

Figure 1: Test accuracy over communication rounds of FedAvg compared to SGD with IID and non-IID data of (a) MNIST (b) CIFAR-10 and (c) KWS datasets. Non-IID(2) represents the 2-class non-IID and non-IID(1) represents the 1-class non-IID.

**0.1. Experimental Results**

For the IID experiments, the convergence curves of FedAvg with a batch size of mostly overlap with the curves of SGD with for all the three datasets (Figure 11). Only a small difference is observed for CIFAR-10 that FedAvg with converges to 82.62% but SGD with converges to (Figure A.1. Thus, FedAvg achieves SGD-level test accuracy for IID data, which is consistent with the results in [3].

Significant reduction in the test accuracy is observed for FedAvg on non-IID data compared to SGD with matched batch size (Figure 1 and A.1). The accuracy reduction of non-IID data is summarized in Table1. The maximum accuracy reduction occurs for the most extreme 1-class non-IID data. Moreover, a larger number of local epochs doesn't reduce the loss. The convergence curves mostly overlap for and . Furthermore, the models pre-trained by SGD doesn't learn from the FedAvg training on non-IID data. For CIFAR-10, the accuracy drops when the pre-trained CNN is trained by FedAvg on non-IID data. Thus, we demonstrate the reduction in the test accuracy of FedAvg for non-IID data. The test accuracy of all the experiments is summarized in Table A.2 Note that the SGD accuracy reported in this paper are not state-of-the-art [6, 21, 22, 1] but the CNNs we train are sufficient for our goal to evaluate federated learning on non-IID data.

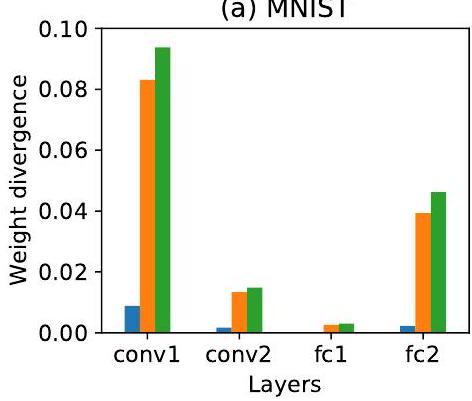
Table 1: The reduction in the test accuracy of FedAvg for non-IID data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Non-IID | B | E | MNIST (%) | CIFAR-10 (%) | KWS (%) |
| Non-IID(1) | large | 1 | 6.52 | 37.66 | 43.64 |
| Non-IID(1) | large | 5 | 6.77 | 37.11 | 43.62 |
| Non-IID(2) | large | 1 | 2.4 | 14.51 | 12.16 |
| Non-IID(1) | small | 1 | 11.31 | 51.31 | 54.5 |
| Non-IID(2) | small | 1 | 1.77 | 15.61 | 15.07 |

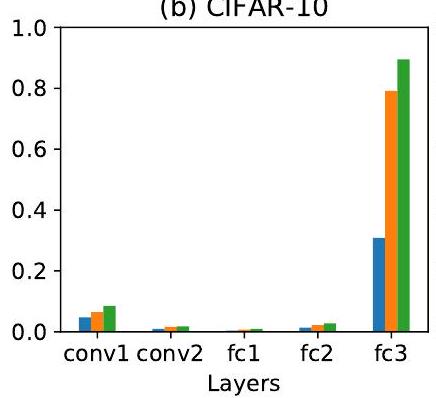
**1. Weight Divergence due to Non-IID Data**

In Figure 1 and A.1, it is interesting to note that the reduction is less for the 2-class non-IID data than for the 1-class non-IID data. It indicates that the accuracy of FedAvg may be affected by the exact data distribution, i.e., the skewness of the data distribution. Since the test accuracy is dictated by the trained weights, another way to compare FedAvg with SGD is to look at the difference of the weights relative to those of SGD, with the same weight initialization. It is termed as weight divergence and it can be computed by the following equation:

As shown in Figure 2, the weight divergence of all the layers increases as the data become more non-IID, from IID to 2-class non-IID to 1-class non-IID. Thus, an association between the weight divergence and the skewness of the data is expected. The accuracy reduction found in Section 2 can be understood in terms of weight divergence, which quantifies the difference of weights from two (a) MNIST



(b) CIFAR-10



(c) KWS

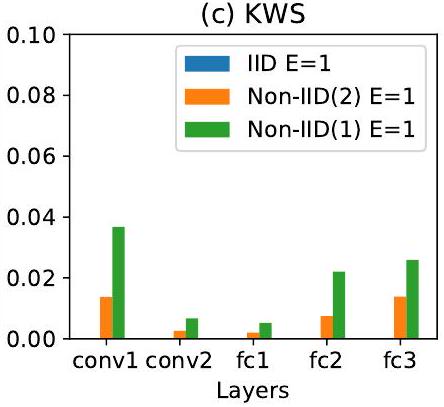


Figure 2: Weight divergence of CNN layers for IID, 2-class non-IID and 1-class non-IID.

different training processes with the same weight initialization. In this section, we formally analyze the origin of the weight divergence. In Section 3.1, we provide an illustrative example and a formal proposition to demonstrate that the root cause of the weight divergence is due to the distance between the data distribution on each client and the population distribution. Specifically, we find such distance can be evaluated with the earth mover's distance (EMD) between the distributions. Then, in Section 3.2. we validate the proposition with experiments that demonstrate the impact of EMD on the weight divergence and test accuracy of FedAvg.

