

Fixing Weight Decay Regularization in Adam

Ilya Loshchilov¹ Frank Hutter¹

Abstract

L_2 regularization and weight decay regularization are equivalent for standard stochastic gradient descent (when rescaled by the learning rate), but as we demonstrate this is *not* the case for adaptive gradient algorithms, such as Adam. While common deep learning frameworks of these algorithms implement L_2 regularization (often calling it “weight decay” in what may be misleading due to the inequivalence we expose), we propose a simple modification to recover the original formulation of weight decay regularization by decoupling the weight decay from the optimization steps taken w.r.t. the loss function. We provide empirical evidence that our proposed modification (i) decouples the optimal choice of weight decay factor from the setting of the learning rate for both standard SGD and Adam, and (ii) substantially improves Adam’s generalization performance, allowing it to compete with SGD with momentum on image classification datasets (on which it was previously typically outperformed by the latter). We also propose a version of Adam with warm restarts (AdamWR) that has strong anytime performance while achieving state-of-the-art results on CIFAR-10 and ImageNet32x32. Our source code is available at <https://github.com/loshchil/AdamW-and-SGDW>

1. Introduction

Adaptive gradient methods, such as AdaGrad (Duchi et al., 2011), RMSProp (Tieleman & Hinton, 2012), and Adam (Kingma & Ba, 2014) have become a default method of choice for training feed-forward and recurrent neural networks (Xu et al., 2015; Gregor et al., 2015; Radford et al., 2015). Nevertheless, state-of-the-art results for popular image classification datasets, such as CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), are still obtained by ap-

*Equal contribution ¹University of Freiburg, Germany. Correspondence to: Ilya Loshchilov <ilya.loshchilov@gmail.com>, Frank Hutter <fh@cs.uni-freiburg.de>.

plying SGD with momentum (Huang et al., 2016; 2017; Loshchilov & Hutter, 2016; Gastaldi, 2017). Furthermore, (Wilson et al., 2017) suggested that adaptive gradient methods do not generalize as well as SGD with momentum when tested on a diverse set of deep learning tasks such as image classification, character-level language modeling and constituency parsing. Different hypotheses about the origins of this worse generalization have been investigated, such as the presence of sharp local minima (Keskar et al., 2016; Dinh et al., 2017) and inherent problems of adaptive gradient methods (Wilson et al., 2017). In this paper, we show that a major factor of the poor generalization of the most popular adaptive gradient method, Adam, is due to the fact that L_2 regularization is not nearly as effective for it as for SGD. We also demonstrate that this problem can be resolved by using the original formulation of weight decay.

Specifically, our analysis of Adam given in this paper leads to the following observations:

L_2 regularization and weight decay are not identical.

Contrary to common belief, the two techniques are not equivalent. For SGD, they can be made equivalent by a reparameterization of the weight decay factor based on the learning rate; this is not the case for Adam. In particular, when combined with adaptive gradients, L_2 regularization leads to weights with large gradients being regularized less than they would be when using weight decay.

L_2 regularization is not effective in Adam. One possible explanation why Adam and other adaptive gradient methods might be outperformed by SGD with momentum is that common deep learning libraries only implement L_2 regularization, not the original weight decay. Therefore, on tasks/datasets where the use of L_2 regularization is beneficial for SGD (e.g., on many popular image classification datasets), Adam leads to worse results than SGD with momentum (for which L_2 regularization behaves as expected).

Weight decay is equally effective in both SGD and Adam. For SGD, it is equivalent to L_2 regularization, while for Adam it is not.

Optimal weight decay depends on the total number of batch passes/weight updates. Our empirical analysis

of Adam suggests that the larger the runtime/number of batch passes to be performed, the smaller the optimal weight decay. This effect tends to be neglected because hyperparameters are often tuned for a fixed or a comparable number of training epochs. As a result, the values of the weight decay found to perform best for short runs do not generalize to much longer runs.

Our contributions are aimed at fixing the issues described above and improving Adam's performance in practice:

Decoupling weight decay from the gradient-based update (Section 2). We suggest to use the original formulation of weight decay to decouple the gradient-based update from weight decay for both SGD and Adam. The resulting SGD version with weight decay (short: SGDW) decouples optimal settings of the learning rate and the weight decay factor (see Figure 2), and *the resulting Adam version with weight decay (short: AdamW) generalizes substantially better than Adam with L₂ regularization*, achieving 15% relative improvement in test error both on CIFAR-10 and ImageNet32x32 (see Figures 2 and 3).

Formal analysis of weight decay vs. L₂ regularization (Section 3). We formalize in which ways the two regularization methods differ and provide intuition as to why weight decay regularization may lead to better generalization performance.

Normalizing the values of weight decay (Section 4). We propose to parameterize the weight decay factor as a function of the total number of batch passes. This leads to a greater invariance of the hyperparameter settings in the sense that the values found to perform best for short runs also perform well for many times longer runs.

Adam with cosine annealing and warm restarts (Section 5). We combine Adam with the cosine annealing learning schedule and warm restarts of the learning rate in order to obtain both strong anytime performance and strong final performance. Empirical results demonstrate that *using cosine annealing reduces Adam's test error by a relative 15%* (see Figure 1), and *warm restarts improve its anytime performance to find good solutions up to 10x faster* (see Figure 4).

The main motivation of this paper is to improve regularization in Adam to make it competitive w.r.t. SGD with momentum even for those problems where it did not use to be competitive. We also improve Adam's practical performance based on cosine annealing and warm restarts, which had previously only been demonstrated to be efficient for

Algorithm 1 SGD with L₂ regularization and SGD with weight decay (SGDW), both with momentum

```

1: given initial learning rate  $\alpha \in \mathbb{R}$ , momentum factor  $\beta_1 \in \mathbb{R}$ , weight decay / L2 regularization factor  $w \in \mathbb{R}$ 
2: initialize time step  $t \leftarrow 0$ , parameter vector  $\mathbf{x}_{t=0} \in \mathbb{R}^n$ , first moment vector  $\mathbf{m}_{t=0} \leftarrow \mathbf{0}$ , schedule multiplier  $\eta_{t=0} \in \mathbb{R}$ 
3: repeat
4:    $t \leftarrow t + 1$ 
5:    $\nabla f_t(\mathbf{x}_{t-1}) \leftarrow \text{SelectBatch}(\mathbf{x}_{t-1})$   $\triangleright$  select batch and return the corresponding gradient
6:    $\mathbf{g}_t \leftarrow \nabla f_t(\mathbf{x}_{t-1}) + w\mathbf{x}_{t-1}$ 
7:    $\eta_t \leftarrow \text{SetScheduleMultiplier}(t)$   $\triangleright$  can be fixed, decay, be used for warm restarts
8:    $\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + \eta_t \alpha \mathbf{g}_t$ 
9:    $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \mathbf{m}_t - \eta_t w \mathbf{x}_{t-1}$ 
10:  until stopping criterion is met
11: return optimized parameters  $\mathbf{x}_t$ 
```

SGD with momentum. We hope that as a result, practitioners do not need to switch between Adam and SGD anymore, which in turn should help to reduce the common issue of selecting dataset/task-specific training algorithms and their hyperparameters.

2. Decoupling the Weight Decay from the gradient-based update

In the weight decay described by Hanson & Pratt (1988), the weights \mathbf{x} decay exponentially as

$$\mathbf{x}_{t+1} = (1 - w)\mathbf{x}_t - \alpha \nabla f_t(\mathbf{x}_t), \quad (1)$$

where w defines the rate of the weight decay per step and $\nabla f_t(\mathbf{x}_t)$ is the t -th batch gradient to be multiplied by a learning rate α . Following Hanson & Pratt (1988), one can also modify the original batch loss $f_t(\mathbf{x}_t)$ and consider a bias term (also referred to as the regularization term) accounting for “costs” on weights which are, e.g., quadratic in the weight values as for L₂ regularization:

$$f_t^{\text{reg}}(\mathbf{x}_t) = f_t(\mathbf{x}_t) + \frac{w}{2} \|\mathbf{x}_t\|_2^2, \quad (2)$$

where w defines the impact of the L₂ regularization. In order to consider the weight decay regularization, one can reformulate the objective function as in Eq. (2) or directly adjust $\nabla f_t(\mathbf{x}_t)$ as

$$\nabla f_t^{\text{reg}}(\mathbf{x}_t) = \nabla f_t(\mathbf{x}_t) + w\mathbf{x}_t. \quad (3)$$

Historically, stochastic gradient descent methods inherited this way of implementing the weight decay regularization.

