks (DNNs). Though there are various types of weight decay schemes, these schemes can be summarized as Algorithm 1. The use of weight decay stems from the ℓ_2 regularization of the original machine learning model [9, 19, 22]. Initially, the training employs the parameter $\alpha = 1-\lambda_{\ell_2}$, where $\lambda_{\ell_2} > 0$ is a small preset number – a regularization hyperparameter. Interestingly, even when applied to the regularized model with regularized training loss $\min_{\boldsymbol{w}\in\mathbb{R}^d} f_S(\boldsymbol{w}) + \frac{\lambda_{\ell_2}}{2}\|\boldsymbol{w}\|^2$, weight decay further improves the generalization of the original model (1) [24].

Algorithm 1 SGD with Weight Decay (SGDW)

Require: parameters $\gamma > 0$, $0 < \alpha < 1$
Initialization: $\boldsymbol{w}^1, \boldsymbol{g}^0 = \boldsymbol{0}$
for $t = 1, 2, \dots$
step 1: sample $i_t \sim \mathrm{Uni}[n]$ and calculate the stochastic gradient $\boldsymbol{g}^t = \nabla f(\boldsymbol{w}^t;\xi_{i_t})$
step 2: update $\boldsymbol{w}^{t+1} = \alpha\boldsymbol{w}^t - \gamma\boldsymbol{g}^t$
end for

In the context of SGD, ℓ_2 regularization and weight decay are equivalent, and the fixed weight decay – with a constant weight decay hyperparameter α – is acceptable. However, for adaptive methods such as Adam [23], the equivalence between weight decay and ℓ_2 regularization breaks – as emphasized by [28]. The authors of [28] also show that ℓ_2 regularization is not efficient for adaptive methods. To address this issue, the authors of [28] introduce the decoupled weight decay – represented as $\alpha = 1-\gamma\lambda_{\text{decoupled}}$ with $\lambda_{\text{decoupled}} > 0$ being a small positive number. The decoupled approach mitigates the aforementioned drawback by introducing a diminishing learning rate γ during the training. Paper [28] empirically shows that decoupled weight decay remarkably improves generalization, particularly when training DNNs using Adam [23].

Despite recent efforts to explore the intricacies of weight decay [7, 35, 36, 38], a theoretical framework to elucidate the effect of weight decay on the convergence and generalization of general nonconvex SGD is still lacking. This paper is dedicated to understanding this fundamental question.

1.2. Our Contributions

This paper contributes to understanding how weight decay operates and why it is beneficial for nonconvex SGD. Our main contributions are summarized as follows:

- We establish the convergence of SGDW for nonconvex optimization under conditions of unbiased gradient estimation and bounded variance when $1-\alpha=\mathcal{O}(\gamma^2)$. Our results demonstrate that weight decay does not accelerate the convergence of nonconvex SGD.
- We investigate the generalization of SGDW and establish its advantages over SGD without weight decay. Our theory shows that weight decay improves the stability bound of SGD when the gradient and weight parameters are not linearly dependent.
- We broaden the scope of our theoretical contributions to encompass sign-based stochastic gradient methods, e.g. SignSGD with momentum [6, 34], providing insights into their convergence and generalization. We demonstrate the superiority of SGDW over sign-based methods with weight decay. It is noteworthy that the sign-based algorithms – unlike SGDW – do not necessitate the bounded gradient assumption in their generalization analysis.
- We substantiate our theoretical results with empirical evidence on Cifar and ImageNet classification using ResNet and VGG. Our investigation confirms that 1) weight decay has a negligible impact on the convergence of SGD, 2) weight decay benefits the generalization of SGD, and 3) SGDW generalizes better than sign-based algorithms with weight decay.

Notations. We call S and S' two neighboring datasets if they have the same number of samples and only differ by one. The ℓ_2-, ℓ_1-, and ℓ_∞-norm of a vector are denoted by $\|\cdot\|$, $\|\cdot\|_1$, and $\|\cdot\|_\infty$, respectively. For positive sequences a_t and b_t, we denote $a_t=\mathcal{O}(b_t)$ if there exists a positive constant $0<C<+\infty$ such that $a_t\leq Cb_t$. Similarly, $a_t\asymp b_t$ implies $a_t=\mathcal{O}(b_t)$ and $b_t=\mathcal{O}(a_t)$. We denote the uniform distribution over the set $\{1,2,\dots,n\}$ as $\mathrm{Uni}[n]$. For a vector $\boldsymbol{a}$, we use $\mathrm{Sign}(\boldsymbol{a})$ to denote the sign of each coordinate of $\boldsymbol{a}$.

2. Additional Related Works

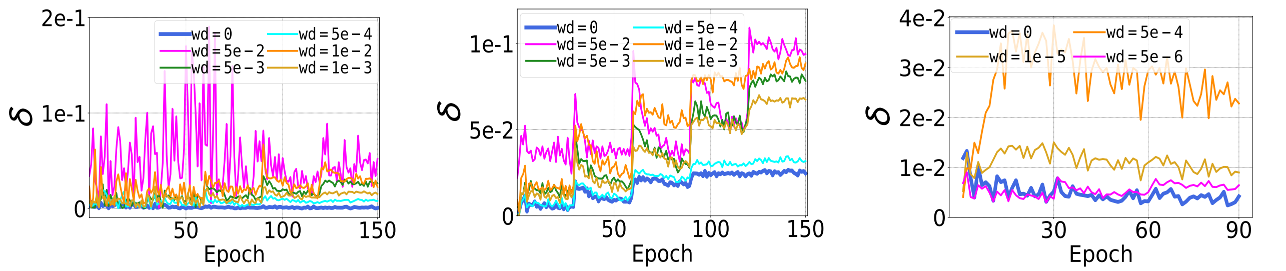
Weight Decay. Weight decay is a fundamental and widely used technique for training DNNs, including large language models [11, 13]. The authors of [24] observe that weight decay can enhance the generalization of DNN. [37] demonstrates its efficacy in improving test accuracy, especially on large datasets like ImageNet [30]. Different regularization schemes are proposed for different optimization algorithms and DNN architectures [38]. The authors of [35] explore normalized neural networks with weight decay, shedding light on equilibrium states. The authors of [7] delve into the decoupled weight decay technique, addressing norm maintenance and gradient mixing issues in adaptive optimizers, notably in Natural Language Processing (NLP) and Reinforcement Learning (RL) tasks. Paper [1] reveals weight decay's role in implicit regularization for overparameterized networks and bias-variance tradeoff in stochastic optimization for underparameterized language models. Paper [36] proposes a scheduled weight decay dynamically adjusting strength based on gradient norms, showcasing superiority over the traditional constant strategy, particularly with ADAM. The authors of [18] introduce adaptive weight decay to dynamically adjust the hyperparameter during each training iteration, resulting in remarkable improvements in adversarial robustness across diverse datasets and architectures. In [14], the authors argue that weight decay in modern deep learning goes beyond traditional regularization. It enhances optimization and stabilizes loss in vision tasks, while improving the bias-variance tradeoff and stability in large language models.

