

Appendix

Table of Contents

A Additional Results on Training Convolutional Networks	12
A.1 Training on CIFAR-10	12
A.2 Training on MNIST	12
A.3 Training on SVHN	13
B Additional Results on Training Residual Networks	15
B.1 Training on MNIST	15
B.2 Training on SVHN	16
C Additional Training Results with Small Learning Rates	17
D Additional Results on Sign Shift of Gradient Entries	18
E Additional Experiments on the Impact of Batch Size	19

A ADDITIONAL RESULTS ON TRAINING CONVOLUTIONAL NETWORKS

A.1 TRAINING ON CIFAR-10

The following figure corresponds to the training of an AlexNet on CIFAR-10 using full-batch gradient descent with the learning rate $\eta = 0.1$ for 1000 epochs. One can observe the zigzag geometry at epoch 595. Also, from the mean gradient correlation figure, the optimization geometry has a sharp transition from smooth geometry to zigzag geometry.

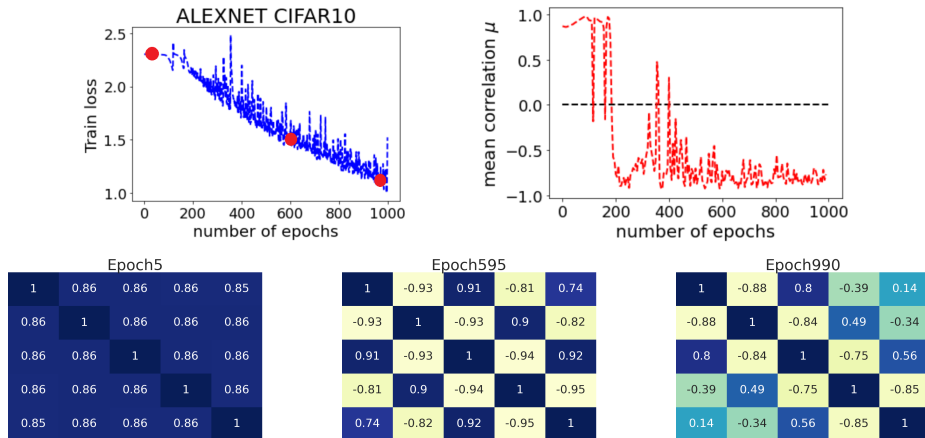


Figure 11: AlexNet training using gradient descent on CIFAR-10

A.2 TRAINING ON MNIST

The following figure corresponds to the training of a CNN on MNIST using full-batch gradient descent with learning rate $\eta = 0.1$ for 2000 epochs.

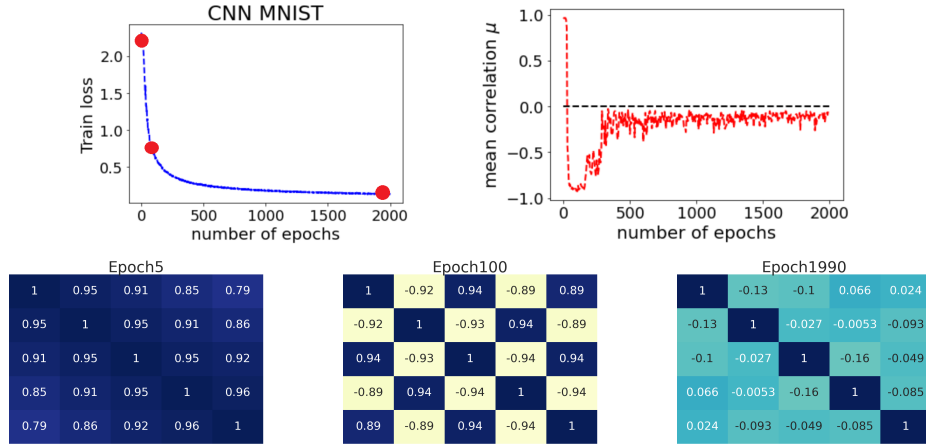


Figure 12: CNN training using gradient descent on MNIST

The following figure corresponds to the training of an AlexNet on MNIST using full-batch gradient descent with learning rate $\eta = 0.1$ for 1000 epochs.

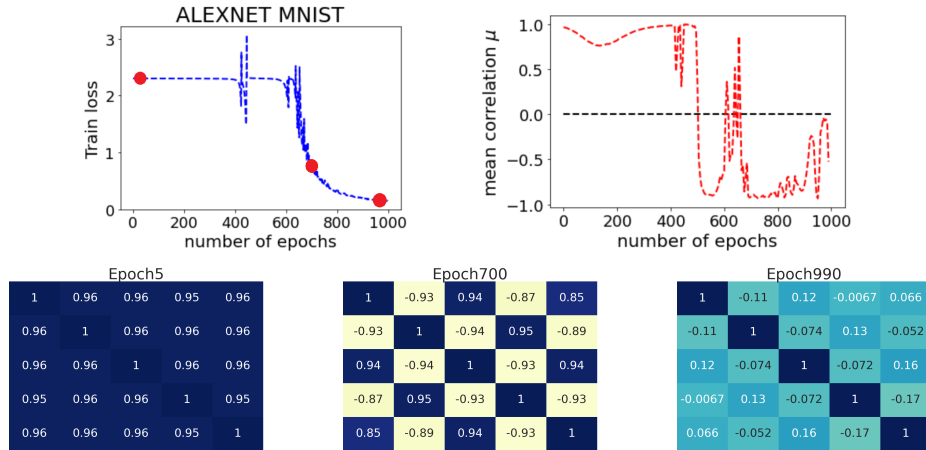


Figure 13: AlexNet training using gradient descent on MNIST

A.3 TRAINING ON SVHN

The following figure corresponds to the training of a CNN on SVHN using full-batch gradient descent with learning rate $\eta = 0.1$ for 1500 epochs.

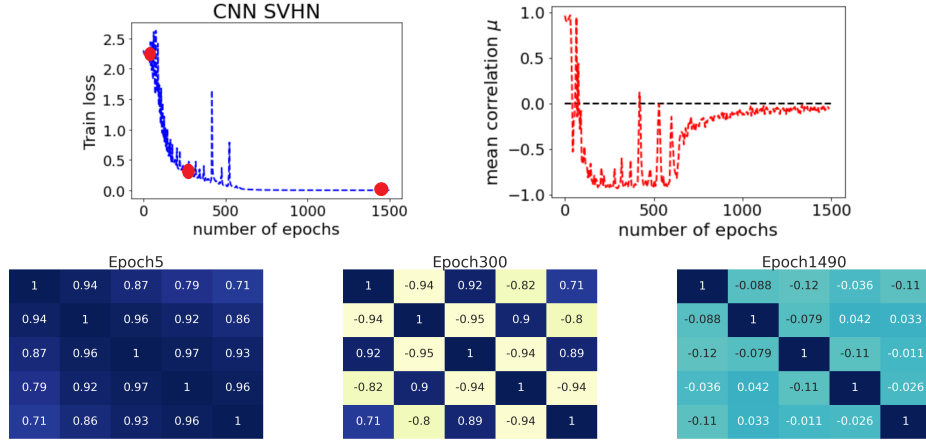


Figure 14: CNN training using gradient descent on SVHN

The following figure corresponds to the training of VGG-16 on SVHN using full-batch gradient descent with learning rate $\eta = 0.1$ for 1000 epochs.

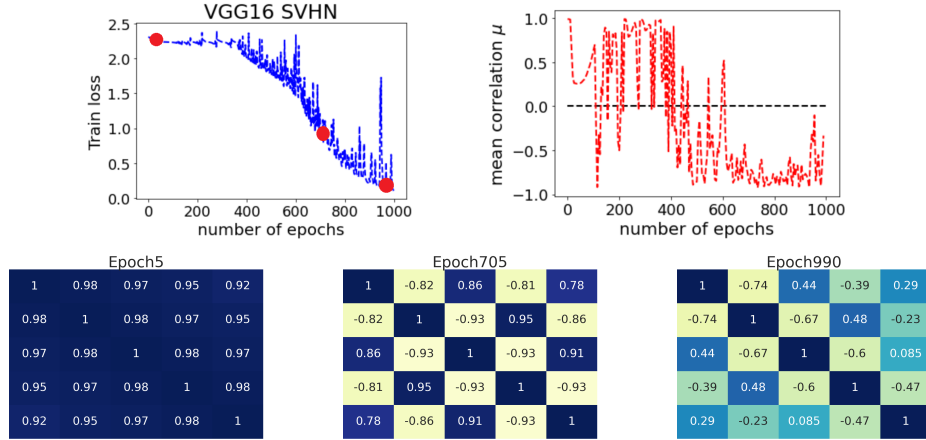


Figure 15: VGG-16 training using gradient descent on SVHN

The following figure corresponds to the training of an AlexNet on SVHN using full-batch gradient descent with learning rate $\eta = 0.1$ for 1000 epochs.

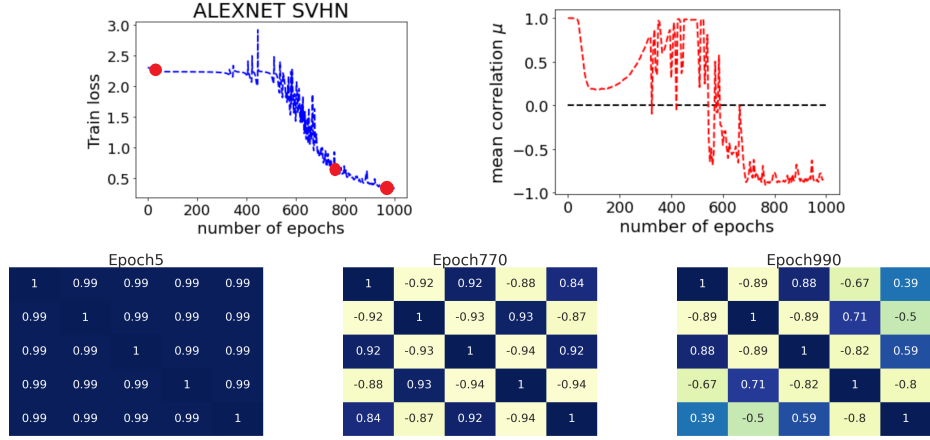


Figure 16: AlexNet training using gradient descent on SVHN

B ADDITIONAL RESULTS ON TRAINING RESIDUAL NETWORKS

B.1 TRAINING ON MNIST

The following figure corresponds to the training of a ResNet-18 on MNIST using full-batch gradient descent with the learning rate $\eta = 0.1$ for 500 epochs.

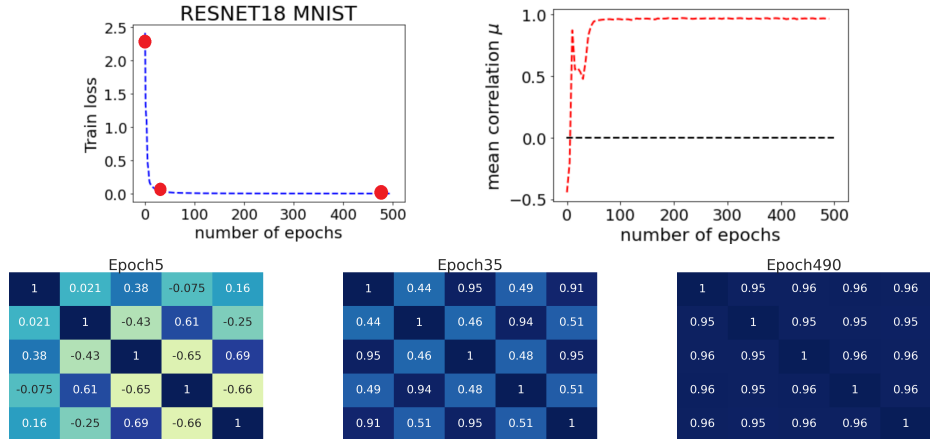


Figure 17: ResNet-18 training using gradient descent on MNIST

The following figure corresponds to the training of a ResNet-34 on MNIST using full-batch gradient descent with learning rate $\eta = 0.01$ for 500 epochs.