The currently most common way (e.g., in popular libraries such as TensorFlow, Keras, PyTorch, Torch, and Lasagne) to introduce weight decay regularization is to use the L₂ regularization term as in Eq. (2) or, often equivalently, to directly

Algorithm 2 Adam with L₂ regularization and Adam with weight decay (AdamW)

```

1: given  $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, w \in \mathbb{R}$ 
2: initialize time step  $t \leftarrow 0$ , parameter vector  $\mathbf{x}_{t=0} \in \mathbb{R}^n$ , first moment vector  $\mathbf{m}_{t=0} \leftarrow \mathbf{0}$ , second moment vector  $\mathbf{v}_{t=0} \leftarrow \mathbf{0}$ , schedule multiplier  $\eta_{t=0} \in \mathbb{R}$ 
3: repeat
4:    $t \leftarrow t + 1$ 
5:    $\nabla f_t(\mathbf{x}_{t-1}) \leftarrow \text{SelectBatch}(\mathbf{x}_{t-1})$   $\triangleright$  select batch and return the corresponding gradient
6:    $\mathbf{g}_t \leftarrow \nabla f_t(\mathbf{x}_{t-1}) + w\mathbf{x}_{t-1}$ 
7:    $\mathbf{m}_t \leftarrow \beta_1\mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t$   $\triangleright$  here and below all operations are element-wise
8:    $\mathbf{v}_t \leftarrow \beta_2\mathbf{v}_{t-1} + (1 - \beta_2)\mathbf{g}_t^2$ 
9:    $\hat{\mathbf{m}}_t \leftarrow \mathbf{m}_t / (1 - \beta_1^t)$   $\triangleright \beta_1$  is taken to the power of  $t$ 
10:   $\hat{\mathbf{v}}_t \leftarrow \mathbf{v}_t / (1 - \beta_2^t)$   $\triangleright \beta_2$  is taken to the power of  $t$ 
11:   $\eta_t \leftarrow \text{SetScheduleMultiplier}(t)$   $\triangleright$  can be fixed, decay, or also be used for warm restarts
12:   $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_t (\alpha\hat{\mathbf{m}}_t / (\sqrt{\hat{\mathbf{v}}_t} + \epsilon) + w\mathbf{x}_{t-1})$ 
13: until stopping criterion is met
14: return optimized parameters  $\mathbf{x}_t$ 

```

modify the gradient as in Eq. (3). Let's first consider the simple case of SGD with momentum; Algorithm 1 demonstrates modifying the gradients directly in this method (see line 6). The weight decay term $w\mathbf{x}_{t-1}$ will first modify \mathbf{g}_t (see line 6) and then affect the momentum term \mathbf{m}_t (see line 8). While the smoothing of the weight decay factor by β_1 (see line 8) might be a feature, we note (for simplicity, we omit η_t) that \mathbf{x}_t will decay by $\alpha w\mathbf{x}_{t-1}$ (see line 9) and not $w\mathbf{x}_{t-1}$ as one could expect according to the definition of weight decay given by Eq. (1). Practically, if one wants to keep the actual weight decay αw fixed while changing α to α' , then w should be modified to $w'_t = \frac{\alpha w}{\alpha'}$. This renders the problem of hyperparameter selection of α and w non-separable.

We propose to fix this problem by following the original definition of weight decay given by Eq. (1) and decay the weights simultaneously with the update of \mathbf{x}_t based on gradient information in Line 9 of Algorithm 1. This yields our proposed variant of SGD with momentum using weight decay (short: **SGDW**). This simple modification explicitly decouples w and α (although some problem-dependent implicit coupling may of course remain as for any two hyperparameters). In order to account for a possible scheduling of both α and w , we introduce a scaling factor η_t delivered by a user-defined procedure *SetScheduleMultiplier*(t). It should be noted that when L₂ regularization is used, weight decay contributes to the batch gradient and thus effectively is scheduled in the same way as the learning rate. Now, since we decouple the two we should also remember to schedule both of them with η_t .

Having shown that using L₂ regularization instead of weight

decay already couples regularization and learning rate in the simple case of SGD with momentum, we now consider adaptive gradient optimizers, such as the Adam algorithm proposed by Kingma & Ba (2014), in which the coupling leads to even more unintended behavior. As an adaptive gradient method, Adam maintains a vector \mathbf{v}_t responsible for storing smoothed amplitudes of parameter-wise gradients \mathbf{g}_t^2 (see line 8 in Algorithm 2). These factors are used to control parameter-wise learning rates by normalizing parameter-wise gradients by $\sqrt{\hat{\mathbf{v}}_t} + \epsilon$ in line 12 of Algorithm 2. The common way to introduce the weight decay $w\mathbf{x}_{t-1}$ to Adam results in an update which only distantly resembles the original weight decay given by Eq. (1), because the \mathbf{v}_t vectors keep track of amplitudes of not only the loss-based gradients, but also the weights. The amplitudes are then used to re-normalize \mathbf{m}_t as given in line 12 of Algorithm 2. To gain a bit of intuition, let us consider the case when t is large, causing β_1^t and β_2^t to go to zero and

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_t \alpha \frac{\beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t}{\sqrt{\beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2 + \epsilon}}, \quad (4)$$

with $\mathbf{g}_t = \nabla f_t(\mathbf{x}_{t-1}) + w\mathbf{x}_{t-1}$,

where operations are performed parameter-wise. Not only the batch gradient $\nabla f_t(\mathbf{x}_{t-1})$ is normalized but also the weight decay $w\mathbf{x}_{t-1}$ itself. Since this formula normalizes updates by their typical amplitudes, weights are not decayed proportionally to their amplitudes anymore, leading to the relative decay being weaker for weights with large gradients. *This is a correct implementation of L₂ regularization, but not of weight decay.* Therefore, it might be misleading to use the two terms interchangeably, as is commonly done in the literature. We note that this difference between the two mechanisms has not been investigated and/or described before. As in the case of SGDW, we propose to follow the original definition of weight decay and perform it simultaneously with the gradient-based update as shown in line 12 of Algorithm 2; this gives rise to our variant of Adam with weight decay (short: **AdamW**). As we will demonstrate experimentally (in Section 6.3), AdamW generalizes much better than Adam with standard L₂ regularization.

3. Weight Decay vs L₂ regularization

We now formalize the findings of the previous section.

Proposition 1 (Weight decay = L₂ reg for standard SGD). *Standard SGD with base learning rate α executes the same steps on batch loss functions $f_t(\mathbf{x})$ with weight decay w (defined in Equation 1) as it executes without weight decay on $f_t^{reg}(\mathbf{x}) = f_t(\mathbf{x}) + \frac{w'}{2} \|\mathbf{x}\|_2^2$, with $w' = \frac{w}{\alpha}$.*

Proposition 2 (Weight decay \neq L₂ reg for adaptive gradients). *Let O denote an optimizer that has iterates $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \alpha \mathbf{M}_t \nabla f_t(\mathbf{x}_t)$ when run on batch loss function $f_t(\mathbf{x})$ without weight decay, and $\mathbf{x}_{t+1} \leftarrow (1-w)\mathbf{x}_t - \alpha \mathbf{M}_t \nabla f_t(\mathbf{x}_t)$*

when run on $f_t(\mathbf{x})$ with weight decay, respectively, with $\mathbf{M}_t \neq k\mathbf{I}$ (where $k \in \mathbb{R}$). Then, for O there exists no L_2 coefficient w' such that running O on batch loss $f_t^{reg}(\mathbf{x}) = f_t(\mathbf{x}) + \frac{w'}{2} \|\mathbf{x}\|_2^2$ without weight decay is equivalent to running O on $f_t(\mathbf{x})$ with decay $w \in \mathbb{R}^+$.