Generalization of Nonconvex Algorithms. Algorithmic stability – rooted in sensitivity analysis – stands out as a widely adopted and effective method for exploring the generalization capability of SGD [10]. Pioneering work on the generalization of nonconvex SGD is undertaken by [20], where they investigate the generalization error under smoothness and bounded gradient assumptions. Subsequent studies [4, 26] relax the smoothness assumption. In [39], the authors establish a high probability bound for the generalization of nonconvex SGD. The concept of Bayes-stability is introduced in [27] to demonstrate algorithm-dependent generalization in nonconvex stochastic gradient Langevin dynamics. Another approach, known as ϵ-uniform stability, is proposed by [25] to study nonconvex SGD. Compared to [20], [25] eliminates the smoothness assumption while achieving improved outcomes for nonconvex scenarios. In recent years, the generalization of distributed algorithms has been extensively studied using algorithmic stability tools [16, 17, 32]. The generalization analysis of SGDW remains open.

3. Weight Decay Does not Accelerate SGD

3.1. Assumptions for Convergence Analysis

We gather two commonly used assumptions: the gradient Lipschitz and bounded variance assumptions.



(a) VGG16 on Cifar10 (b) ResNet20 on Cifar100 (c) ResNet18 on ImageNet

Figure 1. The averaged value of δ per epoch vs. epochs for SGDW under different weight decay values on VGG16, ResNet20, and ResNet18 for Cifar10, Cifar100, and ImageNet classification, respectively.

Assumption 1 *The loss function $f(\cdot;\xi)$ satisfies the gradient Lipschitz, i.e., $\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{y};\xi)\| \leq L\|\boldsymbol{x}-\boldsymbol{y}\|$ as $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^d$ and $\xi \in \Xi$ for some $L > 0$.*

Assumption 2 *The sample selected from dataset S is unform, i.e., $\mathbb{E}\boldsymbol{g}^t = \nabla f_S(\boldsymbol{w}^t)$. The variance of the stochastic gradient is uniformly bounded, i.e., $\sup_{\boldsymbol{w}\in\mathbb{R}^d} \mathbb{E}_{i\sim \mathrm{Uni}[n]}\|\nabla f(\boldsymbol{w};\xi_i) - \nabla f_S(\boldsymbol{w})\|^2 \leq \sigma^2$ for some $\sigma > 0$.*

3.2. Main Convergence Results

We present the convergence results of SGDW in this section.

Theorem 1 *Let Assumptions 1 and 2 hold, and $(\boldsymbol{w}^t)_{t\geq 1}$ be generated by SGDW, and γ is small. Then, as $1-\alpha \leq \min\left\{\gamma^2, \frac{1}{2T}\right\}$, it holds that*

$$\frac{\sum_{t=1}^T \mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2}{T} = \mathcal{O}\Big(\frac{\mathbb{E}f_S(\boldsymbol{w}^1) - \min f_S}{\gamma T} + \frac{L^2\gamma\sigma^2}{2} + \sqrt{1-\alpha}\gamma\sigma^2\Big).$$

If $1-\alpha = \gamma$, we have

$$\frac{\sum_{t=1}^T \mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2}{T} = \mathcal{O}\Big(\frac{\mathbb{E}f_S(\boldsymbol{w}^1) - \min f_S}{(1-\delta_T)\gamma T} + \frac{(L^2+1)\gamma\sigma^2}{2(1-\delta_T)}\Big),$$

where $\delta_T := \sup_{0\leq t\leq T}\left\{\frac{|\langle \nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle|}{\|\nabla f_S(\boldsymbol{w}^t)\|\cdot\|\boldsymbol{w}^t\|}\right\}$.[2]

Theorem 1 presents results under two distinct weight decay scenarios: one involving a small decay rate[3] and the other involving a decay rate equivalent to the learning rate.

[2]We adopted the convention $0 = \frac{0}{0}$ because when the denominator is zero, at least one of $f_S(\boldsymbol{w}^t)$ or $\boldsymbol{w}^t$ becomes $\mathbf{0}$ and $\langle \nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle \leq c\|\nabla f_S(\boldsymbol{w}^t)\|\cdot\|\boldsymbol{w}^t\|$ holds for any $c \in \mathbb{R}^d$.

[3]We call $1-\alpha$ as decay rate in this paper.

Notice that a commonly used learning rate for stochastic optimization is $\gamma = \frac{1}{\sqrt{T}}$. The weight decay satisfies $1-\alpha \leq \min\{\gamma^2, \frac{1}{2T}\} = \frac{1}{2T}$ by Theorem 1, resulting in a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$, which matches the established optimal rate for nonconvex first-order SGD [2]. Notably, in this scenario, we observe that $1-\alpha \asymp \frac{\gamma}{\sqrt{T}}$, echoing commonly used settings in practice. Moreover, such a choice coincides with the decoupled weight decay, explaining why the decoupled approach is more effective than ℓ_2 regularization. In comparison to the previous work [36], our analysis for the small decay rate scenario eliminates the necessity for bounded gradients, leading to more sensible and realistic convergence outcomes. Moreover, our results offer an additional advantage by characterizing the convergence of the original function rather than its ℓ_2 regularized function.

