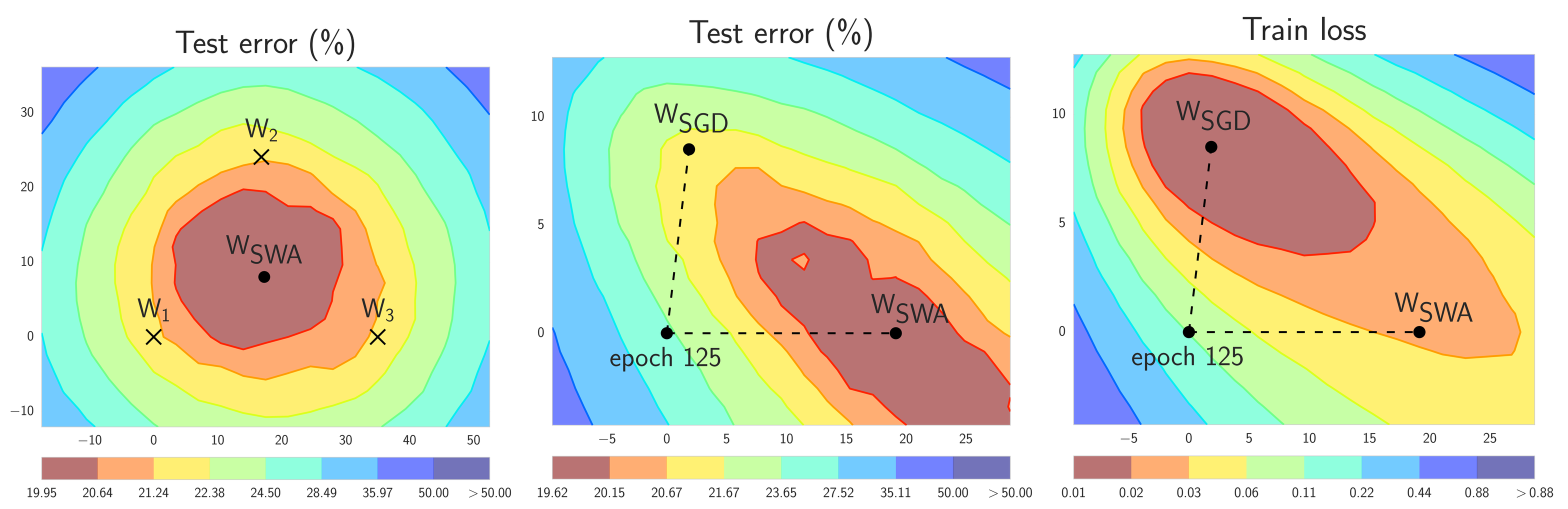
Averaging Weights Leads to Wider Optima and Better Generalization

Pavel Izmailov *[1], Dmitrii Podoprikhin*[2,3], Timur Garipov*[4,5], Dmitry Vetrov[2,3], and Andrew Gordon Wilson[1]

[1]**Cornell University,** [2]**Higher School of Economics,** [3]**Samsung-HSE Laboratory,** [4]**Samsung AI Center in Moscow, and** [5]**Lomonosov Moscow State University**

