Supplementary Materials for Empirically explaining SGD from a line search perspective

Maximus Mutschler and Andreas Zell

University of Tübingen, Sand 1, D-72076 Tübingen, Germany
{maximus.mutschler, andreas.zell}@uni-tuebingen.de

Abstract. In the following the resulting plots of analyzing the full-batch loss along lines in update step direction on a ResNet-18 and Mobilenet-V2 trained on 8% of CIFAR-10 with SGD with and without momentum are given.
Note that ResNet18 has a significantly different block structure compared to ResNet-20. ResNet-18 consists of 4 times 2 basic blocks with average pooling in-between, whereas, ResNet-20 consists of 3 times 3 blocks with average pooling in-between.

Figures start at the next page for easier comparison.

1 Distance Matrices

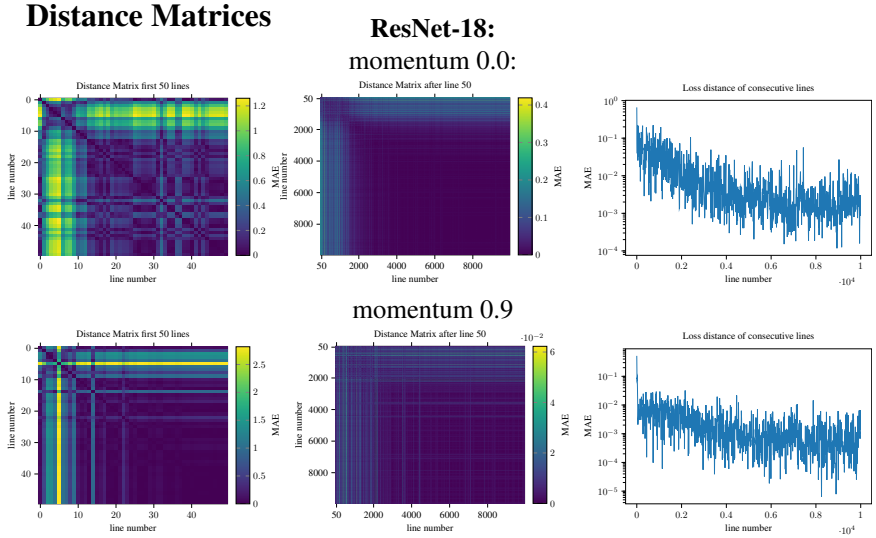

Fig. 1: **ResNet-18:** Distances of the shape of full-batch losses along lines in a window around the current position $s = 0$. **Row 1:** SGD without momentum. **Row 2:** SGD with momentum. Since the offset is not of interest the minimum is shifted to 0 on the y-axis. The distances are rather high for the first 20 lines (left). For the following lines the distances are less than 0.4 MAE (middle) and concentrate around 0.005. The MAEs of the full-batch loss of pairs of consecutive lines are given on the right.