Notice that $\delta_T \leq 1$ according to the Cauchy-Schwarz inequality, and the worst case $\delta_T = 1$ occurs if $|\langle \nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle| = \|\nabla f_S(\boldsymbol{w}^t)\|\cdot\|\boldsymbol{w}^t\|$, i.e., $\nabla f_S(\boldsymbol{w}^t) = c\boldsymbol{w}^t$ for some $c \in \mathbb{R}$ and t. Indeed, vectors that satisfy the condition are rare even for quadratic functions. Let us consider a quadratic function with $\nabla f_S(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$. From basic algebra, we know that $\boldsymbol{A}\boldsymbol{x} = c\boldsymbol{x}$ implies that $\boldsymbol{x}$ is either $\mathbf{0}$ or an eigenvector of $\boldsymbol{A}$. Therefore, $\boldsymbol{x} \in \mathrm{Span}\{(\boldsymbol{v}_i)_{i\in I}\}$, where $(\boldsymbol{v}_i)_{i\in I}$ share the same eigenvalue. $\mathrm{Vol}(\mathrm{Span}\{(\boldsymbol{v}_i)_{i\in I})$ is zero unless $\boldsymbol{A}$ has only one eigenvalue, i.e., $\boldsymbol{A} = c\boldsymbol{I}$ for some $c \neq 0$. Therefore, $|\langle \nabla f_S(\boldsymbol{x}), \boldsymbol{x}\rangle| = \|\nabla f_S(\boldsymbol{x})\|\cdot\|\boldsymbol{x}\|$ holds only for a set with zero measure; in the rest part it follows that $|\langle \nabla f_S(\boldsymbol{x}), \boldsymbol{x}\rangle| < \|\nabla f_S(\boldsymbol{x})\|\cdot\|\boldsymbol{x}\|$. In Figure 1, we show that $\delta = \delta_T < 1$ in some benchmark DNN training tasks. We can employ another kind of weight decay as $\alpha = 1 - \frac{1}{\sqrt{T}}$ with the definition of δ_T. Under this weight decay, SGDW has the convergence rate of $\mathcal{O}\Big(\frac{1}{(1-\delta_T)\sqrt{T}}\Big)$, which is worse than the convergence rate of the classical SGD or SGDW under a small decay rate.

In summary, Theorem 1 demonstrates that weight decay deteriorates the convergence of SGDW with an optimal choice of weight decay hyperparameter 0. This finding is consistent with the results reported by [36]. A natural concern then arises: *does the detrimental effect of weight decay on the convergence of SGD come from our analysis or is it an* To understand this, we conduct numerical experiments to compare the effects of weight decay with different values in Section 6.1. As we will see, the numerical results show that weight decay does slow down SGD in training DNNs, and the numerical results confirm our theoretical analysis.

4. Weight Decay Improves Generalization

Though weight decay does not improve the convergence of SGD, we show that it improves the generalization of SGD in this section. Note that the ERM problem in (1) is an offline problem with the convergence being characterized by the bound of $\|\nabla f_S(\boldsymbol{w}^t)\|$. Indeed, ERM serves as a surrogate for the following expectation minimization model $\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}) := \mathbb{E}_{\xi\sim\mathcal{D}}\nabla f(\boldsymbol{w};\xi)$. Therefore, we will investigate the bound of $\|\nabla f(\boldsymbol{w}^t)\|$ for the generalization analysis of nonconvex SGD.

4.1. Setup for Generalization Analysis

In this subsection, we recap on uniform stability for nonconvex optimization. The current analysis of nonconvex SGD requires an extra assumption of uniform boundedness of gradient and variance of the distribution [25].

Assumption 3 *The gradients in the training are uniformly bounded, that is,* $\sup_{\boldsymbol{w}\in\mathbb{R}^d,\xi\in\Xi}\{\|\nabla f(\boldsymbol{w};\xi)\|\} \leq G$ *for some constant* $G > 0$.

Recently, [25] has proven the generalization bound for the nonconvex SGD based on algorithmic gradient stability. We summarize their results in the following Lemma.

Lemma 1 *[25] Let $\mathcal{A}$ be a stochastic algorithm and the cardinality of the dataset S is n. Assume for any ξ, the function $f(\boldsymbol{w};\xi)$ is differentiable. If Assumption 3 holds and* $\sup_{\boldsymbol{w}\in\mathbb{R}^d}\mathbb{E}_{\xi\sim\mathcal{D}}\|\nabla f(\boldsymbol{w};\xi)-\nabla f(\boldsymbol{w})\|^2 \leq \sigma^2$, *then*

$$\mathbb{E}_{S,\mathcal{A}}\|\nabla f(\mathcal{A}(S)) - \nabla f_S(\mathcal{A}(S))\| \leq \frac{\sigma}{\sqrt{n}} + \mathcal{B}_{\mathcal{A}}, \quad (2)$$

where $f(\cdot) := \mathbb{E}_{\xi\sim\mathcal{D}}\nabla f(\cdot;\xi)$, *and* $\mathcal{B}_{\mathcal{A}} := \sup_{\xi\in\Xi}\sqrt{\mathbb{E}_{\mathcal{A}}\|\nabla f(\mathcal{A}(S);\xi) - \nabla f(\mathcal{A}(S');\xi)\|^2}$ *with S and S' being two neighboring datasets.*

Now, we are ready to present the difference in the gradients between the outputs of SGDW on two neighboring datasets.

Lemma 2 *Let Assumptions 1, 2, and 3 hold, and* $\sup_{\boldsymbol{w}\in\mathbb{R}^d}\mathbb{E}_{\xi\sim\mathcal{D}}\|\nabla f(\boldsymbol{w};\xi) - \nabla f(\boldsymbol{w})\|^2 \leq \sigma^2$, *and let $\mathcal{A}_{\text{SGDW}}$ be SGDW on dataset $S\sim\mathcal{D}$ of n samples with learning rate γ and iteration number T. If* $1-\alpha \leq \min\left\{\gamma^2, \frac{1}{2T}\right\}$, *we have* $\mathcal{B}_{\mathcal{A}_{\text{SGDW}}} = \mathcal{O}\left(\sqrt{\frac{T}{n}}\right)$. *If $1-\alpha=\gamma$ and $L \leq 1$, we have* $\mathcal{B}_{\mathcal{A}_{\text{SGDW}}} = \mathcal{O}\left(\sqrt{\frac{\gamma T}{n}}\right)$.