All proofs are given in the supplementary material, Appendix A. While Proposition 1 simply restates the well-known equivalence of weight decay and L_2 regularization for standard SGD, Proposition 2 shows (for what we believe to be the first time) that this equivalence does not hold for adaptive gradient methods.

Having shown that L_2 regularization and weight decay regularization differ for adaptive gradient algorithms raises the question of how they differ and how to interpret their effects. Their equivalence for standard SGD remains very helpful for intuition: both mechanisms push weights closer to zero, at the same rate. However, for adaptive gradient algorithms they differ: with L_2 regularization, these adapt the summed gradient of the regularizer and the loss function, whereas with weight decay they only adapt the gradients of the loss function (with the weight decay step separated from the adaptive gradient mechanism). With L_2 regularization both types of gradients are normalized by their typical (summed) magnitudes, and therefore weights x with large typical gradient magnitude s are regularized by a smaller relative amount than other weights. In contrast, weight decay regularizes all weights by the same factor, effectively regularizing weights x with large s more than standard L_2 regularization does.

We now demonstrate this formally for a simple special case of adaptive gradient algorithm with a fixed preconditioner:

Proposition 3 (Weight decay = scale-adjusted L2 reg for adaptive gradient algorithm with fixed preconditioner). *Let O denote an algorithm with the same characteristics as in Proposition 2, and using a fixed preconditioner matrix $\mathbf{M}_t = \text{diag}(s)^{-1}$ (with $s_i > 0$ for all i). Then, O with base learning rate α executes the same steps on batch loss functions $f_t(\mathbf{x})$ with weight decay w as it executes without weight decay on the scale-adjusted regularized batch loss*

$$f_t^{sreg}(\mathbf{x}) = f_t(\mathbf{x}) + \frac{w'}{2\alpha} \|\mathbf{x} \odot \mathbf{s}\|_2^2, \quad (5)$$

where \odot denotes element-wise multiplication and $w' = \frac{w}{\alpha}$.

If \mathbf{s} contains the typical gradient magnitudes, the intuition behind regularizing by $\|\mathbf{x} \odot \mathbf{s}\|$ rather than just $\|\mathbf{x}\|$ is that we would like to penalize weights whose gradients tend to be large, because small variations in those weights have large effects, meaning that they lead to brittle solutions. This also relates to the idea of sharp vs. flat local optima (Keskar et al., 2016): penalizing weights with (typically) large gradients more heavily may drive the search away from areas where those gradients are large and thus lead to finding flatter optima with better generalization performance.

In practical adaptive gradient algorithms, the preconditioner matrix $\mathbf{M}_t = \text{diag}(s)^{-1}$ is of course not fixed but is modified over time. Nevertheless, the intuition from above still carries over, just with a preconditioner that adapts at each step. We would like to mention that this adaptation couples the design of the cost function and its optimization, which at first glance appears suboptimal.¹ However, we note that the modification of the preconditioner during the optimization amounts to a regularizer that *automatically adapts* to the shape of the function, an advantage that cannot be encoded into a single fixed cost function.

4. Normalized Weight Decay

Our preliminary experiments showed that different weight decay factors are optimal for different computational budgets (defined in terms of the number of batch passes). Relatedly, Li et al. (2017) demonstrated that a smaller batch size (for the same total number of epochs) leads to the shrinking effect of weight decay being more pronounced. Here, we propose to reduce this dependence by normalizing the values of weight decay. Specifically, we replace the hyperparameter w by a new (more robust) normalized weight decay hyperparameter w_{norm} , and use this to set w as follows:

$$w = w_{norm} \sqrt{\frac{b}{BT}}, \quad (6)$$

where b is the batch size, B is the total number of training points per epoch and T is the total number of epochs.² Thus, w_{norm} can be interpreted as the weight decay to be used if only one batch pass is allowed.

5. Adam with Cosine Annealing and Warm Restarts

We now apply cosine annealing and warm restarts to Adam, following the recent work of Loshchilov & Hutter (2016). There, the authors proposed Stochastic Gradient Descent with Warm Restarts (SGDR) to improve anytime performance of SGD by quickly cooling down the learning rate according to a cosine schedule and periodically increasing it. SGDR has been successfully adopted to lead to new state-of-the-art results for popular image classification benchmarks (Huang et al., 2017; Gastaldi, 2017; Zoph et al., 2017), and we therefore already tried extending it to Adam shortly after it was proposed. However, while our initial version of Adam with warm restarts had better anytime performance than Adam, it was not competitive with SGD with warm

¹There are, however, many other regularizers that also couple the two, such as early stopping, the regularizing effect of gradient noise, or recent regularization techniques by Kawaguchi et al. (2017), which are even specific to the current batch and its size.

²In the context of our AdamWR variant discussed in the next section, T is the total number of epochs in the current restart.

restarts, precisely because L₂ regularization was not working as well as in SGD. Now, having fixed this issue by means of the original weight decay regularization (Section 2) and also having introduced normalized weight decay (Section 4), the original work on cosine annealing and warm restarts by Loshchilov & Hutter (2016) directly carries over to Adam.

In the interest of keeping the presentation self-contained, we briefly describe how SGDR schedules the change of the effective learning rate in order to accelerate the training of DNNs. Here, we decouple the initial learning rate α and its multiplier η_t used to obtain the actual learning rate at iteration t (see, e.g., line 8 in Algorithm 1). In SGDR, we simulate a new warm-started run/restart of SGD once T_i epochs are performed, where i is the index of the run. Importantly, the restarts are not performed from scratch but emulated by increasing η_t while the old value of x_t is used as an initial solution. The amount by which η_t is increased controls to which extent the previously acquired information (e.g., momentum) is used. Within the i -th run, the value of η_t decays according to a cosine annealing learning rate for each batch as follows:

$$\eta_t = \eta_{min}^{(i)} + 0.5(\eta_{max}^{(i)} - \eta_{min}^{(i)})(1 + \cos(\pi T_{cur}/T_i)), \quad (7)$$

where $\eta_{min}^{(i)}$ and $\eta_{max}^{(i)}$ are ranges for the multiplier and T_{cur} accounts for how many epochs have been performed since the last restart. T_{cur} is updated at each batch iteration t and is thus not constrained to integer values. Adjusting (e.g., decreasing) $\eta_{min}^{(i)}$ and $\eta_{max}^{(i)}$ at every i -th restart (see also (Smith, 2016)) could potentially improve performance, but we do not consider that option in our experiments because it would involve additional hyperparameters. For $\eta_{max}^{(i)} = 1$ and $\eta_{min}^{(i)} = 0$, one can simplify Eq. (7) to

$$\eta_t = 0.5 + 0.5 \cos(\pi T_{cur}/T_i). \quad (8)$$

In order to achieve good anytime performance, one can start with an initially small T_i (e.g., from 1% to 10% of the expected total budget) and multiply it by a factor of T_{mult} (e.g., $T_{mult} = 2$) at every restart. The $(i+1)$ -th restart is triggered when $T_{cur} = T_i$ by setting T_{cur} to 0. An example setting of the schedule multiplier is given in Section B of the supplementary material.

Our proposed **AdamWR** algorithm represents AdamW (see Algorithm 2) with η_t following Eq. (8) and w computed at each iteration using normalized weight decay according to Eq. (6). We note that normalized weight decay allowed us to use a constant parameter setting across short and long runs performed within AdamWR. Equivalently to AdamWR, we define SGDWR as SGDW with warm restarts.

6. Experimental Validation

Our experimental setup follows that of Gastaldi (2017), who proposed, in addition to L₂ regularization, to apply the new

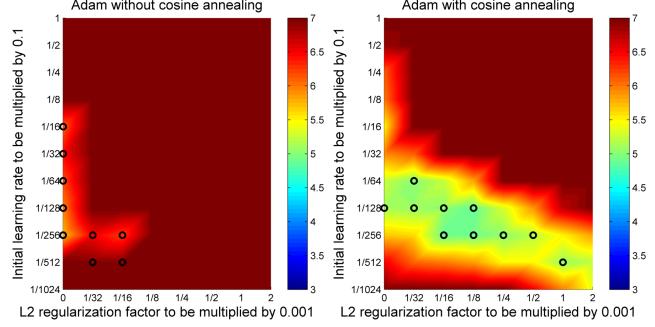


Figure 1. We show the final test error of a 26 2x64d ResNet on CIFAR-10 after 100 epochs of standard Adam with L₂ regularization, with fixed learning rate (left) and with cosine annealing (right). Cosine annealing yields clearly superior results. See Supp-Figure 2 (in the supplementary material) for a comparison with a step-drop learning rate schedule.