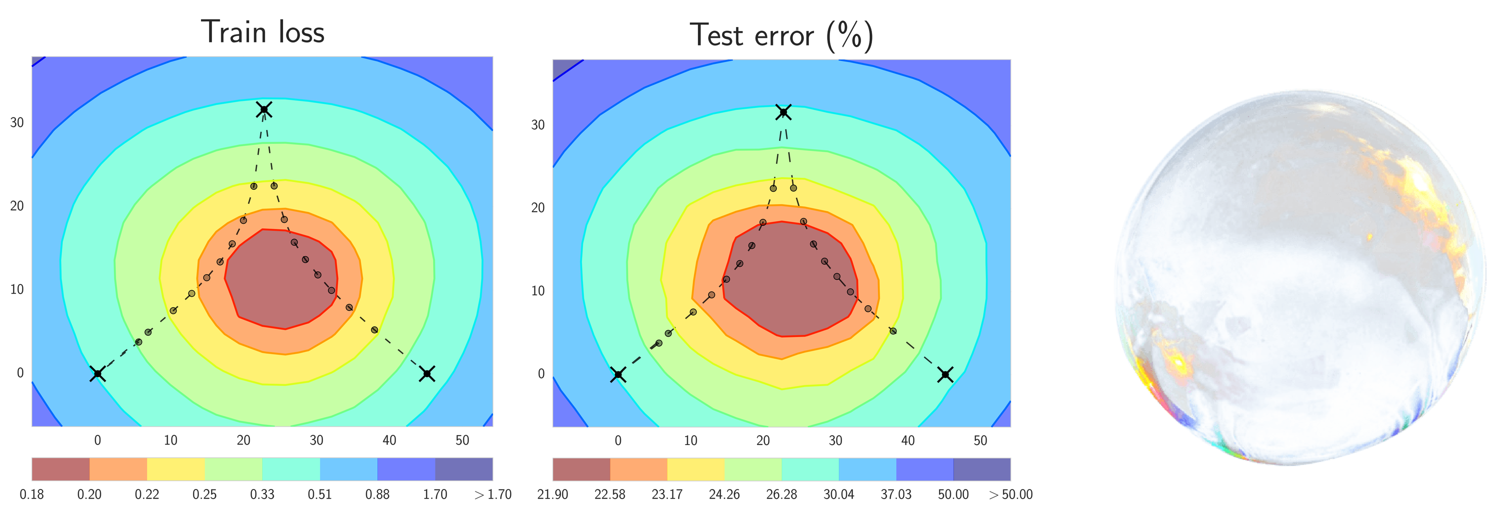
Outline

- SGD with cyclical and constant learning rates traverses regions of weight space corresponding to high-performing networks. While these models are moving around this optimal set they never reach its central points
- We can move into this more desirable space of points by averaging the weights proposed over SGD iterations
- We propose Stochastic Weight Averaging (SWA) – an **equally weighted** running average of parameters (DNN weights) traversed by SGD with a modified (cyclical or high constant) learning rate schedule
- SWA leads to solutions corresponding to wider optima than SGD and achieves notable generalization improvement for a broad range of architectures over several consequential benchmarks with virtually no computational overhead



Motivation

Let's continue to run SGD with a constant learning rate from a pre-trained solution and visualize the trajectory.



SGD oscilates around the region of high-performing solutions and averaging SGD iterates improves test performance.

Explanations:

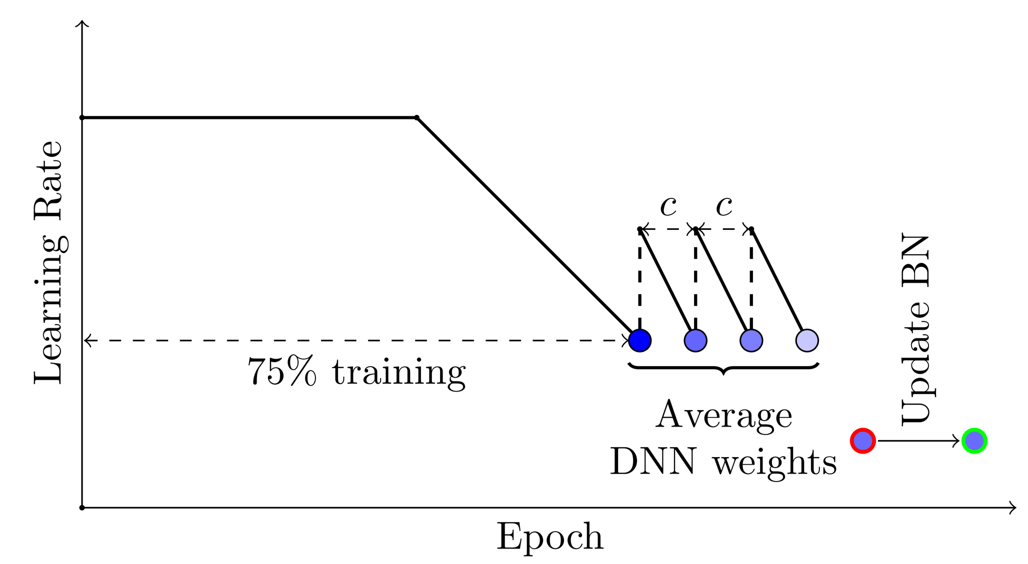
- Soap Bubble: constant learning rate SGD is sampling from a high-dimensional Gaussian, which has most of its mass concentrated in a thin shell
- Averaging weights approximates ensembling predictions by linearization if the weights being averaged are close