**1.1. Mathematical demonstration**

We formally define the problem of federated learning and analyze the origin of the weight divergence. We consider a class classification problem defined over a compact space and a label space , where . The data point distributes over following the distribution . A function maps to the probability simplex , where with denoting the probability for the th class. is parameterized over the hypothesis class , i.e., the weight of the neural network. We define the population loss with the widely used cross-entropy loss as

To simplify the analysis, we ignore the generalization error and assume the population loss is optimized directly. Therefore, the learning problem becomes

To determine , SGD solves the optimization problem iteratively. Let denotes the weight after -th update in the centralized setting. Then, centralized SGD performs the following update:

In federated learning, we assume there are clients. Let denote the amount of data and denote the data distribution on client . On each client, local SGD is conducted separately. At iteration on client , local SGD performs

Then, assume the synchronization is conducted every steps and let denote the weight calculated after the -th synchronization, then, we have

The divergence between and can be understood with the illustration in Figure 3 . When data is IID, for each client , the divergence between and is small and after the -th synchronization, is still close to . When data is non-IID, for each client , due to the distance between the data distribution, the divergence between and becomes much larger and accumulates very fast, which makes the divergence between and much larger. To formally bound the weight divergence between and , we have the following proposition.

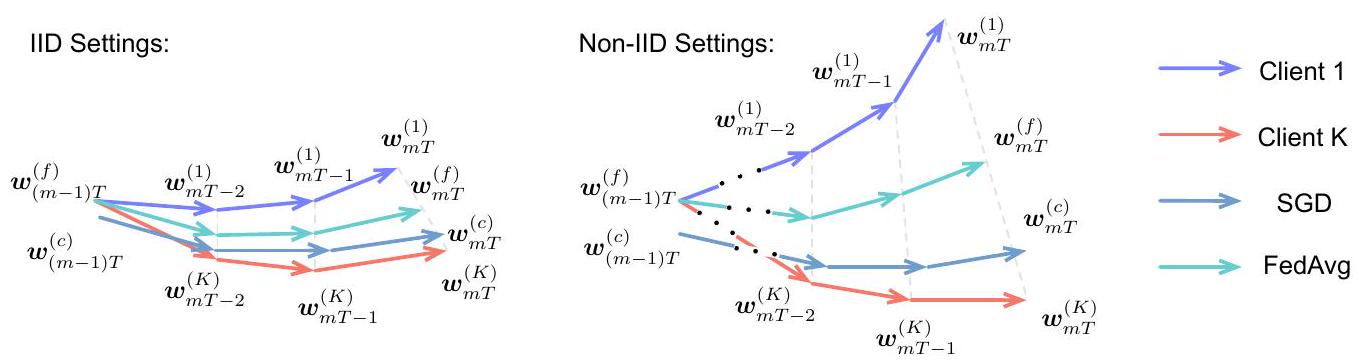


Figure 3: Illustration of the weight divergence for federated learning with IID and non-IID data.

Proposition 3.1. Given clients, each with i.i.d samples following distribution for client . If is -Lipschitz for each class and the synchronization is conducted every steps, then, we have the following inequality for the weight divergence after the -th synchronization,

where and .

Detailed proof of Proposition 3.1 can be found in Appendix A.3 Based on Proposition 3.1 we have the following remarks.

Remark 3.2. The weight divergence after the -th synchronization mainly comes from two parts, including the weight divergence after the -th divergence, i.e., , and the weight divergence induced by the probability distance for the data distribution on client compared with the actual distribution for the whole population, i.e., .

Remark 3.3. The weight divergence after the -th synchronization is amplified by . Since . Hence, if different clients start from different initial in the federated learning, then, even if the data is IID, large weight divergence will still be encountered, which leads to degraded accuracy.

Remark 3.4. When all the clients start from the same initialization as the centralized settings, becomes the root cause of the weight divergence. This term is EMD between the data distribution on client and the population distribution, when the distance measurement is defined as . The impact of EMD is affected by the learning rate , the number of steps before synchronization , and the gradient .