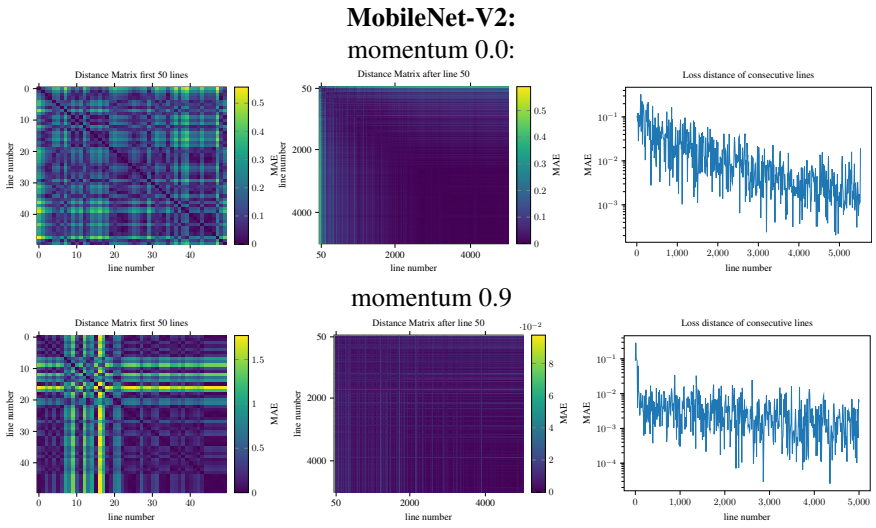

Fig. 2: **MobileNet-V2:** see Figure 1 for explanations. The distances are rather high for the first 25 lines (left). For the following lines the distances are less then 0.6 MAE (middle) and concentrate around 0.01.

2 Parabolic Approximation

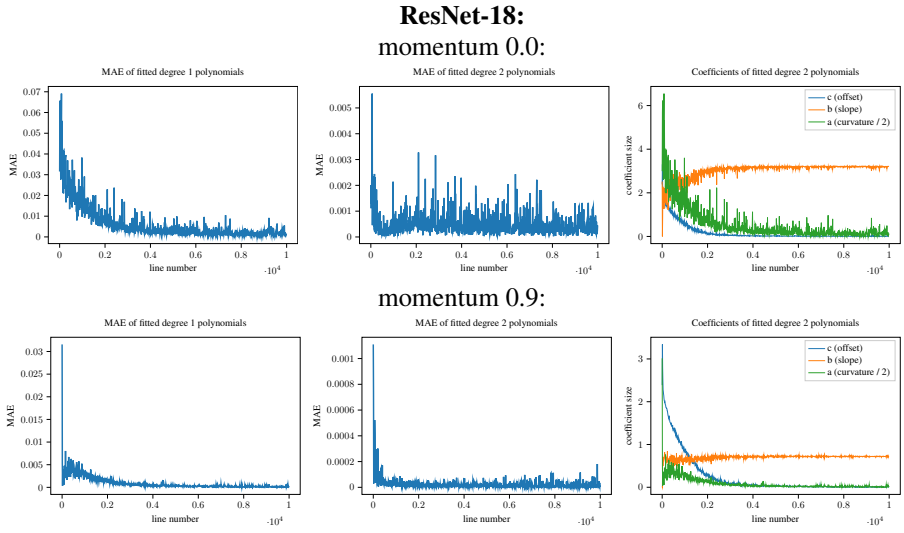

Fig. 3: **ResNet-18:** MAE of polynomial approximations of the full-batch loss of degree 1 and 2. **Row 1:** SGD without momentum. **Row 2:** SGD with momentum. Full-batch losses along lines can be well fitted by polynomials of degree 2. The slope of the approximation stays roughly constant whereas the curvature decreases.