$$f\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) \approx \frac{1}{n}\sum_{i=1}^{n} f(w_i)$$

Stochastic Weight Averaging

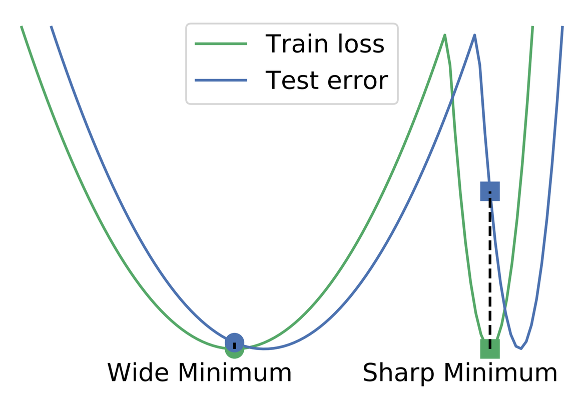
Details of SWA:

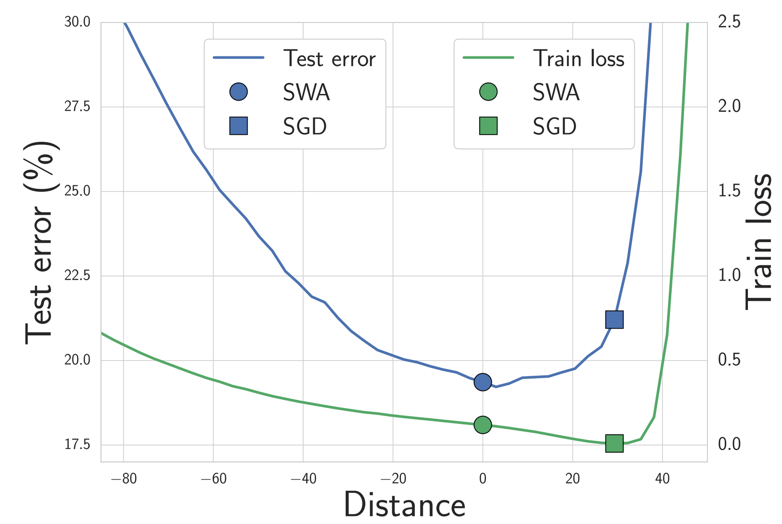
- Use learning rate schedule that doesn't decay to zero, e.g. cyclical or high constant at the end of training
- Average weights at the end of each of the last K epochs or at the end of each cycle
- Recompute Batch Normalization statistics at the end of training