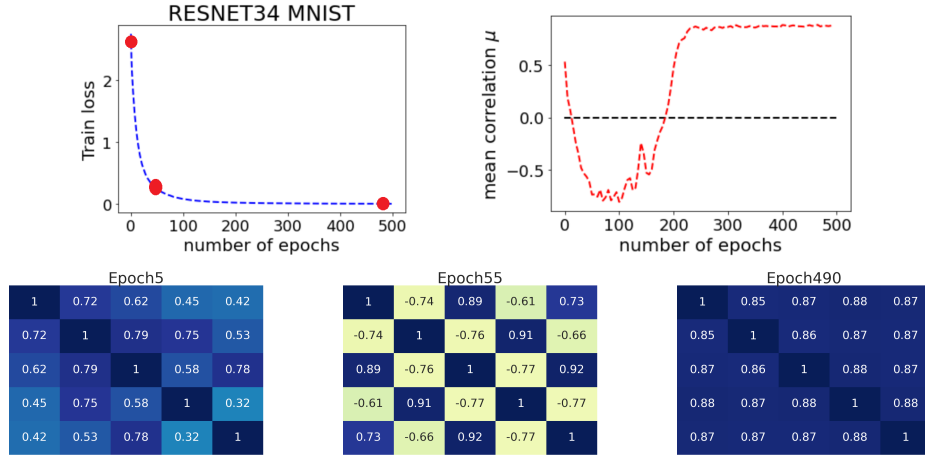


Figure 18: ResNet-34 training using gradient descent on MNIST

B.2 TRAINING ON SVHN

The following figure corresponds to the training of a ResNet-18 on SVHN using full-batch gradient descent with learning rate $\eta = 0.05$ for 500 epochs.

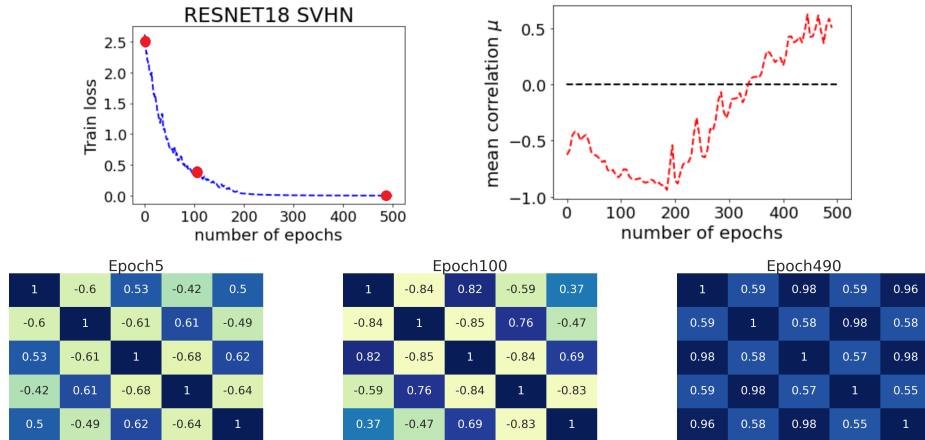


Figure 19: ResNet-18 training using gradient descent on SVHN

The following figure corresponds to the training of a ResNet-34 on SVHN using full-batch gradient descent with learning rate $\eta = 0.1$ for 500 epochs.

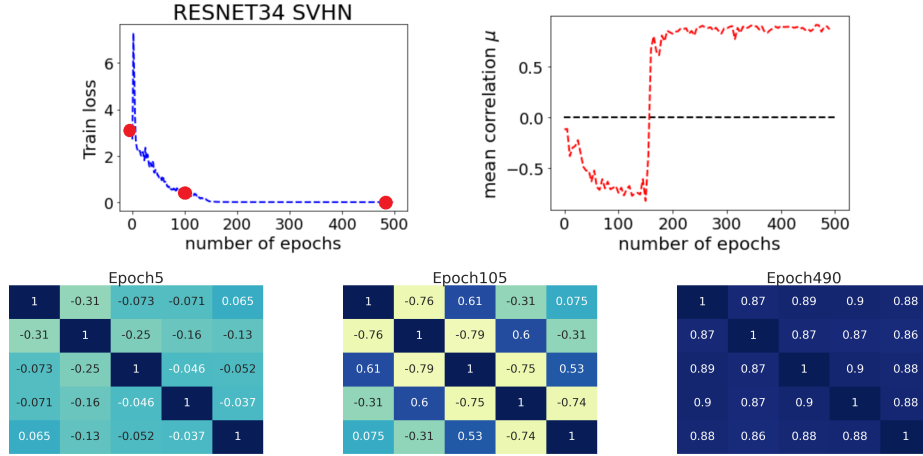
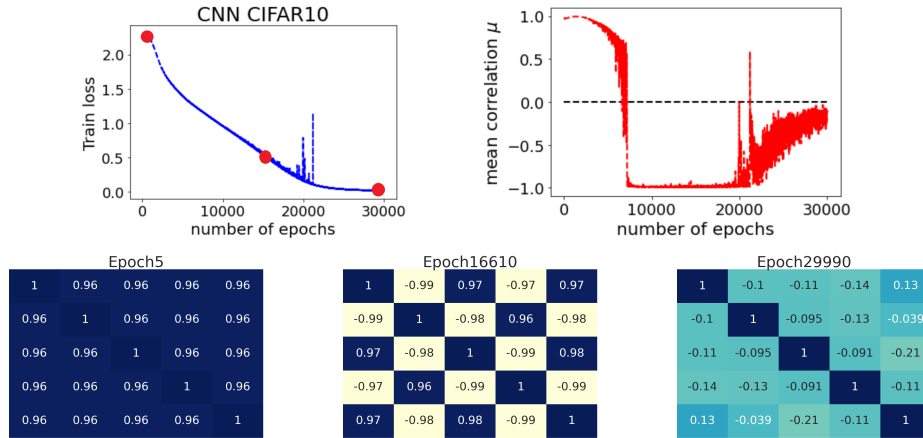


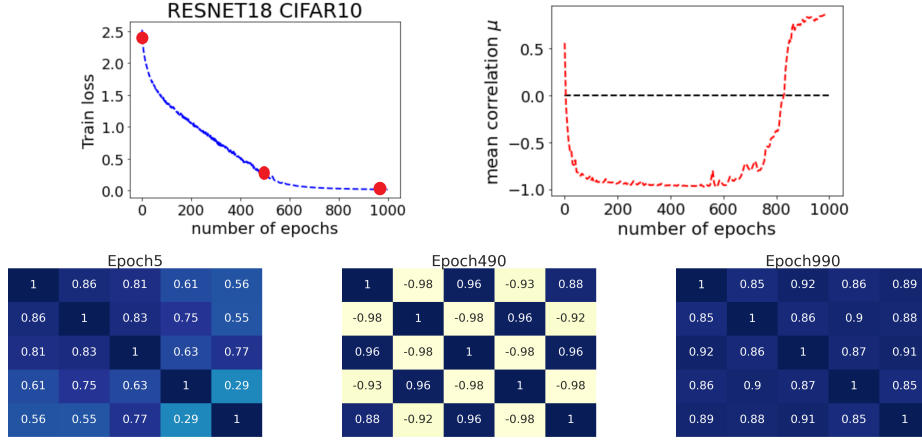
Figure 20: ResNet-34 training using gradient descent on SVHN

C ADDITIONAL TRAINING RESULTS WITH SMALL LEARNING RATES

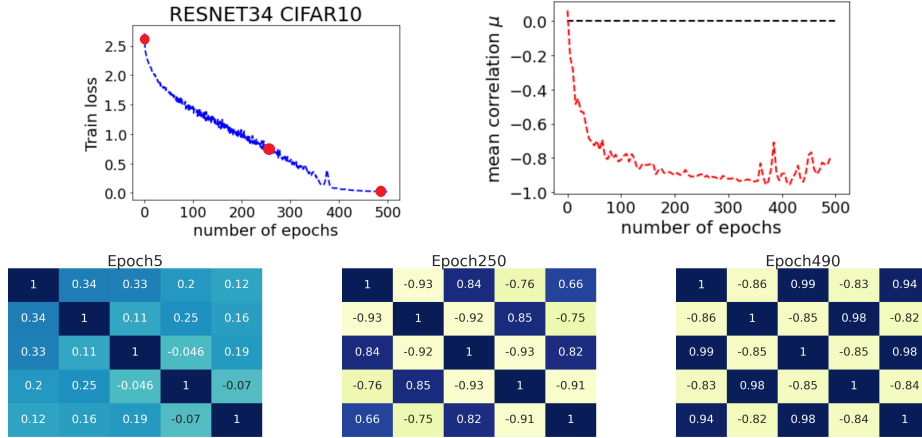
The following figure corresponds to the training of a CNN on CIFAR-10 using full-batch gradient descent with learning rate $\eta = 0.001$ for 30000 epochs.

Figure 21: CNN training using gradient descent with $\eta = 0.001$ on CIFAR-10

The following figure corresponds to the training of a ResNet-18 on CIFAR-10 using full-batch gradient descent with learning rate $\eta = 0.01$ for 1000 epochs.