In the lemma, we impose the condition that Lipschitz constant is less than 1. This prerequisite can be easily satisfied by introducing a factor of $\frac{1}{L}$ to f, given that $L > 1$. Based on Lemma 2, it is evident that the large decay rate gives a remarkably better generalization bound than the small decay rates and decay-free cases. This small bound serves as a crucial explanation for the remarkable generalization advantage observed for SGD with weight decay.

4.2. Main Results for Generalization of SGDW

In what follows, we present the nonconvex generalization result for SGDW.

Theorem 2 *Let Assumption 1, 2, and 3 hold, and* $\sup_{\boldsymbol{w}\in\mathbb{R}^d}\mathbb{E}_{\xi\sim\mathcal{D}}\|\nabla f(\boldsymbol{w};\xi) - \nabla f(\boldsymbol{w})\|^2 \leq \sigma^2$, *and $L \leq 1$, and $(\boldsymbol{w}^t)_{1\leq t\leq T}$ be generated by SGDW applied to dataset S with n samples. If $\gamma = \frac{1}{\sqrt{T}}$ and $1-\alpha \leq \frac{1}{2T}$, it holds that*

$$\frac{\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|}{T} = \mathcal{O}\left(\frac{1}{T^{1/4}} + \sqrt{\frac{T}{n}}\right).$$

If $1-\alpha = \frac{1}{\sqrt{T}}$, we have

$$\frac{\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|}{T} = \mathcal{O}\left(\frac{1}{\sqrt{1-\delta_T}T^{1/4}} + \frac{T^{1/4}}{\sqrt{n}}\right),$$

where δ_T follows the same definition in Theorem 1.

It has been analytically verified that the optimal choice for the number of iterations is $T = n^{2/3}$ for the small decay rates cases. This selection results in a bounded generalization error of $\mathcal{O}(\frac{1}{n^{1/6}})$. These results align with the generalization error established for nonconvex SGD in [25].

In our study, we have not observed any significant advantage of SGDW in terms of convergence or generalization in small decay rate cases. As mentioned earlier, the numerical results in Section 6.1 are consistent with the conclusion that weight decay does not improve convergence. However, weight decay does enhance the generalization of SGD, as numerically demonstrated in Section 6.2. To understand the improved generalization of SGD using weight decay, we consider the case $\alpha = 1-\gamma$. According to Theorem 2, we find that the optimal choice for T is $T = \frac{n}{1-\delta_T}$. In this case, the generalization error bound becomes $\mathcal{O}\left(\frac{1}{(1-\delta_T)^{1/4}n^{1/4}}\right)$. If

$$\delta_T \ll 1 - \frac{1}{T^{1/4}} \Rightarrow \delta_T \ll 1 - \frac{1}{n^{1/3}},$$

it follows that $\frac{1}{(1-\delta_T)^{1/4}n^{1/4}} \ll \frac{1}{n^{1/6}}$. Therefore, under the condition $\delta_T \ll 1-\frac{1}{T^{1/4}}$, Theorem 2 explains that weight decay improves the generalization of SGD for nonconvex optimization.

5. Extension to Sign-Based Algorithms

Sign-based SGD variants utilize gradient signs in optimization. Studies like [3, 6] show ADAM can be simplified to SignSGD with specific parameters. LION, a sign-based method introduced by [12], outperforms ADAM in preventing overfitting. [33] demonstrate nonconvex acceleration with sign-based training, considering weak Lipschitz properties. They link their approach to LION, emphasizing its theoretical benefits. Given the the connection between ADAM and its commendable performance in sign-based methods, this section studies the convergence and generalization of their weight decay versions. It is essential to underscore that the vanilla SignSGD does not converge unless the batch size is sufficiently large [5]. The authors of [34] show that the addition of a simple momentum term ensures convergence for SignSGD. Consequently, we integrate momentum into SignSGD – resulting in SignSGD with Weight decay (SignSGDW) presented in Algorithm 2 – in our analysis.

Algorithm 2 SignSGD with Weight decay (SignSGDW)

Require: parameters $\gamma > 0, 0 < \alpha < 1$
  Initialization: $\boldsymbol{w}^0 = \mathbf{0}, \boldsymbol{m}^0 = \mathbf{0}$
  for $t = 1, 2, \ldots$
    step 1: sample $i_t \sim \text{Uni}[n]$ and calculate the stochastic gradient $\boldsymbol{g}^t = \nabla f(\boldsymbol{w}^t; \xi_{t_t})$
    step 2: update $\boldsymbol{m}^t = \theta \boldsymbol{m}^{t-1} + (1-\theta)\boldsymbol{g}^t$
    step 3: update $\boldsymbol{w}^{t+1} = \alpha \boldsymbol{w}^t - \gamma \text{Sign}(\boldsymbol{m}^t)$
  end for

5.1. Convergence of SignSGDW

The convergence of SignSGDW is presented in Theorem 3.

Theorem 3 *Let Assumptions 1 and 2 hold, and $(\boldsymbol{w}^t)_{t\geq 1}$ be generated by SignSGDW. The learning rate and momentum of SignSGDW are set to be $1-\theta = \frac{1}{\sqrt{T}}$ and $\gamma = \frac{1}{L\sqrt{d}T^{3/4}}$, respectively. Then, as $1-\alpha \leq \frac{1}{T}$, we have*

$$\frac{\sum_{t=1}^T \mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|_1}{T} = \mathcal{O}\Big(\frac{\sqrt{d}(f_S(\boldsymbol{w}^1) - \min f_S + \sigma + \|\nabla f_S(\boldsymbol{w}^1)\|)}{T^{1/4}}\Big).$$

If $1-\alpha = \frac{1}{T^{3/4}}$, it holds that

$$\frac{\sum_{t=1}^T \mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|_1}{T} = \mathcal{O}\Big(\frac{\sqrt{d}(f_S(\boldsymbol{w}^1) - \min f_S + \sigma + \|\nabla f_S(\boldsymbol{w}^1)\|)}{(1-\delta_T)T^{1/4}}\Big),$$

where $\delta_T := \sup_{0\leq t\leq T}\Big\{\frac{|\langle \nabla f(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle|}{\|\nabla f(\boldsymbol{w}^t)\|_1 \cdot \|\boldsymbol{w}^t\|_\infty}\Big\}$.