Based on Proposition 3.1, we validate that EMD is a good metric to quantify the weight divergence and thus the test accuracy of FedAvg with non-IID data in Section 3.2

**1.2. Experimental Validation**

**1.2.1. Experimental Setup**

The training set is sorted and partitioned into 10 clients, examples per client. Eight values are chosen for EMD listed in Table 2. Because there may exist various distributions for one EMD, we aim to generate five distributions to compute the average and variations of the weight divergence and the test accuracy. First, one probability distribution over 10 classes is generated for one EMD. Based on and , we can compute the number of examples over 10 classes for one client. Second, a new distribution is generated by shifting the 10 probabilities of by 1 element. The number of examples for the second client can be computed based on . This procedure is repeated for the other 8 clients. Thus, all the 10 clients have a distribution of examples over 10 classes and each example is only used once. Finally, the above two steps are repeated 5 times to generate 5 distributions for each EMD. The CNNs are trained by FedAvg over 500 communication rounds on the data processed from the above procedures. There are the key parameters used for training: for MNIST, , decay rate ; for CIFAR-10, , decay rate ; for KWS, , decay rate . The weight divergence is computed according to Eq. (1) after 1 synchronization (i.e., 1 communication round).

**1.2.2. Weight Divergence vs. EMD**

The mean and standard deviation of the weight divergence are computed over 5 distributions for each EMD. For all the three datasets, the weight divergence of each layer increases with EMD as shown in Figure 4 The initial weights are identical for all SGD, IID and non-IID experiments on each dataset. Thus, according to Remark 3.2. the weight divergence after 1 synchronization is not affected by the -th divergence, , because it is zero when . Therefore, the results in Figure 4 support Proposition 3.1 that the bound of weight divergence is affected by EMD. This effect is more significant in the first convolutional layer and the last fully connected layer. Moreover, the maximum weight divergence for CIFAR-10 is significantly higher than that for MNIST and KWS, which is affected by the gradient term in Eq. (2) due to the problem itself and different CNN architectures. Note that the initial weights are also identical across the clients to avoid accuracy loss according to Remark 3.3. which is consistent with the significant increase in the loss when averaging models with different initialization [3, 23].

**1.2.3. Test Accuracy vs. EMD**

The mean and standard deviation of the test accuracy are computed over the same 5 distributions for each EMD. The results are summarized in Table 2 and are plotted against EMD in Figure 5 For all the three datasets, the test accuracy decreases with EMD. The decreasing rate is relatively small at first and becomes larger as the data becomes more non-IID. So there is a trade-off between balancing non-IID data towards IID and improving the accuracy of FedAvg. The error bars on the plots represent the variations of test accuracy due to various distributions for each EMD. To take a closer look at the variations, the boxplots show the test accuracy across 5 runs when EMD . Moreover, Table 2 2 shows that the maximum variation is less than for MNIST,

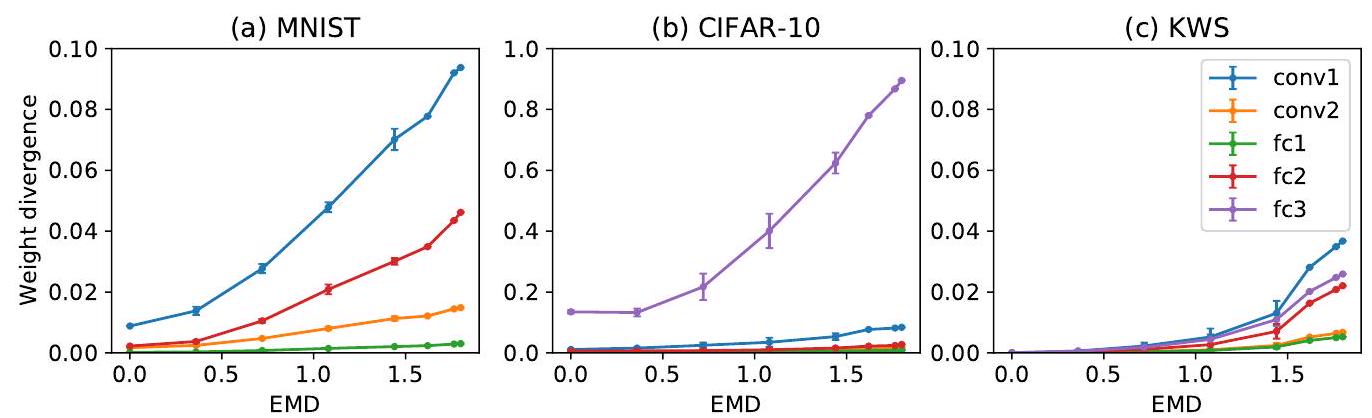


Figure 4: Weight divergence vs. EMD across CNN layers on (a) MNIST, (b) CIFAR-10 and (c) KWS datasets. The mean value and standard deviation are computed over 5 distributions for each EMD.