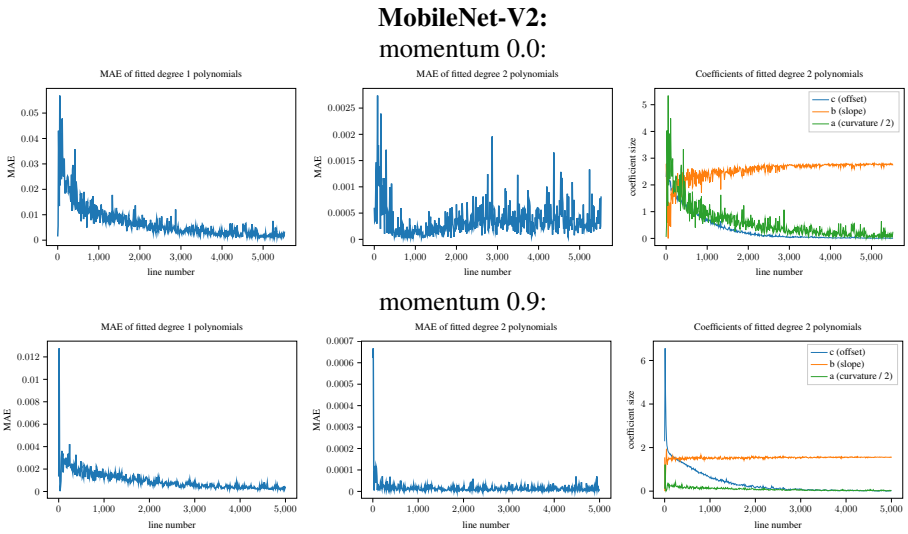

Fig. 4: **MobileNet-V2:** for explanations and interpretations see Figure 3

3 Optimization strategy metrics

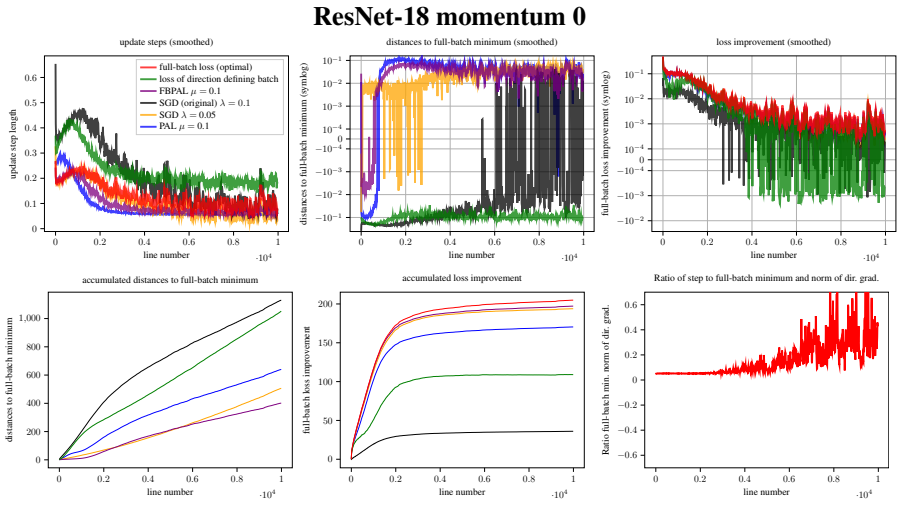

Fig. 5: **ResNet-18 momentum 0:** Several metrics to compare update step strategies: 1. update step sizes. 2. the distance to the minimum of the full batch loss ($s_{opt} - s_{upd}$), which is the optimal update step from a local perspective. 3. the loss improvement per step given as: $l(0) - l(s_{upd})$ where s_{upd} is the update step of a strategy. Average smoothing with a kernel size of 25 is applied. The right lower plot shows almost proportional behavior between s_{opt} and the directional derivative of the direction defining mini-batch loss in the beginning. This is the only case, out of 6 analyzed, where the mean of the ratio changes significantly.

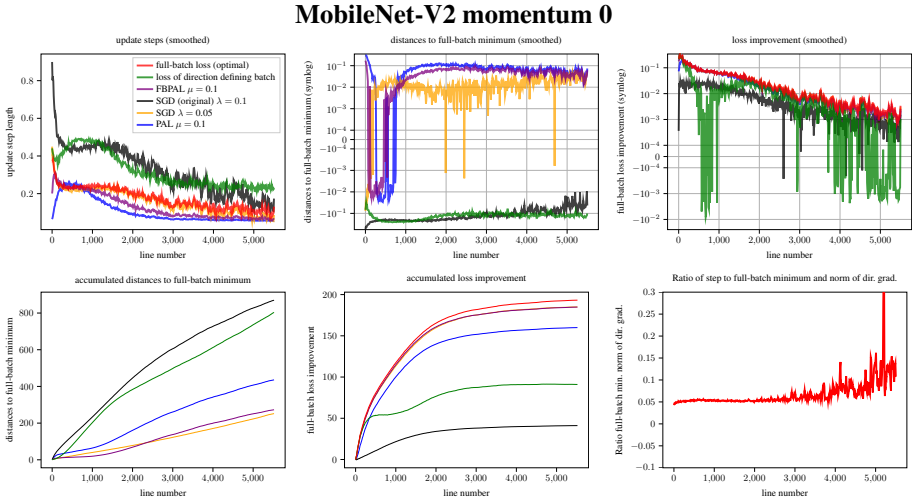

Fig. 6: **MobileNet-V2 momentum 0**. See Figure 5 for explanations.