Shake-Shake regularization to a 3-branch residual neural network. Gastaldi (2017) showed that this regularization allowed to achieve new state-of-the-art results of 2.86% on the CIFAR-10 dataset (Krizhevsky, 2009). The network was trained by SGDR with batch size 128 for 1800 epochs ($T_0 = 1800$) without restarts with the learning rate scheduled by Eq. (7). The regular data augmentation procedure used for the CIFAR datasets was applied. We used the same model/source code based on fb.resnet.torch³. The base networks are a 26 2x64d ResNet (i.e. the network has a depth of 26, 2 residual branches and the first residual block has a width of 64) and a 26 2x96d ResNet with 11.6M and 25.6M parameters, respectively. For a detailed description of the network and the Shake-Shake method, we refer the interested reader to Gastaldi (2017).

6.1. Adam with Cosine Annealing

We first compared Adam’s performance with and without cosine annealing; for each method, we trained a 2x64d ResNet with a budget of 100 epochs with different settings of α and w . As the results in Figure 1 show, the best base learning rate with cosine annealing yields roughly 20% lower test errors than the best fixed learning rate. In SuppFigure 2 (in the supplementary material), we also show, for both Adam and AdamW, that cosine annealing outperforms a step-drop learning rate schedule, which in turn outperforms a fixed learning rate. Based on these results, we scheduled the learning rate with cosine annealing throughout this paper.

6.2. Decoupling The Weight Decay and Initial Learning Rate Parameters

In order to verify our hypothesis about the coupling of the initial learning rate α and the weight decay factor w ,

³<https://github.com/xgastaldi/shake-shake>

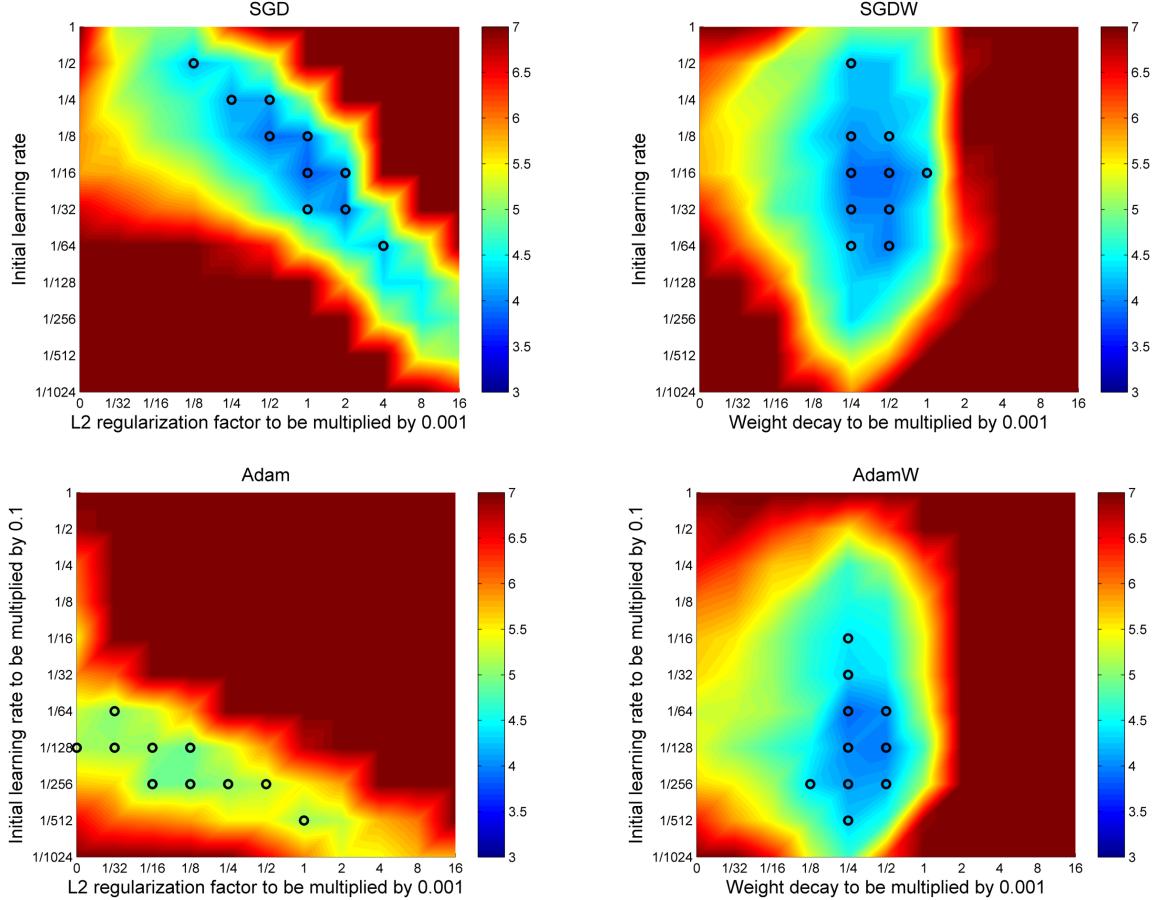


Figure 2. The Top-1 test error of a 26 2x64d ResNet on CIFAR-10 measured after 100 epochs. The proposed SGDW and AdamW (right column) have a more separable hyperparameter space.

we trained a 2x64d ResNet with cosine annealing for 100 epochs, using different settings of α and w . Figure 2 compares the performance of L₂ regularization vs. weight decay in SGD (SGD vs. SGDW, top row) and in Adam (Adam vs. AdamW, bottom row). In SGD (Figure 2, top left), L₂ regularization is not decoupled from the learning rate (the common way as described in Algorithm 1), and the figure clearly shows that the basin of best hyperparameter settings (depicted by color and top-10 hyperparameter settings by black circles) is not aligned with the x-axis or y-axis but lies on the diagonal. This suggests that the two hyperparameters are interdependent and need to be changed simultaneously, while only changing one of them might substantially worsen results. Consider, e.g., the setting at the top left black circle ($\alpha = 1/2$, $w = 1/8 * 0.001$); only changing either α or w by itself would worsen results, while changing both of them could still yield clear improvements. We note that this coupling of initial learning rate and L₂ regularization factor might have contributed to SGD’s reputation of being very sensitive to its hyperparameter settings.

In contrast, the results for SGD with weight decay (SGDW) in Figure 2 (top right) show that weight decay and initial learning rate are decoupled. The proposed approach renders the two hyperparameters more separable: even if the learning rate is not well tuned yet (e.g., consider the value of 1/1024 in Figure 2, top right), leaving it fixed and only optimizing the weight decay factor would yield a good value (of 1/4*0.001). This is not the case for SGD with L₂ regularization, shown in Figure 2 (top left).

The results for Adam with L₂ regularization are given in Figure 2 (bottom left). Adam’s best hyperparameter settings performed clearly worse than SGD’s best ones (compare Figure 2, top left). While both methods used L₂ regularization, Adam did not benefit from it at all: its best results obtained for non-zero L₂ regularization factors were comparable to the best ones obtained without the L₂ regularization, i.e., when $w = 0$. Similarly to the original SGD, the shape of the hyperparameter landscape suggests that the two hyperparameters are coupled.

