Supplementary Material - Game of Gradients: Mitigating Irrelevant clients in Federated Learning

A.1. Shapley values is an unique solution to four Axioms

In the main paper, we presented the Game Theory preliminaries in Section 4.1. There, we stated that Shapley values is an unique solution to four Axioms. Here we state the four Axioms.

(i) *efficiency* — all the payoff of the grand coalition is distributed among players;

(ii) *symmetry* — if two agents play the same role in any coalition they belong to (i.e. they are symmetric) then their payoff should also be symmetric;

(iii) *null player* — agents with no marginal contributions to any coalitions whatsoever should receive no payoff from the grand coalition; and

(iv) *additivity* — values of two uncorrelated games sum up to the value computed for the sum of both games.

A.2. More Datasets and Corresponding Experimental Configuration

We show the experimental results for *KWS* (Warden 2018) dataset in this supplementary material. KWS is a speech command recognition dataset. It consists of several speech commands that span 1 second each. The original dataset has 35 labels. We select a subset of it with speech commands for digits 0 to 9. We consider the labels corresponding to even digits as *relevant labels* and the ones corresponding to odd digits as *irrelevant labels*. As part of pre-processing, we first ensure that all the commands span 1 second. We pad the speech sequences that fall short with zeros and truncate the commands that exceed 1 second. We then convert each speech signal into a spectrogram of dimension 99×81 which is then used as input to the neural network model.

The objective of central server is to build a 5 class classification model. The data partitioning scheme for relevant and irrelevant cases is exactly as mentioned in Section 3.1 in the main paper.

We build a CNN based model at the central server with the following architecture. It has input layer followed by 3 convolutional layers having $8, 16, 32$ filters of size 3×3 respectively. Each of these convolutional layers are followed by a max pooling layer with filter of size 2×2. We flatten the output of the last pooling layer and pass it to a fully connected layer with 64 neurons. It is then followed with the softmax output layer. We use ReLU activations in all the layers. The cross entropy loss is minimized by Adam optimizer.

Hyper-parameter Values: The values of hyper-parameters that we use in these experiments are: $K = 10; m = 5; B = 32; \eta_i = 0.001; \alpha = 0.75; \beta = 0.5; \zeta = 5; R = 10; 0.\gamma = 0.95; \lambda = 2\%; \beta = 5$.

With the above experimental setup, the standard Federated averaging algorithm yields an accuracy of around 63% as shown in Figure 1 with the irrelevant data case.

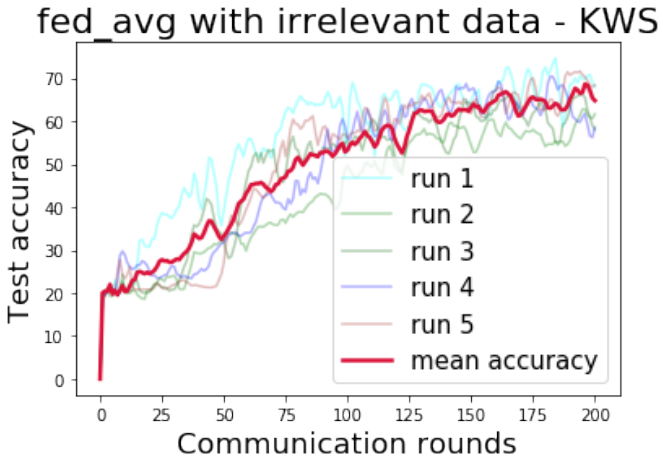

Figure 1: Test accuracy of the standard Federated Averaging algorithm using the KWS dataset in the irrelevant data case

A.3. More Experimental Results on S-FedAvg selects Relevant clients

Figure 2 shows the results of S-FedAvg algorithm using KWS dataset in the irrelevant non-iid setting. Figure 2.(a) shows the accuracy of the model on the test set across each communication round. It can be seen that the accuracy converges smoothly using S-FedAvg algorithm unlike the standard federated averaging algorithm (Figure 1). Figure 2.(b)

Table 1: This shows the way data is distributed among clients in relevant and irrelevant cases for the MNIST dataset. For Relevant data case, all ten are relevant clients. For irrelevant data case, first six are relevant and the last four are irrelevant clients respectively.

Relevant Data Case

Client	label0	label2	label4	label6	label8
1	2949	0	0	0	0
2	2949	0	0	0	0
3	25	2924	0	0	0
4	0	2949	0	0	0
5	0	85	2864	0	0
6	0	0	2949	0	0
7	0	0	29	2920	0
8	0	0	0	2949	0
9	0	0	0	40	2900
10	0	0	0	0	2949

Irrelevant Data Case

Client	label0	label2	label4	label6	label8
1	4915	0	0	0	0
2	1008	3907	0	0	0
3	0	2051	2864	0	0
4	0	0	2978	1937	0
5	0	0	0	3981	934
6	0	0	0	0	4915
7	6131	1496	0	0	0
8	0	4769	2858	0	0
9	0	0	2563	5064	0
10	0	0	0	885	6742

Table 2: This shows the way data is distributed among clients in relevant and irrelevant cases for the KWS dataset. For Relevant data case, all ten are relevant clients. For irrelevant data case, first six are relevant and the last four are irrelevant clients respectively.