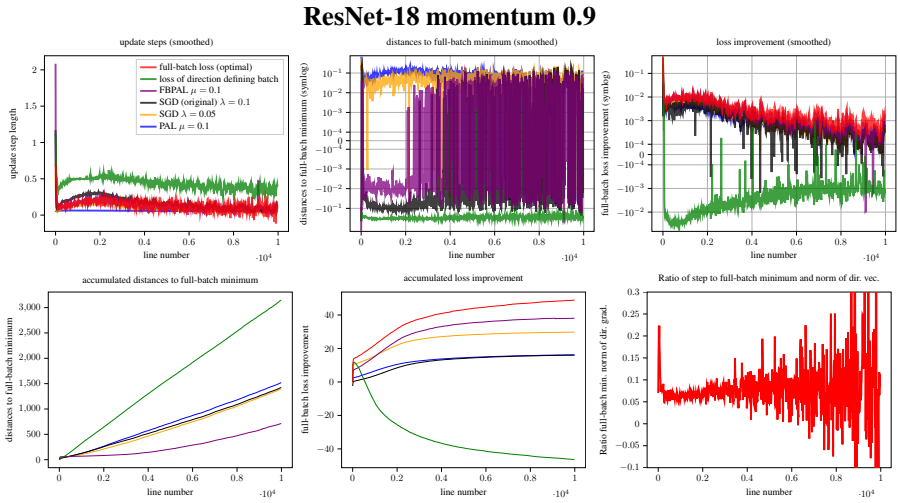

Fig. 7: **ResNet-18 momentum 0.9**. See Figure 5 for explanations. In the case of momentum SGD is not able to perform such an exact line search as in the case without monemtum since the norm of the momentum vector is not directly related to the loss of the current line considered.

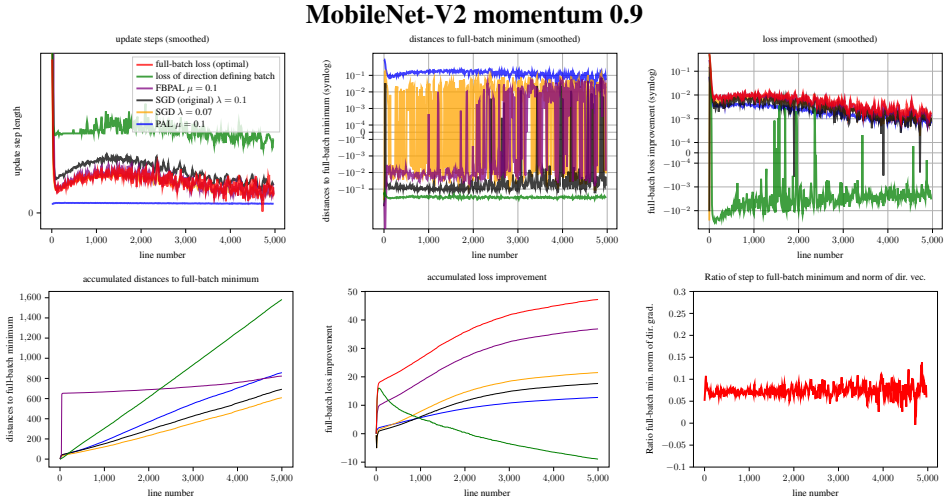

Fig. 8: **MobileNet-V2 momentum 0.9**. See Figure 5 and 7 for explanations and interpretations.