In contrast, the results for our new variant of Adam with

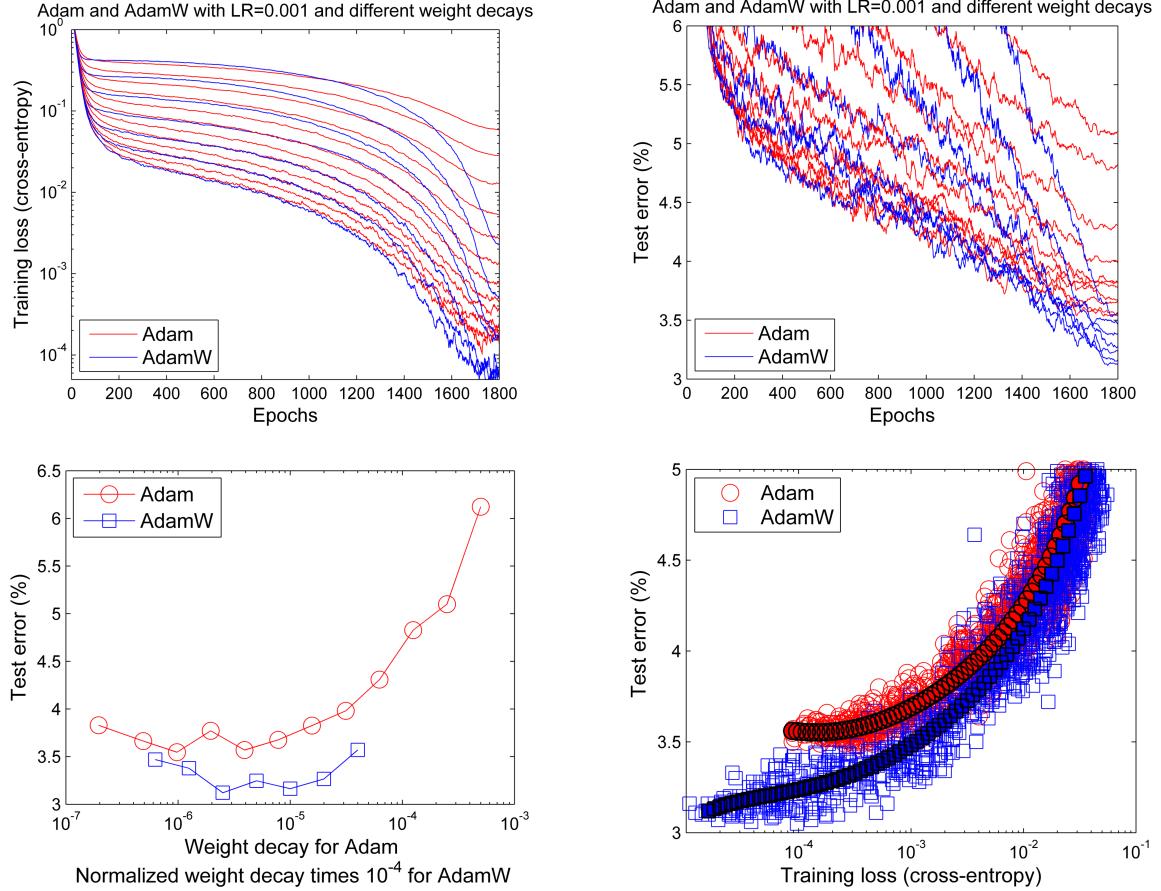


Figure 3. Learning curves (top row) and generalization results (bottom row) obtained by a 26 2x96d ResNet trained with Adam and AdamW on CIFAR-10. See text for details. SuppFigure 5 (supplemental material) shows the same qualitative results for ImageNet32x32.

weight decay (AdamW) in Figure 2 (bottom right) show that AdamW largely decouples weight decay and learning rate. The results for the best hyperparameter settings were substantially better than the best ones of Adam with L₂ regularization and rivaled those of SGD and SGDW.

In summary, the results in Figure 2 support our hypothesis that the weight decay and learning rate hyperparameters can be decoupled, and that this in turn simplifies the problem of hyperparameter tuning in SGD and improves Adam's performance to be competitive w.r.t. SGD with momentum.

6.3. Better Generalization of AdamW

While the previous experiment suggested that the basin of optimal hyperparameters of AdamW is broader and deeper than the one of Adam, we next investigated the results for much longer runs of 1800 epochs to compare the generalization capabilities of AdamW and Adam.

We fixed the initial learning rate to 0.001 which represents both the default learning rate for Adam and the one which showed reasonably good results in our experiments. Figure 3 shows the results for 12 settings of the L₂ weight decay

of Adam and 7 settings of the normalized weight decay of AdamW. Interestingly, while the dynamics of the learning curves of Adam and AdamW often coincided for the first half of the training run, AdamW often led to lower training loss and test errors (see Figure 3 top left and top right, respectively). Importantly, the use of weight decay in Adam did not yield as good results as in AdamW (see also Figure 3, bottom left). Next, we investigated whether AdamW's better results were only due to better convergence or due to better generalization. *The results in Figure 3 (bottom right) for the best settings of Adam and AdamW suggest that AdamW did not only yield better training loss but also yielded better generalization performance for similar training loss values.* The results on ImageNet32x32 (see SuppFigure 5 in the supplementary material) lead to the same conclusion of substantially improved generalization performance.

6.4. Easier Hyperparameter Selection due to Normalized Weight decay

Our experimental results with Adam and SGD suggested that the total runtime in terms of the number of epochs affect the basin of optimal hyperparameters (see SuppFigure 4 in

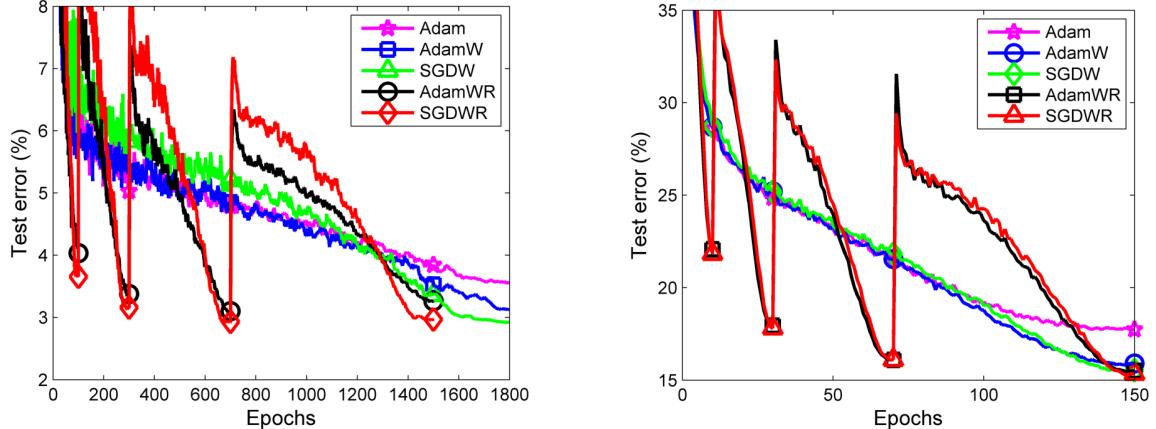


Figure 4. Top-1 test error on CIFAR-10 (left) and Top-5 test error on ImageNet32x32 (right).

the supplementary material). More specifically, the greater the total number of epochs the smaller the values of the weight decay should be. SuppFigure 4 shows that our remedy for this problem, the normalized weight decay defined in Eq. (8), simplifies hyperparameter selection because the optimal values observed for short runs are similar to the ones for much longer runs. We used our initial experiments on CIFAR-10 to suggest the square root normalization we proposed in Eq. (8) and double-checked that this is not a coincidence on the ImageNet32x32 dataset (Chrabaszcz et al., 2017), a downsampled version of the original ImageNet dataset with 1.2 million 32×32 pixels images, where an epoch is 24 times longer than on CIFAR-10. This experiment also supported the square root scaling: the best values of the normalized weight decay observed on CIFAR-10 represented nearly optimal values for ImageNet32x32 (see SuppFigure 4). In contrast, had we used the same raw weight decay values w for ImageNet32x32 as for CIFAR-10 and for the same number of epochs, *without the proposed normalization*, w would have been roughly 5 times too large for ImageNet32x32, leading to much worse performance. The optimal normalized weight decay values were also very similar (e.g., $w_{norm} = 0.025$ and $w_{norm} = 0.05$) across SGDWR and AdamW.