Relevant Data Case

Client	label0	label2	label4	label6	label8
1	929	0	0	0	0
2	929	0	0	0	0
3	8	921	0	0	0
4	0	929	0	0	0
5	0	23	906	0	0
6	0	0	929	0	0
7	0	0	4	925	0
8	0	0	0	929	0
9	0	0	0	9	929
10	0	0	0	0	929

Irrelevant Data Case

Client	label0	label2	label4	label6	label8
1	1548	0	0	0	0
2	318	1230	0	0	0
3	0	643	905	0	0
4	0	0	934	614	0
5	0	0	0	1249	299
6	0	0	0	0	1548
7	1875	456	0	0	0
8	0	1388	943	0	0
9	0	0	949	1382	0
10	0	0	0	459	1872

shows the evolution of Relevance values of the clients across different rounds. In this figure, *green curves* correspond to Relevance values of the relevant clients and *red curves* correspond to the irrelevant clients respectively. We observe flat red curves because of the use of softmax function to convert Relevance values to probability distribution (Gao and Pavel 2017).

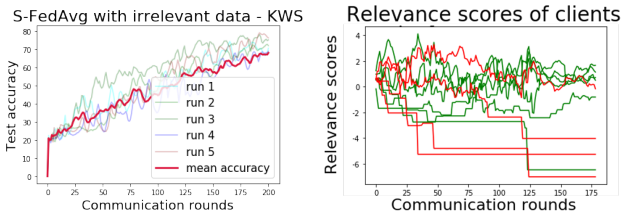

Figure 2: Performance of the proposed S-FedAvg Algorithm using the KWS dataset in irrelevant data setting

A.4. More Experimental Results on Class-specific Best Client Selection Problem

In the main paper, we showed results only for class 2 in Section 5.3 Class-specific Best Client Selection. Here we show the results for all classes for MNIST relevant data setting in Figure 5 and irrelevant data setting in Figure 6 respectively.

The experimental results using KWS dataset are shown in Figure 7 for relevant data configuration and Figure 8 for the irrelevant data configuration respectively. In each of these plots, the green curves correspond to Relevance values of the clients with data of class c. It is easy to see from the plots that the clients that possess data of label c do have relatively higher Relevance values than the ones without data of label c. This separation is more explicit in the relevant data setting as the gap between the green and red curves is more significant.

A.5. More Experimental Results on Data Label Standardization

The results for data label standardization are shown in Figure 4. We consider the irrelevant data case and manually corrupt the labels of client 3 (in 2) by permuting the labels 2 and 4. Figures 4 (a) and (b) show the Relevance values of clients across communication rounds using S-FedAvg and S-FedAvg-Label-Std respectively. The green curves correspond to relevant clients and red curves correspond to irrelevant clients respectively. Here the Relevance value of the corrupted client is denoted by the blue curve. From Figure 4 (b), we observe that label permutation occurred at communication round 73. As expected, the Relevance value for the corrupted client (i.e. client 3) is heavily penalized by S-FedAvg. Figure 4 (c) shows the confusion matrix of predictions obtained from the downloaded server model using parameters θ_{73}.

As we can see in the column corresponding to label 2, predictions from the model are in favor of label 4 and hence the corrupted client swaps label 2 with label 4. Figure 4 (d)

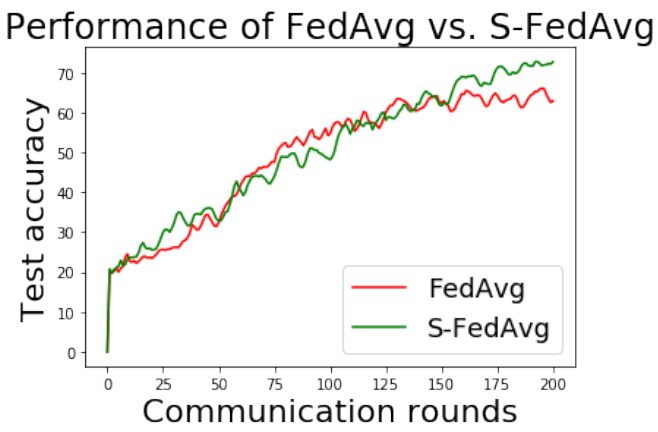

Figure 3: Performance comparison of FedAvg and S-FedAvg on KWS dataset

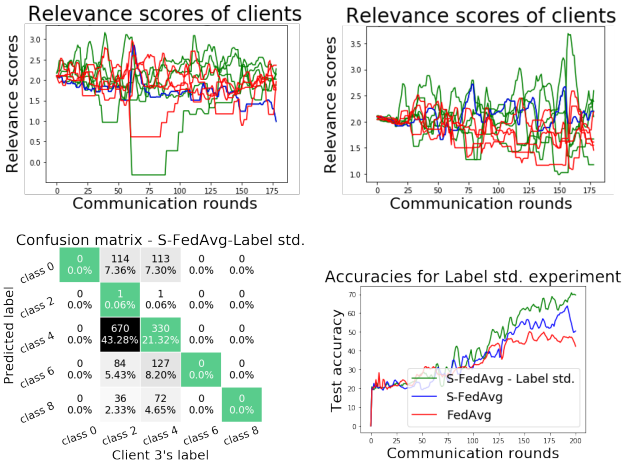

Figure 4: Experimental results for Data label standardization experiment using KWS dataset