For the small decay rate case ($1-\alpha \leq \frac{1}{T}$), the convergence rate of SignSGDW is precisely characterized as $\mathcal{O}\left(\frac{1}{T^{1/4}}\right)$, aligning with the optimal convergence rate for the first-order stochastic gradient algorithm. This observed convergence rate stands out as it experiences an augmentation by a factor of $\sqrt{d}$ compared with SGDW. This augmentation arises due to the adoption of $\|\cdot\|_1$ in SignSGDW rather than $\|\cdot\|$ in SGDW. Indeed, the factor $\sqrt{d}$ is substantiated by the fact that $\|\cdot\|_1 \leq \sqrt{d}\|\cdot\|$, showing that a comparable impact of $\sqrt{d}$ would persist if $\|\cdot\|_1$ was incorporated into SGDW.

Notice that $\delta_T \leq 1$ due to the Hölder inequality, which states that $|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|_1 \cdot \|\boldsymbol{y}\|_\infty$. Compared to SGDW, SignSGDW exhibits notable distinctions in its hyperparameter configuration. In particular, the learning rate employed in SignSGDW is significantly smaller than SGDW, concurrently ensuring a comparably large decay rate in SignSGDW.

5.2. Generalization of SignSGDW

We present the generalization error of SignSGDW below.

Theorem 4 *Let Assumptions 1 and 2 hold, and $\sup_{\boldsymbol{w}\in\mathbb{R}^d} \mathbb{E}_{\xi\sim\mathcal{D}}\|\nabla f(\boldsymbol{w};\xi) - \nabla f(\boldsymbol{w})\|^2 \leq \sigma^2$. Let $(\boldsymbol{w}^t)_{t\geq 1}$ be generated by SignSGDW, and the learning rate and momentum are set to be $1-\theta = \frac{1}{\sqrt{T}}$ and $\gamma = \frac{1}{L\sqrt{d}T^{3/4}}$. If $1-\alpha \leq \frac{1}{T}$, we have*

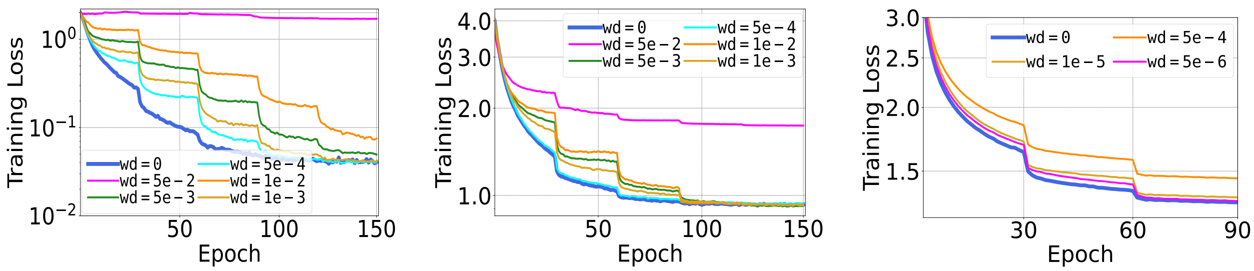
$$\frac{\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|_1}{T} = \mathcal{O}\Big(\frac{\sqrt{d}}{T^{1/4}} + \frac{\sqrt{d}T^{3/4}}{\sqrt{n}}\Big).$$

If $1-\alpha = \frac{1}{T^{3/4}}$, it holds that

$$\frac{\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|_1}{T} = \mathcal{O}\Big(\frac{\sqrt{d}}{(1-\delta_T)T^{1/4}} + \sqrt{\frac{T}{n}}\Big),$$

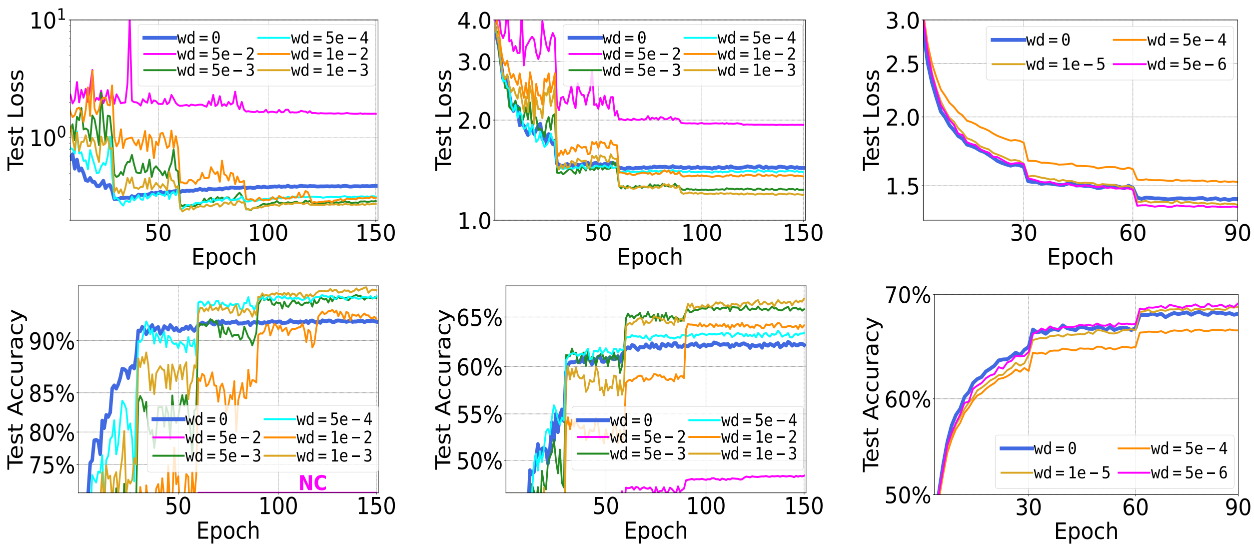
where δ_T follows the same definition in Theorem 3.