6.5. AdamWR with Warm Restarts for better anytime performance

Finally, we investigated the strong anytime performance AdamWR obtains from warm restarts. As Figure 4 shows, AdamWR greatly sped up AdamW on CIFAR-10 and ImageNet32x32, up to a factor of 10 (see the results at the first restart). For the default learning rate of 0.001, AdamW achieved 15% relative improvement in test errors compared to Adam both on CIFAR-10 (also see Figure 3) and ImageNet32x32 (also see SuppFigure 5). AdamWR achieved the same improved results but with a much better anytime performance. These improvements closed most of the gap

between Adam and SGDWR on CIFAR-10 and yielded comparable performance on ImageNet32x32.

7. Conclusion and Future Work

Following suggestions that adaptive gradient methods such as Adam might lead to worse generalization than SGD with momentum (Wilson et al., 2017), we identified at least one possible explanation to this phenomenon: the inequivalence of L_2 regularization and weight decay we expose. We empirically showed that our version of Adam with the original formulation of weight decay yields substantially better generalization performance than the common implementation of Adam with L_2 regularization. We also proposed normalized weight decay and the use of cosine annealing and warm restarts for Adam, resulting in a more robust hyperparameter selection, a better final performance and a better anytime performance, respectively.

Our results obtained on image classification datasets must be verified on a wider range of tasks, especially ones where the use of regularization is expected to be important. It would be interesting to integrate our findings on weight decay into other methods which attempt to improve Adam, e.g., normalized direction-preserving Adam (Zhang et al., 2017). While we focussed our experimental analysis on Adam, we believe that similar results also hold for other adaptive gradient methods, such as AdaGrad (Duchi et al., 2011), RMSProp (Tieleman & Hinton, 2012), and AMSGrad (Reddi et al., 2018).

Advani & Saxe (2017) analytically showed that in the limited data regime of deep networks the presence of eigenvalues that are zero forms a frozen subspace in which no learning occurs and thus smaller (e.g., zero) initial weight norms should be used to achieve best generalization results. We thus plan to consider adapting initial weight norms or weight norm constraints (Salimans & Kingma, 2016) at each warm restart.

References

- Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv:1710.03667*, 2017.
- Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of ImageNet as an alternative to the CIFAR datasets. *arXiv:1707.08819*, 2017.
- Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *arXiv:1703.04933*, 2017.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.
- Xavier Gastaldi. Shake-Shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.
- Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv:1502.04623*, 2015.
- Stephen José Hanson and Lorien Y Pratt. Comparing biases for minimal network construction with back-propagation. In *Proceedings of the 1st International Conference on Neural Information Processing Systems*, pp. 177–185, 1988.
- Gao Huang, Zhuang Liu, and Kilian Q Weinberger. Densely connected convolutional networks. *arXiv:1608.06993*, 2016.
- Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv:1704.00109*, 2017.
- Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv:1710.05468*, 2017.
- Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv:1609.04836*, 2016.
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.
- Hao Li, Zheng Xu, Gavin Taylor, and Tom Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.
- Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. *arXiv:1608.03983*, 2016.
- Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.
- Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *International Conference on Learning Representations*, 2018.
- Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 901–909, 2016.
- Leslie N Smith. Cyclical learning rates for training neural networks. *arXiv:1506.01186v3*, 2016.
- 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.
- Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *arXiv:1705.08292*, 2017.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057, 2015.
- Zijun Zhang, Lin Ma, Zongpeng Li, and Chuan Wu. Normalized direction-preserving adam. *arXiv:1709.04546*, 2017.
- Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *arXiv:1707.07012 [cs.CV]*, 2017.

Supplementary material

A. Proofs

Proof of Proposition 1

The proof for this well-known fact is straight-forward. SGD without weight decay has the following iterates on $f_t^{\text{reg}}(\mathbf{x}) = f_t(\mathbf{x}) + \frac{w'}{2} \|\mathbf{x}\|_2^2$:

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \alpha \nabla f_t^{\text{reg}}(\mathbf{x}_t) = \mathbf{x}_t - \alpha \nabla f_t(\mathbf{x}_t) - \alpha w' \mathbf{x}_t. \quad (9)$$

SGD with weight decay has the following iterates on $f_t(\mathbf{x})$:

$$\mathbf{x}_{t+1} \leftarrow (1-w)\mathbf{x}_t - \alpha \nabla f_t(\mathbf{x}_t). \quad (10)$$

These iterates are identical since $w' = \frac{w}{\alpha}$. \square

Proof of Proposition 2

Similarly to the Proof of Proposition 1, the iterates of O without weight decay on $f_t^{\text{reg}}(\mathbf{x}) = f_t(\mathbf{x}) + \frac{1}{2}w' \|\mathbf{x}\|_2^2$ and O with weight decay w on f_t are, respectively:

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \alpha w' \mathbf{M}_t \mathbf{x}_t - \alpha \mathbf{M}_t \nabla f_t(\mathbf{x}_t). \quad (11)$$

$$\mathbf{x}_{t+1} \leftarrow (1-w)\mathbf{x}_t - \alpha \mathbf{M}_t \nabla f_t(\mathbf{x}_t). \quad (12)$$

The equality of these iterates for all \mathbf{x}_t would imply $w\mathbf{x}_t = \alpha w' \mathbf{M}_t \mathbf{x}_t$. This can only hold for all \mathbf{x}_t if $\mathbf{M}_t = k\mathbf{I}$, with $k \in \mathbb{R}$, which is not the case for O . Therefore, no L₂ regularizer $w' \|\mathbf{x}\|_2^2$ exists that makes the iterates equivalent. \square

Proof of Proposition 3

O without weight decay has the following iterates on $f_t^{\text{reg}}(\mathbf{x}) = f_t(\mathbf{x}) + \frac{w'}{2} \|\mathbf{x} \odot \mathbf{s}\|_2^2$:

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \alpha \nabla f_t^{\text{reg}}(\mathbf{x}_t)/s \quad (13)$$

$$= \mathbf{x}_t - \alpha \nabla f_t(\mathbf{x}_t)/s - \alpha w' \mathbf{x}_t \odot \mathbf{s}/s \quad (14)$$

$$= \mathbf{x}_t - \alpha \nabla f_t(\mathbf{x}_t)/s - \alpha w' \mathbf{x}_t, \quad (15)$$

where the division by s is element-wise. O with weight decay has the following iterates on $f_t(\mathbf{x})$:

$$\mathbf{x}_{t+1} \leftarrow (1-w)\mathbf{x}_t - \alpha \nabla f(\mathbf{x}_t)/s \quad (16)$$

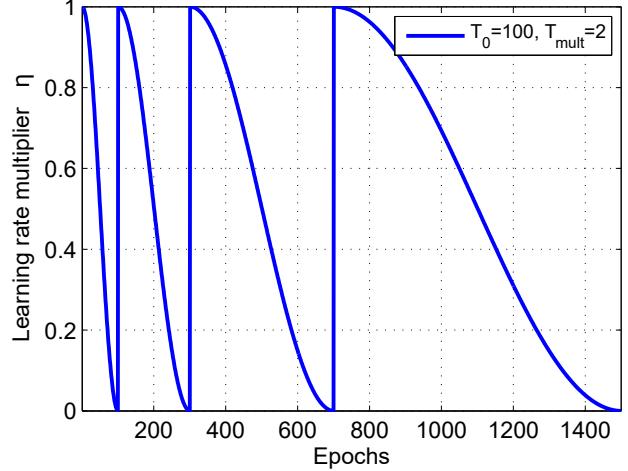
$$= \mathbf{x}_t - \alpha \nabla f(\mathbf{x}_t)/s - w\mathbf{x}_t, \quad (17)$$

These iterates are identical since $w' = \frac{w}{\alpha}$. \square

The iterates of the algorithm are as in Equation 12. A fixed point \mathbf{x}_* is attained when $\mathbf{x}_* = -\frac{\alpha}{w} \mathbf{M}_t \nabla f(\mathbf{x}_*)$. Since $\mathbf{M}_t = \text{diag}(s_1, \dots, s_N)^{-1}$, this simplifies to $\mathbf{x}_* = -\frac{\alpha}{w} \nabla f(\mathbf{x}_*)/s$, where the division is element-wise. Since $\nabla f_t^{\text{reg}}(\mathbf{x}) = \nabla f_t(\mathbf{x}) + \frac{w}{\alpha} \mathbf{x} \odot \mathbf{s}$ evaluates to zero at $\mathbf{x}_* = -\frac{\alpha}{w} \nabla f(\mathbf{x}_*)/s$, this is a minimizer of the function. \square

