Supplementary Material: Resilient Peer-to-peer Learning based on Adaptive Aggregation

A Derivation of Optimal Aggregation Weights

The derivation of the optimal aggregation weights is presented here. The optimization problem was formulated in Section III as

$$\begin{aligned} &\min_{C_k} \left\| \sum_{l \in \mathcal{N}_k} c_t(l,k) \hat{w}_k^t - w_k^* \right\|^2 \\ &\text{subject to} \sum_{l \in \mathcal{N}_k} c_t(l,k) = 1, \quad c_t(l,k) \geq 0, \quad c_t(l,k) = 0 \text{ for } l \notin \mathcal{N}_k. \end{aligned} \tag{A.1}$$

The optimization function in (A.1) can be rewritten as

$$\left\| \sum_{l \in \mathcal{N}_k} c_t(l,k) \hat{w}_l^t - w_k^* \right\|^2 \approx \sum_{l \in \mathcal{N}_k} c_t(l,k)^2 \left\| \hat{w}_l^t - w_k^* \right\|^2. \tag{A.2}$$

With the initialization of the parameters of each worker within $\mathbb{B}(w_s^*, \Gamma)$, the risk function of each worker acts as m-strictly convex. Thus, it holds that

$$r_k(\hat{w}_l^t) - r_k(w_k^*) \geq \langle \nabla r_k(w_k^*), y - x \rangle + \frac{m}{2} \left\| \hat{w}_l^t - w_k^* \right\|.$$

Since $\nabla r_k(w_k^*) = 0$, we obtain

$$\left\| \hat{w}_l^t - w_k^* \right\|^2 \leq \frac{2}{m} \Big(r_k(\hat{w}_l^t) - r_k(w_k^*) \Big). \tag{A.3}$$

Instead of minimizing $\left\| \hat{w}_l^t - w_k^* \right\|^2$, we aim to minimize its upper bound, as given by (A.3). Thus, we reformulate the optimization problem (A.1) using (A.2) as

$$\begin{aligned} &\min_{C_k} \sum_{l \in \mathcal{N}_k} c_t(l,k)^2 \cdot \Big(r_k(\hat{w}_l^t) - r_k(w_k^*) \Big) \\ &\text{subject to} \sum_{l \in \mathcal{N}_k} c_t(l,k) = 1, \quad c_t(l,k) \geq 0, \quad c_t(l,k) = 0 \text{ for } l \notin \mathcal{N}_k. \end{aligned} \tag{A.4}$$

As $r_k(w_k^*)$ is small compared to $r_k(\hat{w}_l^t)$, we use $r_k(w_k^*) = 0$, which gives us the modified optimization problem [1]

$$\begin{aligned} &\min_{C_k} \sum_{l \in \mathcal{N}_k} \frac{1}{2} c_t(l,k)^2 \cdot r_k(\hat{w}_l^t) \\ &\text{subject to} \sum_{l \in \mathcal{N}_k} c_t(l,k) = 1, \quad c_t(l,k) \geq 0, \quad c_t(l,k) = 0 \text{ for } l \notin \mathcal{N}_k. \end{aligned} \tag{A.5}$$

[1] The factor of 1/2 is introduced for simplification of solution.

After incorporating the constraint $\sum_{l\in\mathcal{N}_k} c_t(l,k) = 1$, the Lagrangian of (A.5) is given by

$$\mathcal{L}\big(c_t(l,k),\lambda\big) = \sum_{l\in\mathcal{N}_k} \frac{1}{2} c_t^2(l,k) r_k(\hat{w}_l^t) + \lambda\Big(1 - \sum_{l\in\mathcal{N}_k} c_t(l,k)\Big), \tag{A.6}$$

where λ is the Lagrange multiplier. Now, taking the gradient of the Lagrangian w.r.t. $c_t(l,k)$ and equating it to zero, we get

$$c_t(l,k) r_k(\hat{w}_l^t) - \lambda = 0, \quad \forall l \in \mathcal{N}_k. \tag{A.7}$$

This yields

$$c_t(l,k) = \frac{\lambda}{r_k(\hat{w}_l^t)}, \quad \forall l \in \mathcal{N}_k.$$