Like the convergence result of SignSGDW, the presence of the factor $\sqrt{d}$ in the generalization bound is a consequence of using the $\|\cdot\|_1$ norm as the complexity measure. By letting $T = \sqrt{n}$ in Theorem 4 yields a generalization error of $\mathcal{O}\left(\frac{1}{n^{1/8}}\right)$ when $1-\alpha \leq \frac{1}{T}$. Although this convergence rate is marginally inferior to that of SGDW and SGD, it is crucial to highlight that this result is established without assuming bounded gradients, i.e., Assumption 3. It is noteworthy that the assumptions used to analyze the generalization of SignSGDW align with those employed for the convergence analysis of SGD, signifying that the assumptions for generalization analysis are relatively mild. When $0 < \delta_T < 1$, by setting $T = \frac{n^{2/3}}{(1-\delta_T)^{4/3}}$, weight decay significantly improves the generalization of SignSGD – up to $\mathcal{O}(\frac{1}{(1-\delta_T)^{2/3}n^{1/6}})$. SignSGDW exhibits worse generalization ability over SGDW if the δ_Ts are close for two algorithms.



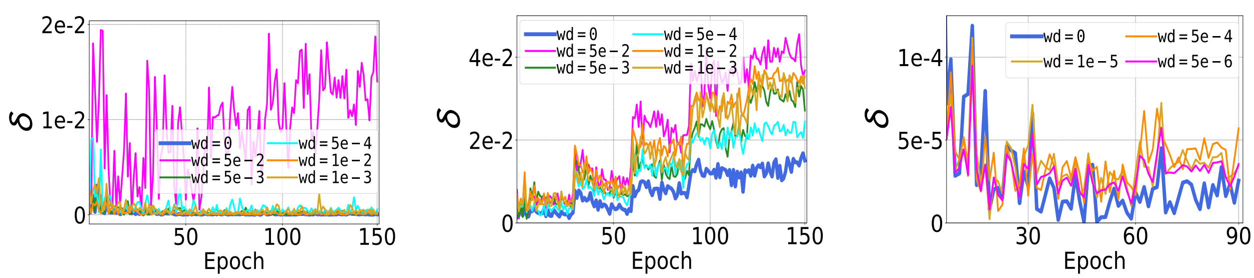
(a) VGG16 on Cifar10 (b) ResNet20 on Cifar100 (c) ResNet18 on ImageNet

Figure 2. Training loss of SGDW under different values of weight decay on VGG16 for Cifar10, ResNet20 for Cifar100, and ResNet18 for ImageNet classification. As the weight decay hyperparameter increases, the convergence of SDGW slows down.



(a) VGG16 on Cifar10 (b) ResNet20 on Cifar100 (c) ResNet18 on ImageNet

Figure 3. The generalization of SGDW, in terms of test accuracy and test loss, under different weight decay values on VGG16, ResNet20, and ResNet18 for Cifar10, Cifar100, and ImageNet classification, respectively. For the VGG16 on the Cifar10 classification task, the SGDW failed to converge when using the largest weight decay value 5e-2 ('NC' is 'Not Converged'), resulting in poor generalization.

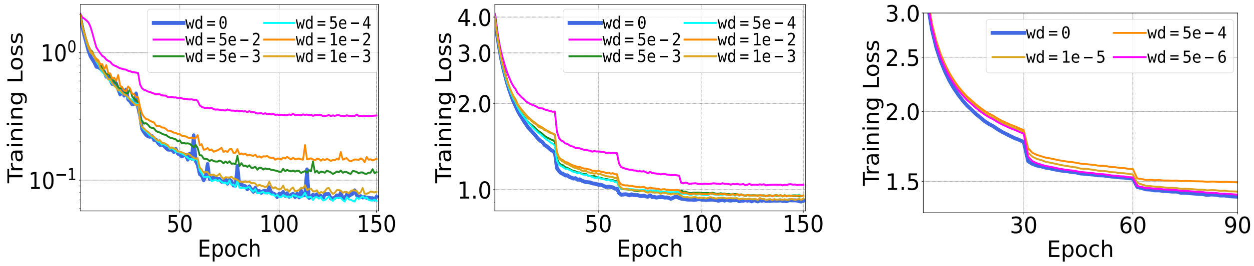


(a) VGG16 on Cifar10 (b) ResNet20 on Cifar100 (c) ResNet18 on ImageNet

Figure 4. The averaged value of δ per epoch vs. epochs for SignSGDW under different weight decay values on VGG16, ResNet20, and ResNet18 for Cifar10, Cifar100, and ImageNet classification, respectively.