B. An example setting of the schedule multiplier

An example schedule of the schedule multiplier η_t is given in SuppFigure 1 for $T_{i=0} = 100$ and $T_{\text{mult}} = 2$. After the



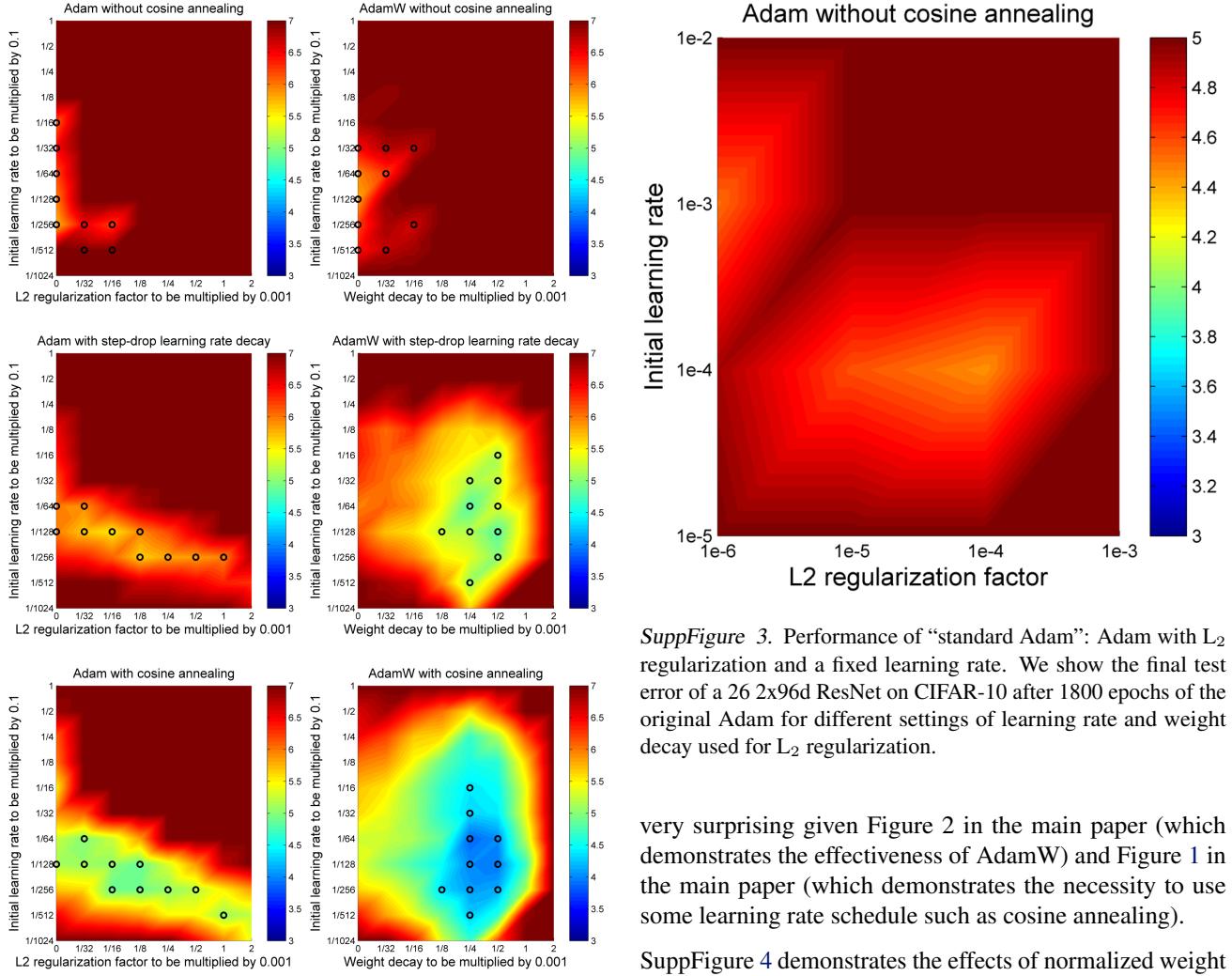
SuppFigure 1. An example schedule of the learning rate multiplier as a function of epoch index. The first run is scheduled to converge at epoch $T_{i=0} = 100$, then the budget for the next run is doubled as $T_{i=1} = T_{i=0}T_{\text{mult}} = 200$, etc.

initial 100 epochs the learning rate will reach 0 because $\eta_{t=100} = 0$. Then, since $T_{\text{cur}} = T_{i=0}$, we restart by resetting $T_{\text{cur}} = 0$, causing the multiplier η_t to be reset to 1 due to Eq. (8). This multiplier will then decrease again from 1 to 0, but now over the course of 200 epochs because $T_{i=1} = T_{i=0}T_{\text{mult}} = 200$. Solutions obtained right before the restarts, when $\eta_t = 0$ (e.g., at epoch indexes 100, 300, 700 and 1500 as shown in SuppFigure 1) are recommended by the optimizer as the solutions, with more recent solutions prioritized.

C. Additional results

In SuppFigure 2, we assess the performance of different learning rate schedules for both standard Adam with L₂ regularization and our variant of Adam with the original formulation of weight decay (AdamW). The results show that (a) cosine annealing outperforms the step-drop learning rate schedule, which in turn outperforms fixed learning rates, and (b) AdamW outperforms Adam, with larger differences for better learning rate schedules.

Fixing Weight Decay Regularization in Adam



SuppFigure 2. We show the final test error of a 26 2x64d ResNet on CIFAR-10 after 100 epochs of Adam (left column) and AdamW (right column) with fixed learning rate (i.e., without cosine annealing, see first row), step-drop learning rate (with drops at epoch indexes 30, 60 and 80, see second row) and cosine annealing (see third row). AdamW leads to a more separable hyperparameter search space, especially when a learning rate schedule such as step-drop and cosine annealing is applied. Cosine annealing yields clearly superior results.

We also investigated whether the use of much longer runs (1800 epochs) of “standard Adam” (Adam with L₂ regularization and a fixed learning rate) makes the use of cosine annealing unnecessary. SuppFigure 3 shows the results of standard Adam for a 4 by 4 logarithmic grid of hyperparameter settings (the coarseness of the grid is due to the high computational expense of runs for 1800 epochs). Even after taking the low resolution of the grid into account, the results appear to be at best comparable to the ones obtained with AdamW with 18 times less epochs and a smaller network (see SuppFigure 4, top row, middle). These results are not