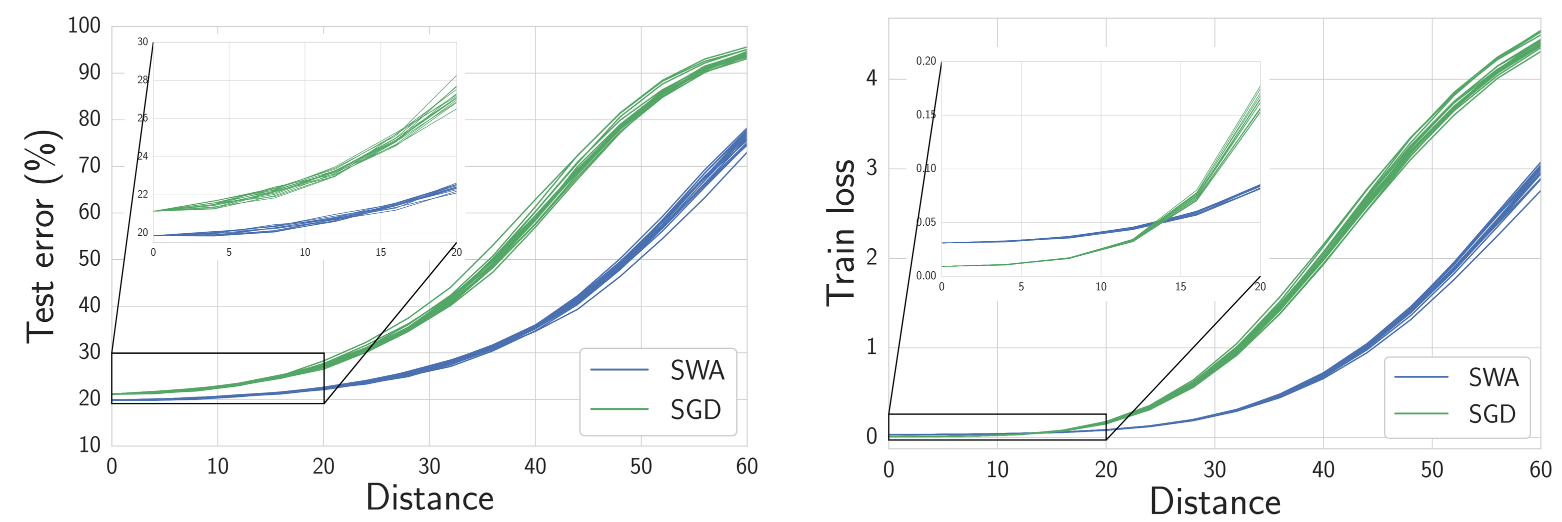
Optima Width

Optima width is conjectured to be highly correlated with generalization.





SWA leads to wider optima centered in the region of high-performing networks.



Results

DNN (Budget)	SGD	SWA 1 Budget	SWA 1.5 Budget
CIFAR-100			
VGG-16 (200)	72.55 ± 0.10	73.91 ± 0.12	$\mathbf{74.27 \pm 0.25}$
ResNet-164 (150)	78.49 ± 0.36	79.77 ± 0.17	$\mathbf{80.35 \pm 0.16}$
WRN-28-10 (200)	80.82 ± 0.23	81.46 ± 0.23	$\mathbf{82.15 \pm 0.27}$
PyramidNet-272 (300)	83.41 ± 0.21	–	$\mathbf{84.16 \pm 0.15}$
CIFAR-10			
VGG-16 (200)	93.25 ± 0.16	93.59 ± 0.16	$\mathbf{93.64 \pm 0.18}$
ResNet-164 (150)	95.28 ± 0.10	95.56 ± 0.11	$\mathbf{95.83 \pm 0.03}$
WRN-28-10 (200)	96.18 ± 0.11	96.45 ± 0.11	$\mathbf{96.79 \pm 0.05}$
ShakeShake-2x64d (1800)	96.93 ± 0.10	–	$\mathbf{97.12 \pm 0.06}$
Imagenet			
DNN	SGD	SWA 5 epochs	SWA 10 epochs
ResNet-50	76.15	76.83 ± 0.01	$\mathbf{76.97 \pm 0.05}$
ResNet-152	78.31	78.82 ± 0.01	$\mathbf{78.94 \pm 0.07}$
DenseNet-161	77.65	78.26 ± 0.09	$\mathbf{78.44 \pm 0.06}$

Code

Code available at `https://github.com/timgaripov/swa`