Using this in the constraint, we get

$$\lambda \sum_{l\in\mathcal{N}_k} \frac{1}{r_k(\hat{w}_l^t)} = 1.$$

Thus, the Lagrange multiplier is given by

$$\lambda = \frac{1}{\sum\limits_{l\in\mathcal{N}_k} r_k(\hat{w}_l^t)^{-1}}.$$

Substituting this in (A.7), we get the optimal aggregation weights derived in Section III, given as

$$c_t(l,k) = \begin{cases} \frac{r_k(\hat{w}_l^t)^{-1}}{\sum\limits_{p\in\mathcal{N}_k} r_k(\hat{w}_p^t)^{-1}} & \text{for } l \in \mathcal{N}_k, \\ 0 & \text{for } l \notin \mathcal{N}_k. \end{cases} \tag{A.8}$$

B Detailed Proofs of Lemmas and Theorems

B.1 Proof of Lemma 1

Using the aggregation step $w_k^t = \sum_{l \in \mathcal{N}_k} c_t(l,k) \cdot \hat{w}_l^t$, the risk of worker k in epoch t is given by

$$r_k(w_k^t) = r_k\Big(\sum_{l \in \mathcal{N}_k} c(l,k) \cdot \hat{w}_l^t\Big)$$

Using Jensen's inequality, we can write

$$r_k(w_k^t) \leq \sum_{l \in \mathcal{N}_k} c_t(l,k) \cdot r_k(\hat{w}_l^t)$$

Subtracting $r_k(w_k^*)$ from both sides and taking expectations over the joint distributions ξ_k, we obtain

$$\begin{aligned} \mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*)\right] &\leq \sum_{l \in \mathcal{N}_k} \mathbb{E}\left[c_t(l,k)\right] \cdot \mathbb{E}\left[r_k(w_l^t) - r_k(w_k^*)\right] \\ &\leq \frac{\sum\limits_{l \in \mathcal{N}_k} \mathbb{E}\left[r_k(\hat{w}_l^t)\right]^{-1} \mathbb{E}\left[r_k(w_l^t) - r_k(w_k^*)\right]}{\sum\limits_{p \in \mathcal{N}_k} \mathbb{E}\left[r_k(\hat{w}_p^t)\right]^{-1}}, \end{aligned} \tag{B.1}$$

using (A.8). We now show that the right side of (B.1) is smaller than $\frac{1}{|\mathcal{N}_k|} \sum\limits_{l \in \mathcal{N}_k} \mathbb{E}\left[r_k(\hat{w}_l^t) - r_k(w_k^*)\right]$. Let us denote $\mathbb{E}\left[r_k(\hat{w}_l^t)\right]^{-1}$ as χ_l^t and $\mathbb{E}\left[r_k(w_l^t) - r_k(w_k^*)\right]$ as Δ_l^t. We aim to prove

$$\frac{\sum\limits_{l \in \mathcal{N}_k} \chi_l^t \cdot \Delta_l^t}{\sum\limits_{p \in \mathcal{N}_k} \chi_p^t} \leq \frac{1}{|\mathcal{N}_k|} \sum_{l \in \mathcal{N}_k} \Delta_l^t,$$

or, equivalently,

$$|\mathcal{N}_k| \sum_{l \in \mathcal{N}_k} \chi_l^t \cdot \Delta_l^t \leq \sum_{p \in \mathcal{N}_k} \chi_p^t \sum_{l \in \mathcal{N}_k} \Delta_l^t.$$

When $|\mathcal{N}_k| = 1$, it is trivial to see that this condition holds. When $|\mathcal{N}_k| \geq 2$, let l_1^t be the one with smallest risk, i.e., $r_k(\hat{w}_{l_1}^t) = \min\limits_{l \in \mathcal{N}_k} r_k(\hat{w}_l^t)$ and l_2^t be the one with the second smallest risk, i.e., $r_k(\hat{w}_{l_2}^t) = \min\limits_{l \in \mathcal{N}_k \setminus l_1^t} r_k(\hat{w}_l^t)$. As the risk function is locally m-strongly convex and also under the assumption of parameter initialization explained earlier, it

holds that $\chi_{l_1^t}^t \geq \chi_{l_2^t}^t \geq \chi_l^t$ and $\Delta_{l_1^t}^t \leq \Delta_{l_2^t}^t \leq \Delta_l^t$ for $l \in \mathcal{N}_k \setminus l_1^t, l_2^t$. Therefore,

$$
\begin{aligned}
&|\mathcal{N}_k| \sum_{l\in\mathcal{N}_k} \chi_l^t \Delta_l^t - \sum_{p\in\mathcal{N}_k} \chi_p^t \sum_{l\in\mathcal{N}_k} \Delta_l^t \\
&= \sum_{l\in\mathcal{N}_k} \chi_l^t \left(|\mathcal{N}_k|\, \Delta_l^t - \sum_{p\in\mathcal{N}_k} \Delta_p^t \right) \\
&= \chi_{l_1^t}^t \left(\Big(|\mathcal{N}_k| - 1\Big) \Delta_{l_1^t}^t - \sum_{l\in\mathcal{N}_k\setminus l_1^t} \Delta_l^t \right) + \sum_{l\in\mathcal{N}_k\setminus l_1^t} \chi_l^t \left(|\mathcal{N}_k|\, \Delta_l^t - \sum_{p\in\mathcal{N}_k} \Delta_p^t \right) \\
&\leq \chi_{l_1^t}^t \left(\Big(|\mathcal{N}_k| - 1\Big) \Delta_{l_1^t}^t - \sum_{l\in\mathcal{N}_k\setminus l_1^t} \Delta_l^t \right) + \chi_{l_2^t}^t \left(\sum_{l\in\mathcal{N}_k\setminus l_1^t} |\mathcal{N}_k|\, \Delta_l^t - \Big(|\mathcal{N}_k| - 1\Big) \sum_{p\in\mathcal{N}_k} \Delta_p^t \right) \\
&= \chi_{l_1^t}^t \left(\Big(|\mathcal{N}_k| - 1\Big) \Delta_{l_1^t}^t - \sum_{l\in\mathcal{N}_k\setminus l_1^t} \Delta_l^t \right) + \chi_{l_2^t}^t \left(\sum_{l\in\mathcal{N}_k\setminus l_1^t} \Delta_l^t - \Big(|\mathcal{N}_k| - 1\Big) \Delta_{l_1^t}^t \right) \\
&= \Big(\chi_{l_1^t}^t - \chi_{l_2^t}^t\Big) \left(\Big(|\mathcal{N}_k| - 1\Big) \Delta_{l_1^t}^t - \sum_{l\in\mathcal{N}_k\setminus l_1^t} \Delta_l^t \right) \\
&= \Big(\chi_{l_1^t}^t - \chi_{l_2^t}^t\Big) \left(\sum_{l\in\mathcal{N}_k\setminus l_1^t} \Big(\Delta_{l_1^t}^t - \Delta_l^t\Big) \right) \leq 0.
\end{aligned}
$$

Therefore, $\frac{\sum_{l\in\mathcal{N}_k} \chi_l^t \Delta_l^t}{\sum_{p\in\mathcal{N}_k} \chi_p^t} \leq \frac{1}{|\mathcal{N}_k|} \sum_{l\in\mathcal{N}_k} \Delta_l^t$. By plugging it back to (B.1), we obtain

$$
\mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*) \right] \leq \frac{1}{|\mathcal{N}_k|} \sum_{l\in\mathcal{N}_k} \mathbb{E}\left[r_k(\hat{w}_l^t) - r_k(w_k^*) \right], \tag{B.2}
$$

which completes the proof. ■

B.2 Proof of Theorem 1

As the set of selected neighbors $\mathcal{N}_k^+$ is a subset of $\mathcal{N}_k$, we know that $\left|\mathcal{N}_k^+\right| \leq |\mathcal{N}_k|$. Thus,

$$
\frac{1}{|\mathcal{N}_k|} \sum_{l\in\mathcal{N}_k} \mathbb{E}\left[r_k(\hat{w}_l^t) - r_k(w_k^*) \right] \leq \frac{1}{\left|\mathcal{N}_k^+\right|} \sum_{l\in\mathcal{N}_k} \mathbb{E}\left[r_k(\hat{w}_l^t) - r_k(w_k^*) \right].
$$

Using this in Lemma 1, we have

$$
\mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*) \right] \leq \frac{1}{\left|\mathcal{N}_k^+\right|} \sum_{l\in\mathcal{N}_k} \mathbb{E}\left[r_k(\hat{w}_l^t) - r_k(w_k^*) \right]. \tag{B.3}
$$

Let $\mathbb{E}[\cdot]$ denote the expected value taken with respect to the joint distribution of all random variables ξ_k and ξ_l for $l \in \mathcal{N}_k^+$, i.e., $\mathbb{E}[\cdot] = \mathbb{E}_{\xi_k} \mathbb{E}_{\{\xi_l | l\in\mathcal{N}_k^+\}}[\cdot]$. For every