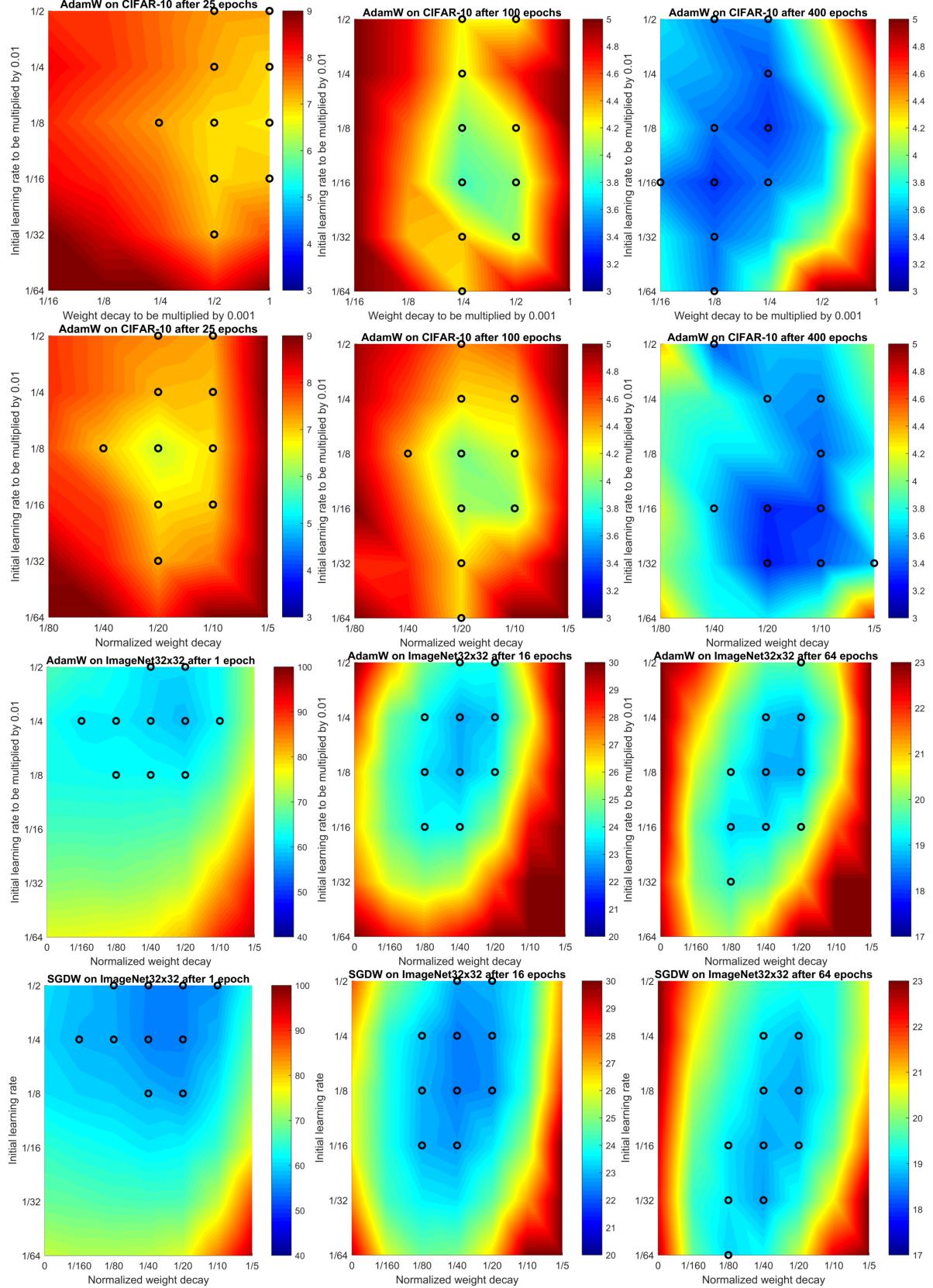
SuppFigure 3. Performance of “standard Adam”: Adam with L₂ regularization and a fixed learning rate. We show the final test error of a 26 2x96d ResNet on CIFAR-10 after 1800 epochs of the original Adam for different settings of learning rate and weight decay used for L₂ regularization.

very surprising given Figure 2 in the main paper (which demonstrates the effectiveness of AdamW) and Figure 1 in the main paper (which demonstrates the necessity to use some learning rate schedule such as cosine annealing).

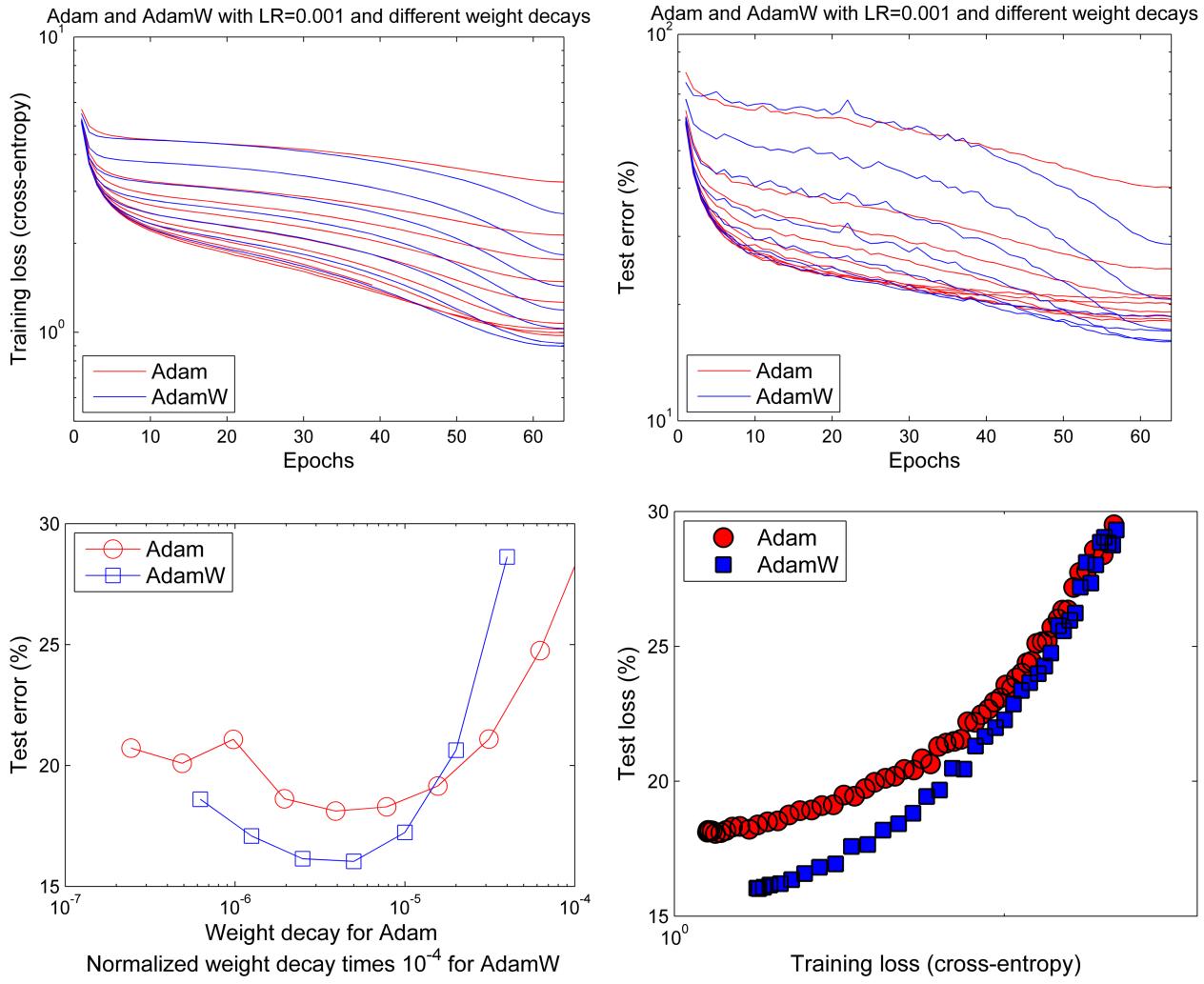
SuppFigure 4 demonstrates the effects of normalized weight decay: the optimal setting of weight decay differs substantially for different runtime budgets (see the first row), the values for normalized weight decay remain very similar (see the second row). They also remain similar across datasets and even across AdamW and SGDW (see the third and fourth rows).

Finally, SuppFigure 5 is the equivalent of Figure 2 in the main paper, but for ImageNet32x32 instead of for CIFAR-10. The qualitative results are identical: weight decay leads to better training loss (cross-entropy) than L₂ regularization, and to an even greater improvements of test error.

Fixing Weight Decay Regularization in Adam



SuppFigure 4. Effect of normalized weight decay. We show the final test Top-1 error on CIFAR-10 (first two rows for AdamW without and with normalized weight decay) and Top-5 error on ImageNet32x32 (last two rows for AdamW and SGDW, both with normalized weight decay) of a 26 2x64d ResNet after different numbers of epochs (see columns). While the optimal settings of the raw weight decay change significantly for different runtime budgets (see the first row), the values of the normalized weight decay remain very similar for different budgets (see the second row) and different datasets (here, CIFAR-10 and ImageNet32x32), and even across AdamW and SGDW.



SuppFigure 5. Learning curves (top row) and generalization results (Top-5 errors in bottom row) obtained by a 26 2x96d ResNet trained with Adam and AdamW on ImageNet32x32.