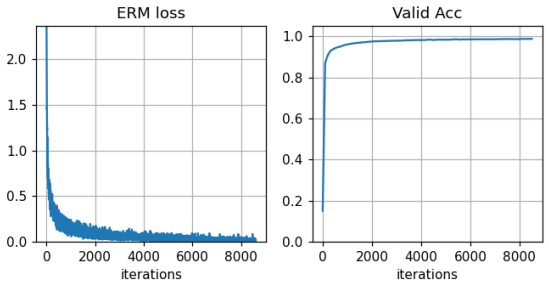
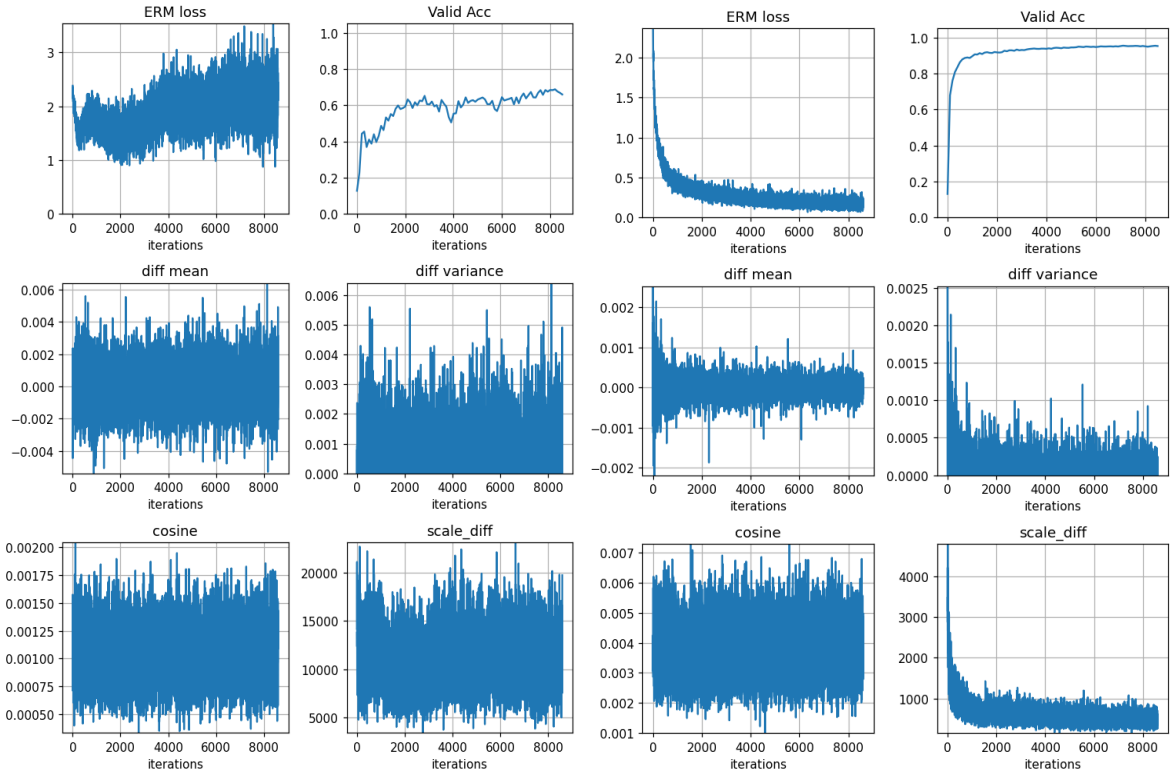
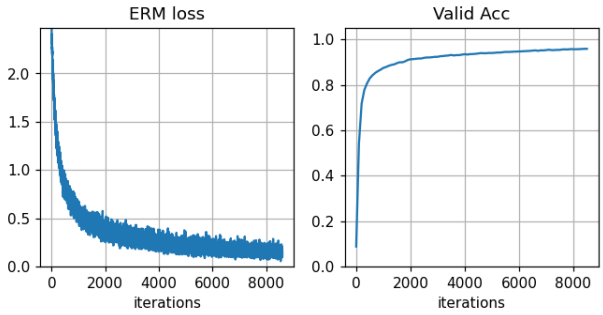
Linear Porjection with CNN ~9 million parameters, On mnist, SGD, lr = 1e-3/1e-4/1e-5

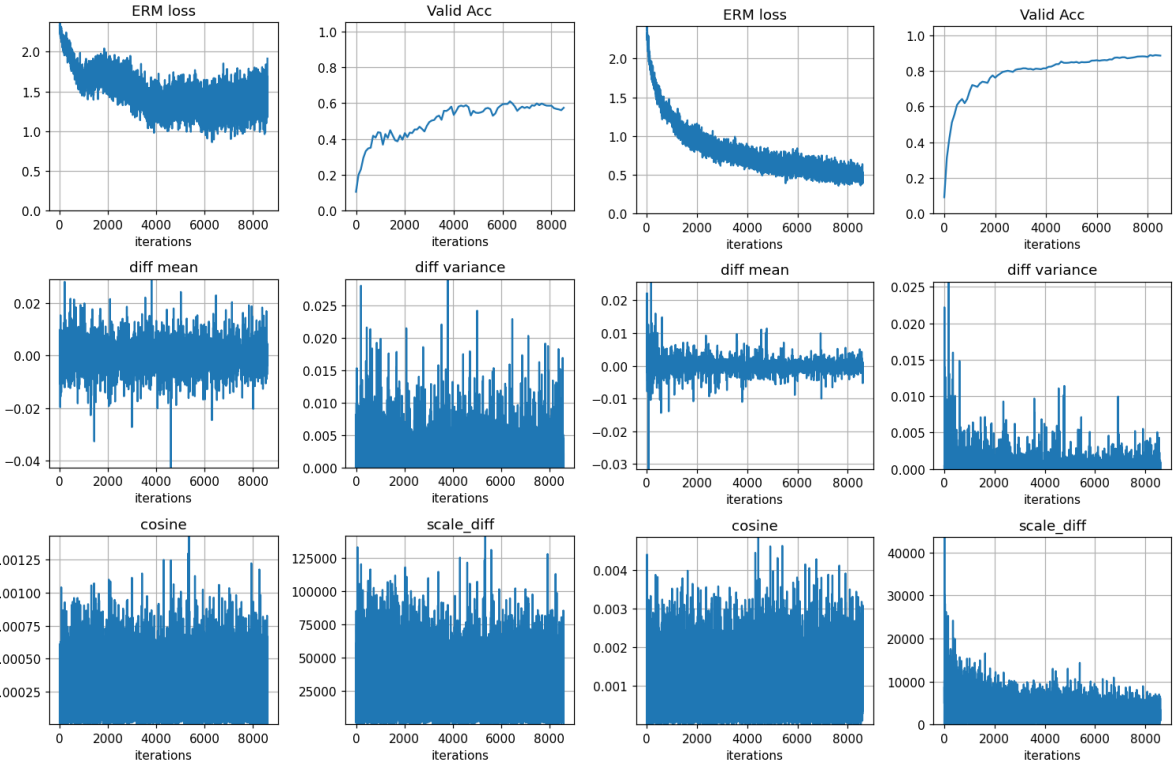
Lr = 1e-3 （num_dirs = 5）：



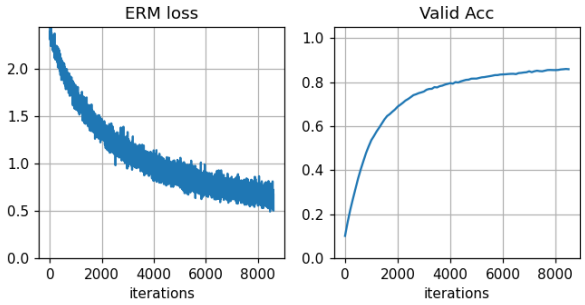


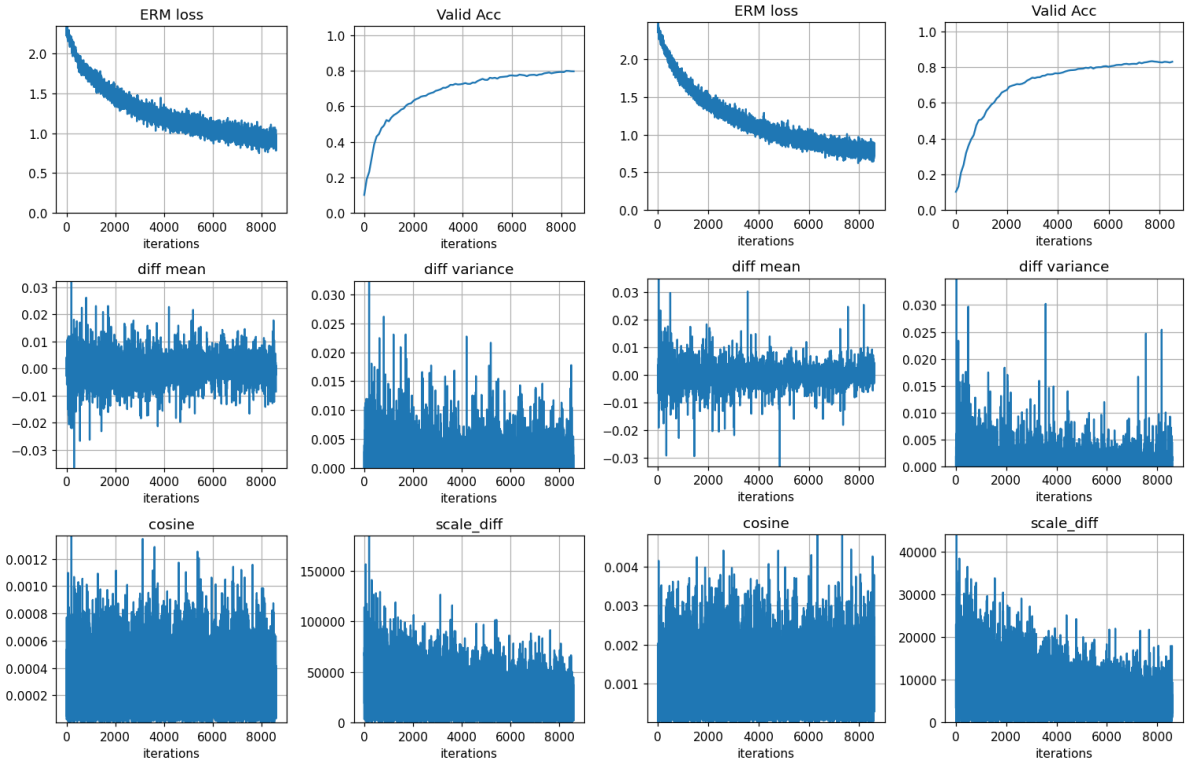
Lr = 1e-4


ERM loss
iterations
Valid Acc
iterations

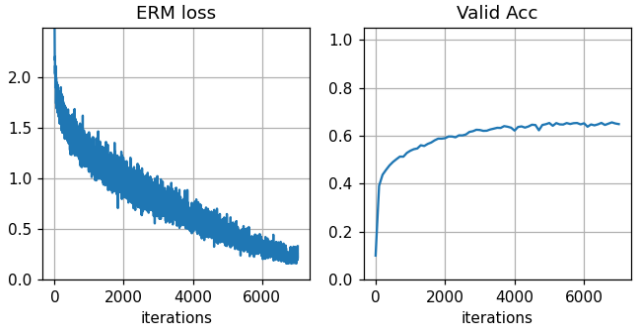

ERM loss
iterations
Valid Acc
iterations
ERM loss
iterations
Valid Acc
iterations
diff mean
iterations
diff variance
iterations
diff mean
iterations
diff variance
iterations
cosine
iterations
scale_diff
iterations
cosine
iterations
scale_diff
iterations

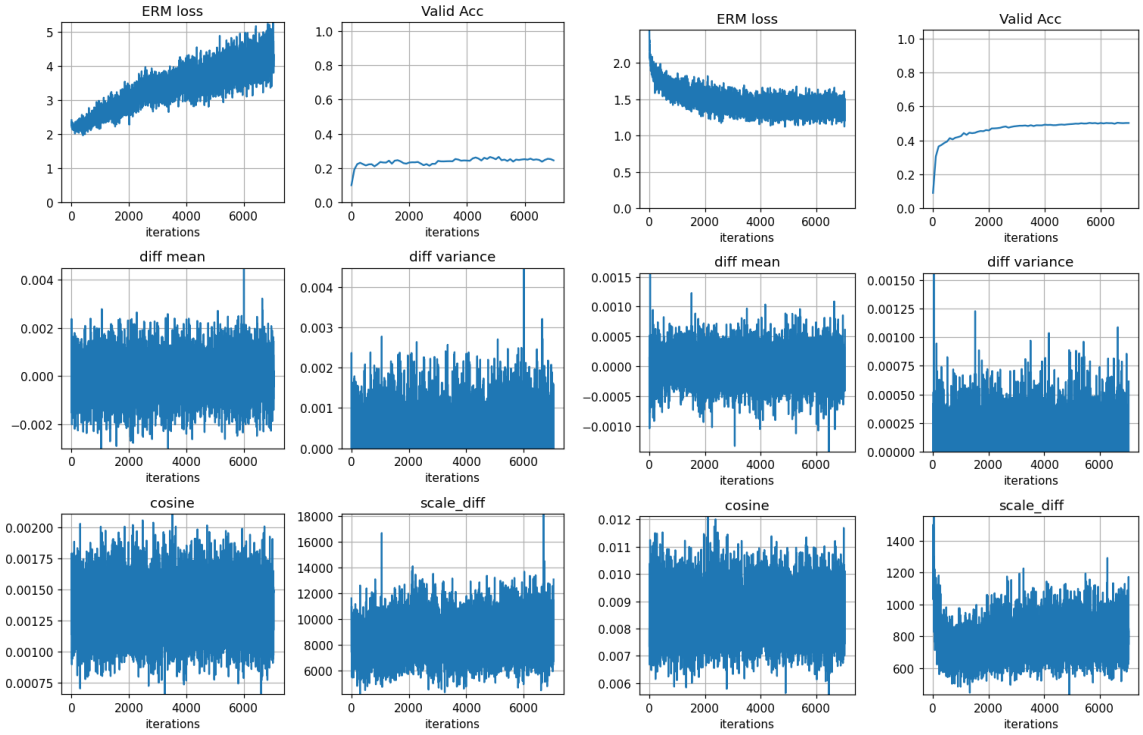
Lr = 1e-5:


ERM loss
iterations
Valid Acc
iterations

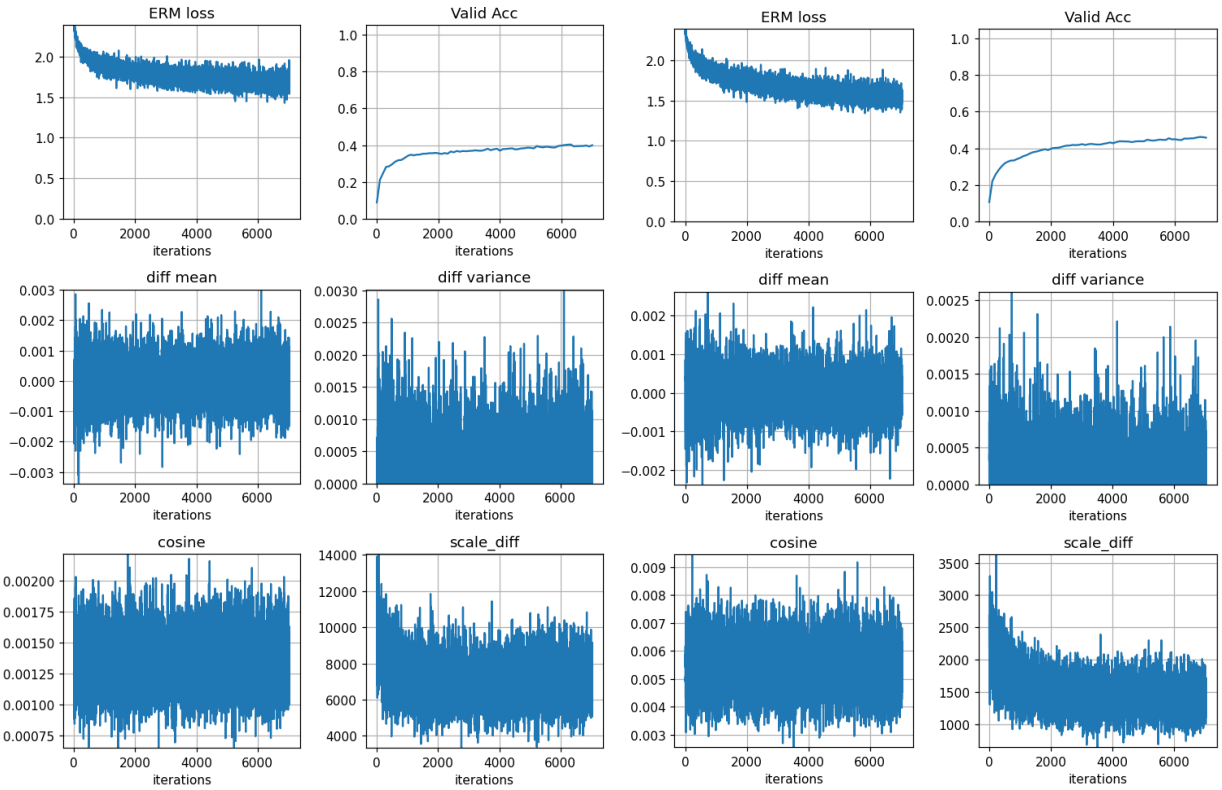

ERM loss
iterations
Valid Acc
iterations
ERM loss
iterations
Valid Acc
iterations
diff mean
iterations
diff variance
iterations
diff mean
iterations
diff variance
iterations
cosine
iterations
scale_diff
iterations
cosine
iterations
scale_diff
iterations

Cifar-10, VGG like, ~12 milli parameters: num_dir = 10, lr=1e-3





Lr = 1e-4:



Conclusion:

1. We can get the unbiased estimation, but do we have to do that?
2. The gain of the performance may not come from the decrease of variance but from the increase of cosine similarity.