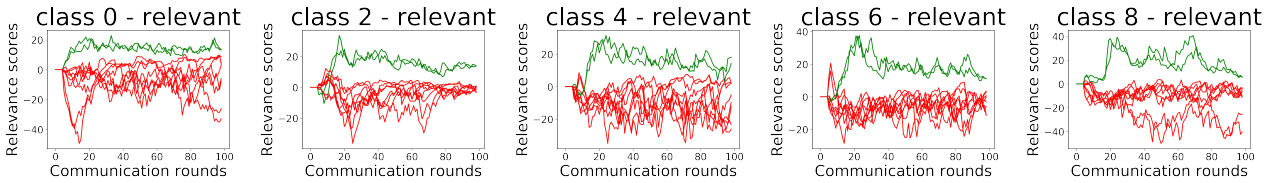

Figure 5: Class-specific Relevance value (φ_c) for MNIST in Relevant data case for classes $C = \{0, 2, 4, 6, 8\}$.

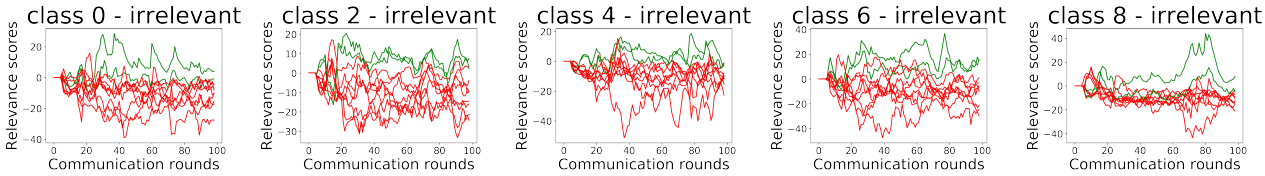

Figure 6: Class-specific Relevance value (φ_c) for MNIST in Irrelevant data case for classes $C = \{0, 2, 4, 6, 8\}$.

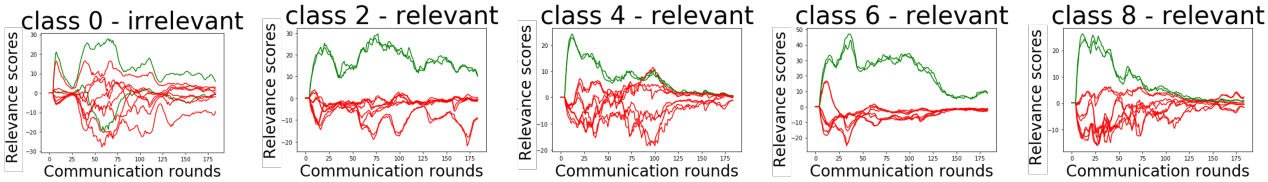

Figure 7: Class-specific Relevance value (φ_c) for KWS in Relevant data case for classes $C = \{0, 2, 4, 6, 8\}$.

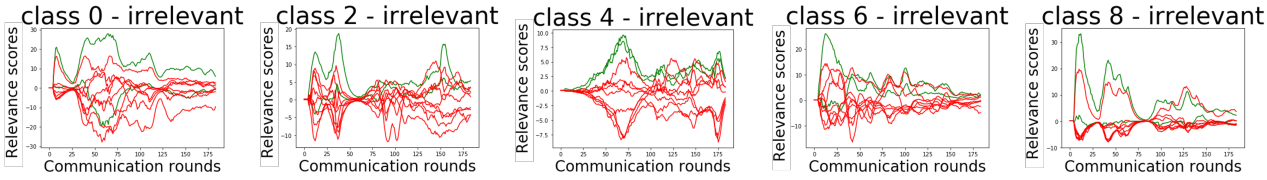

Figure 8: Class-specific Relevance value (φ_c) for KWS in Irrelevant data case for classes $C = \{0, 2, 4, 6, 8\}$.

shows the test performance of three algorithms viz., FedAvg, S-FedAvg, and S-FedAvg-Label-Std. The performance of S-FedAvg-Label-Std is superior to that of S-FedAvg and FedAvg; and this performance gain is due to our proposed data label standardization procedure.

References

Gao, B.; and Pavel, L. 2017. On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning.

Warden, P. 2018. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition.