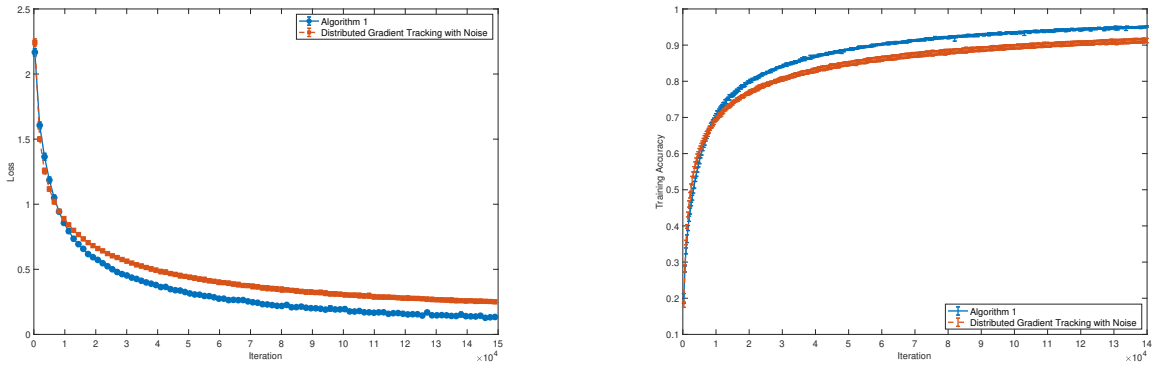

(a) **(b)**

Fig. 1. Comparison between the proposed algorithm and distributed gradient tracking with noise: (a) Loss function values; (b) Training accuracy on CIFAR-10 dataset. For this experiment, we consider the training of a convolutional neural network (CNN) for the classification of the CIFAR-10 dataset, which contains 50,000 training images across 10 different classes. We evenly spread the CIFAR-10 dataset to the five agents and set the batch size as 32. Our baseline CNN architecture is a deep network ResNet-18, the training of which is a highly nonconvex problem characterized by the presence of many saddle points. In the experiments, we train the CNN using both the proposed Algorithm 1 and the distributed gradient tracking method with noise. For our algorithm, the noise amplitude, noise-injection interval and clipping threshold were set to $\theta = 1 \times 10^{-4}$, $\mathcal{K}_0 = 100$ and $c_0 = 0.7$, respectively. The stepsize α used for algorithm 1 is 0.05 and for distributed gradient tracking is 0.02 (the largest stepsize that can still ensure convergence). In order to ensure fairness in comparison, both algorithms use the same noise amplitude.