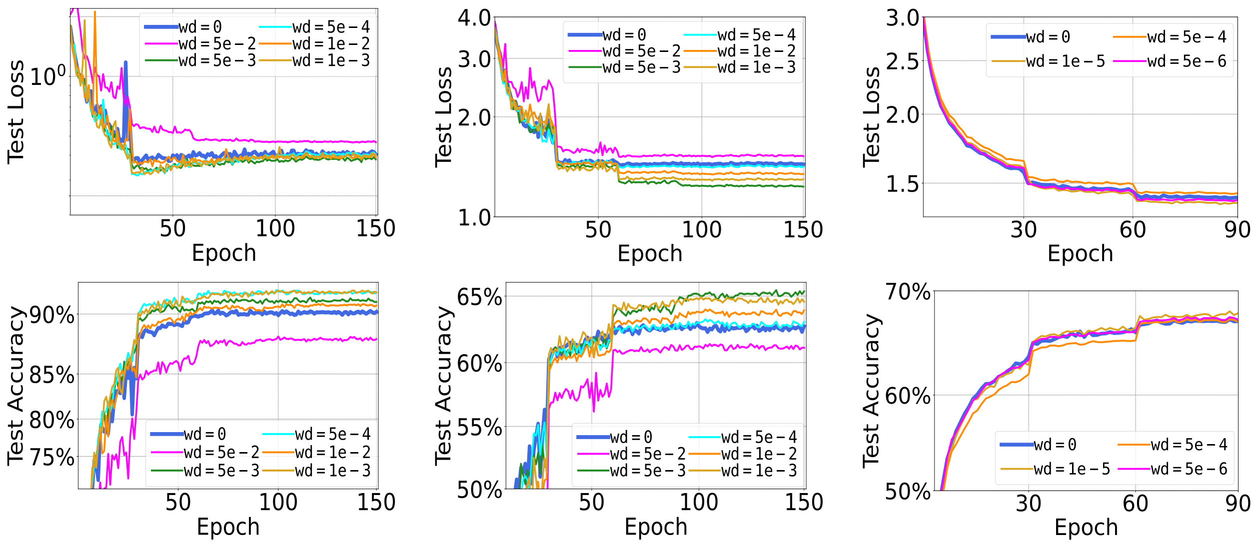
6. Numerical Results

In this section, we provide numerical experiments on training DNNs for a few benchmark image classification tasks to further solidify our theoretical analysis. In particular, we aim to numerically verify our theory on how weight decay affects the convergence and generalization of SGD and SignSGD.



(a) VGG16 on Cifar10 (b) ResNet20 on Cifar100 (c) ResNet18 on ImageNet

Figure 5. The convergence of SignSGDW under different weight decay values on VGG16 for Cifar10, ResNet20 for Cifar100, and ResNet18 for ImageNet classification.



(a) VGG16 on Cifar10 (b) ResNet20 on Cifar100 (c) ResNet18 on ImageNet

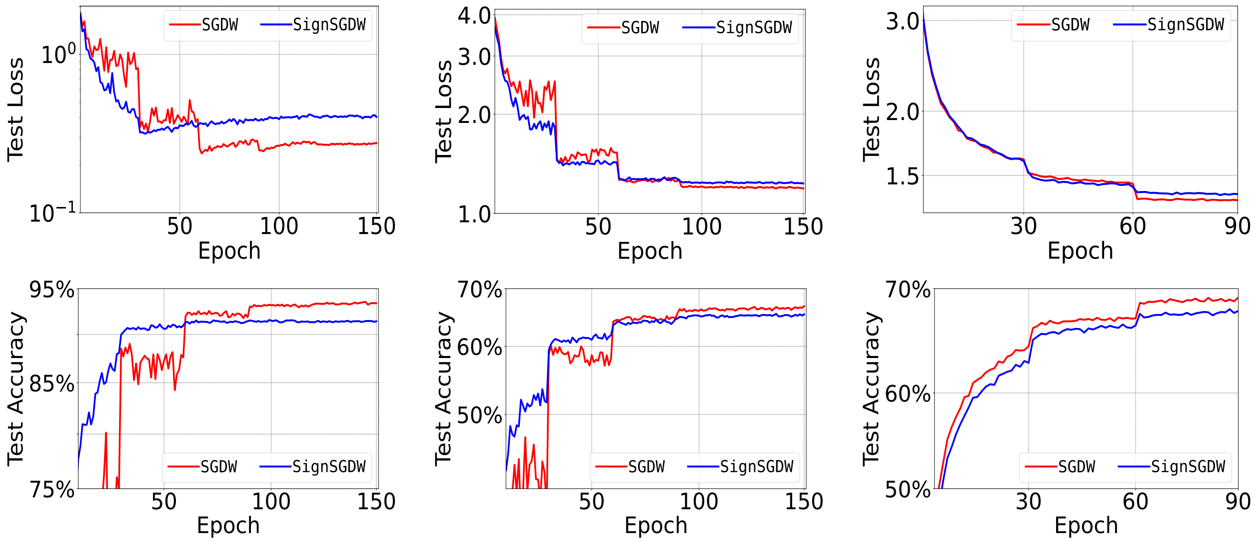
Figure 6. The generalization of SignSGDW, in terms of test accuracy and test loss, under different weight decay values on VGG16, ResNet20, and ResNet18 for Cifar10, Cifar100, and ImageNet classification, respectively

6.1. Effects of Weight Decay on Convergence of SGD

Theorem 1 implies that weight decay slows down the convergence of SGD when $\delta := \sup_{t\geq 0}\left\{\frac{|\langle \nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle|}{\|\nabla f_S(\boldsymbol{w}^t)\|\cdot\|\boldsymbol{w}^t\|}\right\} < 1$. We aim to confirm Theorem 1 by using VGG16 [31], ResNet20 [21], and ResNet18 [21] for Cifar10, Cifar100, and ImageNet [15] classification, respectively. In particular, we first monitor the value of δ in the training process and then contrast the convergence of SGD and SGDW.

Numerical Exploration for δ. For Cifar experiments, we use batch size 128 and the initial learning rate of 0.1 – which is decayed by a factor of 0.2 after every 30 epochs – for SGD. For ImageNet experiments, we use batch size 256 and an initial learning rate of 0.1, which is decayed by a factor of 0.2 after every 30 epochs – for SGD. We use the decoupled version of weight decay for SGD, which is given by $\alpha = 1 - \text{wd}\cdot\gamma$. In the Cifar experiment, weight decay (wd) values for SGDW are 5e-2, 1e-2, 5e-3, 1e-3, 5e-4, and 0 (no weight decay). In the ImageNet experiment, weight decay (wd) values for SGDW are 5e-4, 1e-5, 5e-6, and 0 (no weight decay). Figure 1 plots δ (averaged over each epoch) over training epochs, showing that δ is significantly less than 1 in all three experiments for all tested weight decay values. Indeed, we notice δ is much less than 1 for each training iteration as well.

Convergence of SGD vs. SGDW. The value of δ in Figure 1 together with Theorem 1 indicate that weight decay slows down the convergence of SGD. The convergence of SGDW under different weight decay values in the three benchmark experiments is depicted in Figure 2. These numerical results clearly show that increasing the weight decay does not improve the convergence of the training process; in contrast, it slows down the convergence. SGD without weight decay (wd=0) achieves the fastest convergence among all different weight decay values.