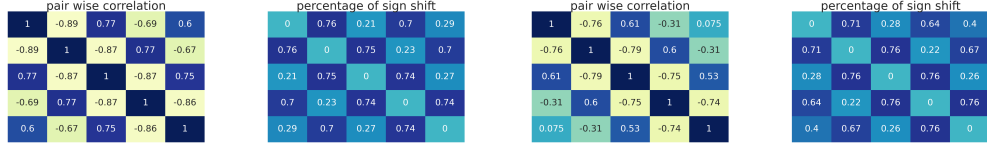
Figure 22: ResNet-18 training using gradient descent with $\eta = 0.01$ on CIFAR-10

The following figure corresponds to the training of a ResNet-34 on CIFAR-10 using full-batch gradient descent with learning rate $\eta = 0.01$ for 500 epochs.

Figure 23: ResNet-34 training using gradient descent with $\eta = 0.01$ on CIFAR-10

D ADDITIONAL RESULTS ON SIGN SHIFT OF GRADIENT ENTRIES

In Figure 24, we show the pairwise gradient correlation matrix when encountering the zigzag geometry in training residual networks. It can be seen that when encountering the zigzag geometry in ResNet-18 training, the gradient correlations of adjacent iterations stay around $-0.89 \sim -0.86$, and about 75% of the gradient entries have opposite signs. As a comparison, in ResNet-34 training, the gradient correlations of adjacent iterations stay around $-0.93 \sim -0.92$, and about 58% of the gradient entries have opposite signs. This implies that the magnitudes of gradient entries of residual networks have a relatively concentrated distribution, and such a complex geometry is across those parameter dimensions with relatively high magnitudes.



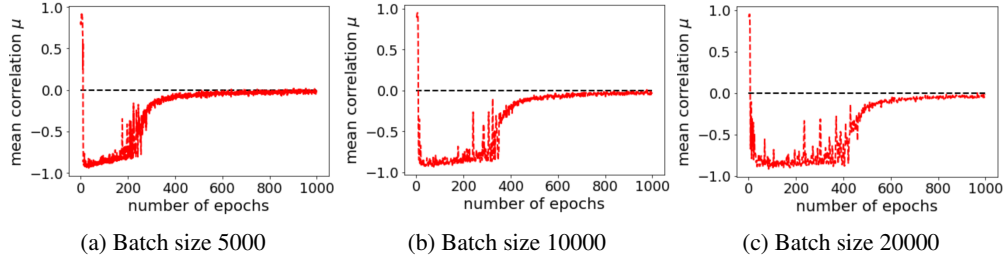
(a) Epoch 60 of ResNet-18 training on CIFAR-10

(b) Epoch 105 of ResNet-34 training on SVHN

Figure 24: Gradient sign shift in training residual networks

E ADDITIONAL EXPERIMENTS ON THE IMPACT OF BATCH SIZE

Figures 25 and 26 show the mean gradient correlation in training CNN and ResNet-34 with different large batch sizes on CIFAR-10.

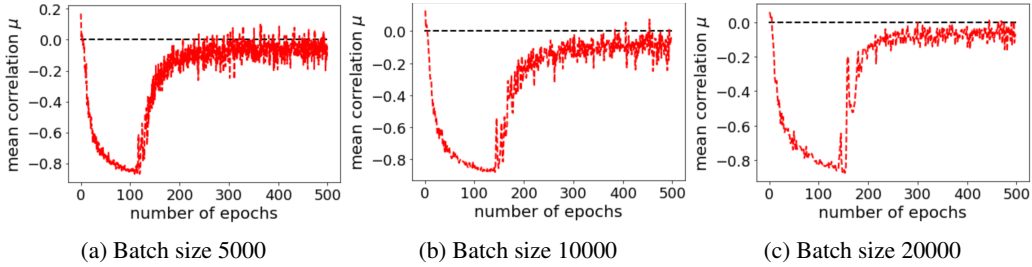


(a) Batch size 5000

(b) Batch size 10000

(c) Batch size 20000

Figure 25: Gradient correlation in CNN training on CIFAR-10 under different batch sizes



(a) Batch size 5000

(b) Batch size 10000

(c) Batch size 20000

Figure 26: Gradient correlation in ResNet-34 training on CIFAR-10 under different batch sizes