On the Distributional Properties of Adaptive Gradients (Supplementary Material)

Zhang Zhiyi
1School of Statistics, Xi’an University of Finance and Economics
Liu Ziyin
2Department of Physics, University of Tokyo

A ADDITIONAL EXPERIMENTS

A.1 DISTRIBUTION ACROSS DIFFERENT RANDOM SEEDS

We also plot compare the distribution over different random initializations in Figure 1. We plot the overlap of three different random seed in dark red, which also agrees well with the prediction. We see that the variance of the distribution across different seeds is relatively small, and all agree well with the theoretical prediction.

Figure 1: Update distribution for different random seeds.

A.2 OTHER EXPERIMENTS ON CIFAR-10

In this section, we present more experiments to validify our theory. We first plot the distribution of update for a different layer (from what appeared in the main text) RegNetX-200MF trained on CIFAR-10, and excellent agreement between our theory and experiment is observed. We then plot the distribution of updates from random layers (with more than $10^4$ parameters) of VGG [Simonyan and Zisserman, 2014], ResNet-18 [He et al., 2016], ShuffleNet-V2 [Zhang et al., 2018], ResNeXt-29 [Xie et al., 2017], MobileNet [Howard et al., 2017], EfficientNet-B0 [Tan and Le, 2019], DenseNet-121 [Huang et al., 2017]. We first plot the variance of $\theta$ and $|\theta|$ in Figure 2. While some nets deviate from the theory, others agree quite well. For those that deviate from the prediction, we observe that they all have smaller variances than the predicted value, this agrees with our message that the adaptive gradient methods do not have exploding variance problem at the beginning of training. We then plot the distributions of $\theta$ for a single trajectory for each of these models.
Figure 2: Left: variance of $\theta$. Right: variance of $|\theta|$. We see that, some architectures agree quite well our analysis, while other architectures deviates in a rather consistent way.
Figure 3: Distribution of the update distribution of another non-hand-picked layer of a RegNetX-200MF trained on CIFAR-10.
Figure 4: Distribution of the update distribution of another non-hand-picked layer of a MobileNet trained on CIFAR-10.

Figure 5: Distribution of the update distribution of another non-hand-picked layer of a EfficientNet trained on CIFAR-10.
Figure 6: Distribution of the update distribution of another non-hand-picked layer of a ResNet-18 trained on CIFAR-10.

Figure 7: Distribution of the update distribution of another non-hand-picked layer of a VGG trained on CIFAR-10.
Figure 8: Distribution of the update distribution of another non-hand-picked layer of a ShuffleNet trained on CIFAR-10.

Figure 9: Distribution of the update distribution of another non-hand-picked layer of a ResNeXt29 trained on CIFAR-10.
References