(a) VGG16 on Cifar10 (b) ResNet20 on Cifar100 (c) ResNet18 on ImageNet

Figure 7. The comparison between SGDW and SignSGDW with respect to generalization performance on VGG16, ResNet20, and ResNet18 for Cifar10, Cifar100, and ImageNet classification, respectively. We select the weight decay studied before that results in optimal test accuracy for SGDW and SignSGDW, respectively.

6.2. Effects of Weight Decay on Generalization of SGD

Theorem 2 shows that weight decay improves SGD generalization when $\delta < 1$, which aligns with our expectation that weight decay boosts test accuracy in the image classification tasks of Section 6.1, as $\delta < 1$ in all experiments. Figure 3 compares test accuracy and loss for SGD with different weight decay values. Except for wd = 5e-2, SGDW consistently outperforms SGD without weight decay in both test accuracy and loss. These results confirm that moderate weight decay enhances SGD generalization, as indicated by Theorem 2.

6.3. Numerical Results of SignSGDW

We follow the experimental setup in Section 6.1 and present three types of results: 1) Monitoring δ during SignSGDW training, 2) Convergence and generalization of SignSGDW with different weight decay values, and 3) A comparison of generalization between SignSGDW and SGDW. Using the hyperparameters from [34], we set the initial learning rate to 1e-3, momentum to 0.9 for SignSGDW, and applied the same learning rate decay schedule as in the SGDW experiments.

Numerical Exploration for δ. Figure 4 plots the values of δ along the trajectory of SignSGDW under different weight decay values. It is clear that δ is far less than one for all three tasks.

Convergence and Generalization of SignSGDW. We plot training loss vs. iteration in Figure 5 and plot the generalization of SignSGD in Figure 6. Similar to SGDW, weight decay does not accelerate the convergence of SignSGD but improves its generalization. Unlike SGDW, SignSGDW still converges and generalizes well even using a large weight decay value of 5e-2, and these empirical results are elucidated by Theorem 3: SignSGDW converges under a relatively larger weight decay value compared to SGDW when $\delta < 1$.

Comparison Between SignSGDW and SGDW. In previous experiments, we set the weight decay value (wd) of SGDW to be 1e-3, 1e-3, and 5e-6 for VGG16, ResNet20, and ResNet18, respectively. For SignSGDW, we set wd as 1e-3, 5e-3, and 1e-5 for VGG16, ResNet20, and ResNet18, respectively. These parameter choices result in satisfactory algorithmic generalization performance. The comparison between SGDW and SignSGDW is shown in Figure 7, showing that SGDW outperforms SignSGDW, resonating with Theorem 3.

7. Conclusion

This paper investigates the impact of weight decay on the convergence and generalization of SGD in nonconvex optimization, focusing on training DNNs without regularization. Our findings clarify that weight decay does not speed up training but enhances generalization, with SGDW generalizing better than SignSGDW. The convergence results are derived under mild assumptions, demonstrating that weight decay improves generalization for both SGD and SignSGD when gradient and weight parameters are not linearly dependent.

Acknowledgements

Tao Sun is sponsored in part by the National Natural Science Foundation of China under Grant No. 62376278, Young Elite Scientists Sponsorship Program by CAST (No. 2022QNRC001).

References

[1] Maksym Andriushchenko, Francesco D'Angelo, Aditya Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? *arXiv preprint arXiv:2310.04415*, 2023. 2

[2] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023. 3

[3] Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pages 404–413. PMLR, 2018. 5

[4] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020. 2

[5] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, pages 560–569. PMLR, 2018. 5

[6] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant. *arXiv*, 2018. 2, 5

[7] Johan Bjorck, Kilian Q Weinberger, and Carla Gomes. Understanding decoupled and early weight decay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6777–6785, 2021. 2

[8] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010. 1

[9] Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436. Springer, 2012. 1

[10] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2: 499–526, 2002. 2

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[12] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023. 5

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[14] Francesco D'Angelo, Maksym Andriushchenko, Aditya Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[16] Xiaoge Deng, Tao Sun, Shengwei Li, Dongsheng Li, and Xicheng Lu. Stability and generalization of asynchronous sgd: Sharper bounds beyond lipschitz and smoothness. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2

[17] Xiaoge Deng, Tao Sun, Shengwei Li, and Dongsheng Li. Stability-based generalization analysis of the asynchronous decentralized sgd. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7340–7348, 2023. 2

[18] Amin Ghiasi, Ali Shafahi, and Reza Ardekani. Improving robustness with adaptive weight decay. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[19] Stephen Hanson and Lorien Pratt. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, 1, 1988. 1

[20] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016. 2

[21] 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

[22] Geoffrey E Hinton. Learning translation invariant recognition in a massively parallel networks. In *International conference on parallel architectures and languages Europe*, pages 1–13. Springer, 1987. 1

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 1

[24] Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991. 1, 2

[25] Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 191–227. PMLR, 2023. 2, 4, 17

[26] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020. 2

[27] Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2019. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[29] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 1

[30] 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*, 115:211–252, 2015. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 7

[32] Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9756–9764, 2021. 2

[33] Tao Sun, Congliang Chen, Peng Qiao, Li Shen, Xinwang Liu, and Dongsheng Li. Rethinking sign training: Provable nonconvex acceleration without first-and second-order gradient lipschitz. *arXiv preprint arXiv:2310.14616*, 2023. 5

[34] Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of signsgd under weaker assumptions. *ICML*, 2023. 2, 5, 8, 19

[35] Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using sgd and weight decay. *Advances in Neural Information Processing Systems*, 34:6380–6391, 2021. 2

[36] Zeke Xie, Jingzhao Zhang, Issei Sato, Masashi Sugiyama, et al. On the overlooked pitfalls of weight decay and how to mitigate them: A gradient-norm perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3

[37] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 2

[38] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2018. 2

[39] Yi Zhou, Yingbin Liang, and Huishuai Zhang. Understanding generalization error of sgd in nonconvex optimization. *Machine Learning*, pages 1–31, 2022. 2