4 Batch Size comparision

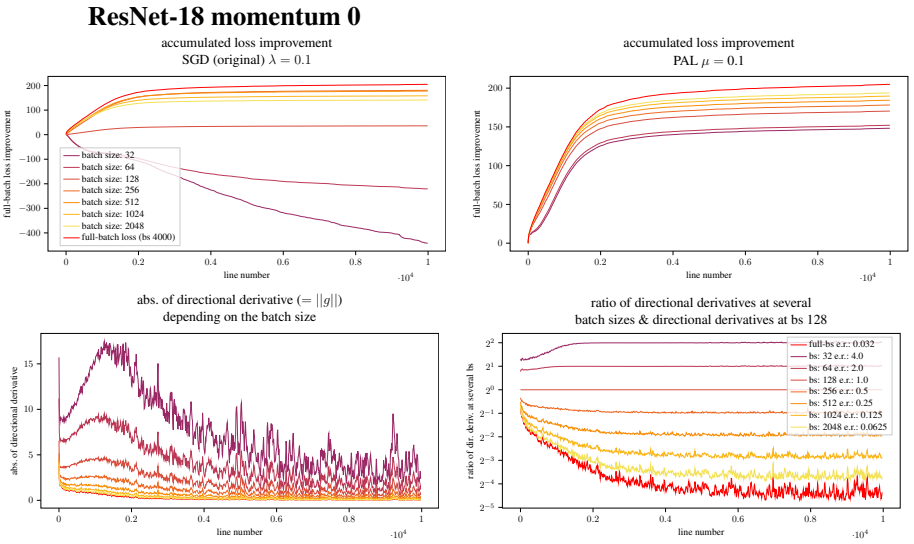

Fig. 9: **Row 1** Comparing the influence of the batch size on the loss improvement. Left: SGD with the original learning rate of 0.1. Right: parabolic approximation line search (PAL). **Row 2:** Analysis of the relation of the batch size to the absolute directional derivative (=gradient norm) which shows in detail that increasing the batch size has a similar effect as decreasing the learning rate by the same factor. **e.r.** stands for expected ratio.

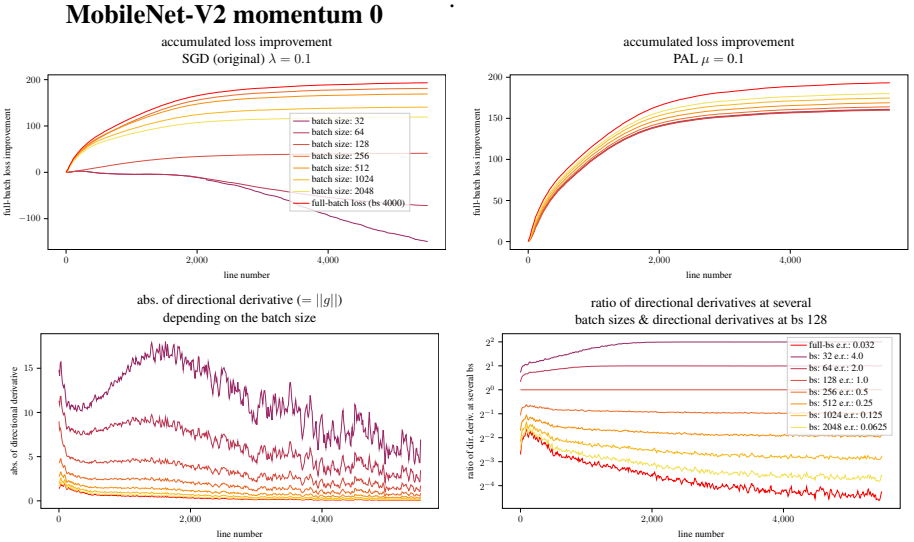

Fig. 10: **MobileNet-V2 momentum 0:** See Figure 9 for explanations.