Supplementary Materials for Averaging Weights Leads to Wider Optima and Better Generalization

Pavel Izmailov[*1] **Dmitrii Podoprikhin**[*2,3] **Timur Garipov**[*4,5] **Dmitry Vetrov**[2,3] **Andrew Gordon Wilson**[1]
[1]Cornell University, [2]Higher School of Economics, [3]Samsung-HSE Laboratory,
[4]Samsung AI Center in Moscow, [5]Lomonosov Moscow State University

A Appendix

A.1 EXPERIMENTAL DETAILS

For the experiments on CIFAR datasets (section **??**) we used the following implementations (embedded links):

- Shake-Shake-2x64d
- PyramidNet-272
- VGG-16
- Preactivation-ResNet-164
- Wide ResNet-28-10

Models for ImageNet are from here. Pretrained networks can be found here.

SWA learning rates. For PyramidNet SWA uses a cyclic learning rate with $\alpha_1 = 0.05$ and $\alpha_2 = 0.001$ and cycle length 3. For VGG and Wide ResNet we used constant learning $\alpha_1 = 0.01$. For ResNet we used constant learning rates $\alpha_1 = 0.01$ on CIFAR-10 and 0.05 on CIFAR-100.

For Shake-Shake Net we used a custom cyclic learning rate based on the cosine annealing used when training Shake-Shake with SGD. Each of the cycles replicate the learning rates corresponding to epochs $1600 - 1700$ of the standard training and the cycle length $c = 100$ epochs. The learning rate schedule is depicted in Figure 4 and follows the formula

$$\alpha(i) = 0.1 \cdot \left(1 + \cos\left(\pi \cdot \frac{1600 + \text{epoch}(i) \bmod 100)}{1800}\right)\right),$$

where epoch(i) is the number of data passes completed before iteration i.

[*]Equal contribution.

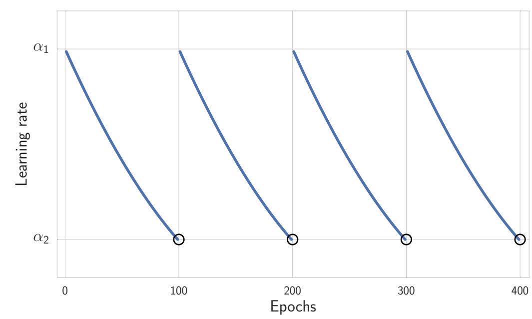

Figure 8: Cyclical learning rate used for Shake-Shake as a function of iteration.

For all experiments with ImageNet we used cyclic learning rate schedule with the same hyperparameters $\alpha_1 = 0.001$, $\alpha_2 = 10^{-5}$ and $c = 1$.

SGD learning rates. For conventional SGD training we used SGD with momentum 0.9 and with an annealed learning rate schedule. For VGG, Wide ResNet and Preactivation ResNet we fixed the learning rate to α_1 for the first half of epochs ($0B$–$0.5B$), then linearly decreased the learning rate to $0.01\alpha_1$ for the next 40% of epochs ($0.5B$–$0.9B$), and then kept it constant for the last 10% of epochs ($0.9B$ – $1B$). For VGG we set $\alpha_1 = 0.05$, and for Preactivation ResNet and Wide ResNet we set $\alpha_1 = 0.1$. For Shake-Shake Net and PyramidNets we used the cosine and piecewise-constant learning rate schedules described in Gastaldi [2017] and Han et al. [2016] respectively.

A.2 TRAINING RESNET WITH A CONSTANT LEARNING RATE

In this section we present the experiment on training Preactivation ResNet-164 using a constant learning rate. The experimental setup is the same as in section **??**. We set the learning rate to $\alpha_1 = 0.1$ and start averaging after epoch 200. The results are presented in Figure 5.

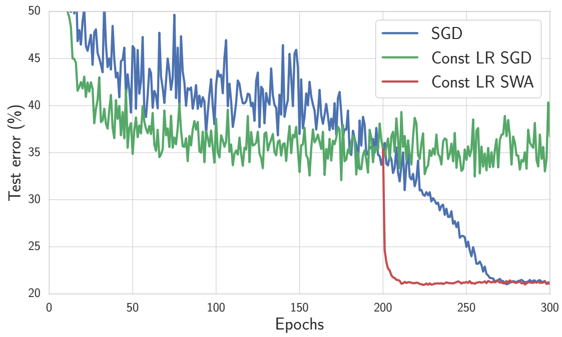

Figure 9: Test error as a function of training epoch for constant (green) and decaying (blue) learning rate schedules for a Preactivation ResNet-164 on CIFAR-100. In red we average the points along the trajectory of SGD with constant learning rate starting at epoch 200.

References

Xavier Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.

Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. *arXiv preprint arXiv:1610.02915*, 2016.