$l \in \mathcal{N}_k^+$, we have $r_k(\hat{w}_l^t) \leq r_k(\hat{w}_k^t)$. Thus,

$$\frac{1}{\left|\mathcal{N}_k^+\right|} \sum_{l \in \mathcal{N}_k} \mathbb{E}\left[r_k(\hat{w}_l^t) - r_k(w_k^*)\right] \leq \mathbb{E}\left[r_k(\hat{w}_k^t) - r_k(w_k^*)\right].$$

Using this in Lemma 1, we get

$$\mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*)\right] \leq \mathbb{E}\left[r_k(\hat{w}_k^t) - r_k(w_k^*)\right], \quad \forall k \in \mathcal{V}, t \in \mathbb{N}. \tag{B.4}$$

We now prove the convergence of the parameters with the proposed aggregation method. Given Assumptions 1-4, also when the parameters of the normal workers are initialized within $\mathbb{B}(w_s^*, \Gamma)$, we obtain from [1] that using constant step size $\mu_k \in (0, \frac{1}{La_k}]$, it holds that

$$\mathbb{E}\left[r_k(\hat{w}_k^t) - r_k(w_k^*)\right] - \frac{\mu_k L \sigma_k^2}{2m} \leq (1 - \mu_k m)\left(\mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*)\right] - \frac{\mu_k L \sigma_k^2}{2m}\right)$$

This, combined with (B.4), yields

$$\mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*)\right] - \frac{\mu_k L \sigma_k^2}{2m} \leq (1 - \mu_k m)\left(\mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*)\right] - \frac{\mu_k L \sigma_k^2}{2m}\right). \tag{B.5}$$

For $\mu_k \in (0, \frac{1}{La_k}]$ with $a_k \geq 1, m \leq L$, it holds that $(1 - \mu_k m) \in [0, 1)$. Applying (B.5) repeatedly through iteration $t \in \mathbb{N}$, we obtain

$$\mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*)\right] \leq \frac{\mu_k L \sigma_k^2}{2m} + (1 - \mu_k m)^t \left(r_k(w_k^0) - r_k(w_k^*) - \frac{\mu_k L \sigma_k^2}{2m}\right)$$

As $(1 - \mu_k m) \in [0, 1)$, when $t \to \infty$, we get $\mathbb{E}\left[r_k(w_k^t) - r_k(w_k^*)\right] \to \frac{\mu_k L \sigma_k^2}{2m}$. This means that w_k^t converges towards its optimal value w_k^* with the expected regret bounded by $\frac{\mu_k L \sigma_k^2}{2m}$. ■

C Details of Empirical Evaluation

– **Human Activity Recognition on UCI-HAR Dataset:** Human activity recognition is a popular ML task where the goal is to predict activities of a human based on the sensor readings. We use the UCI HAR dataset which contains data collected from the accelerometer and gyroscope sensors embedded in smartphones worn by participants during various physical activities. The data is collected from 30 humans performing the following six activities: walking, walking-upstairs, walking-downstairs, sitting, standing, lying-down.
We consider 30 worker machines where each worker corresponds to one user, which makes the data distribution non-iid. The private data of each worker is split into 75% training data and 25% testing data. The sensor readings are embedded in the form of a feature vector of length 561, which is the input to a fully connected neural network model. The details of this network are given in Table. C.1. For this task, we used cross-entropy loss, a learning rate of 0.01 and a batch-size of 10. Two attacked scenarios are simulated here - 13 attacked workers, which is the highest number of adversarial workers allowed in Krum aggregation, and another scenario with $6(\ 20\%)$ attacked neighbors.

Table C.1: Network architecture for activity recognition task.

#	Layer(type)	Details
1	Linear	number of neurons: 100
2	ReLU	-
3	Linear	number of neurons: 6
4	Softmax	-

– **Digit Classification on MNIST Dataset:** The MNIST dataset consists of a large collection of 28×28 pixel grayscale images of handwritten digits (0 through 9), along with their corresponding labels. The dataset contains a total of 60000 images, among which 50000 images are used for training, and the remaining 10000 are used for testing. We use the pathological non-iid data distribution among the workers. For this, only two out of the ten labels are available at each worker. To achieve this, the data is first sorted by digit label, and then divided into 200 shards of size 300. We then assign each of the 10 clients 20 shards. A 5-layer neural network with three convolution layers and two linear connected layers is used for this task. The detailed description is given in Table. C.2. Cross-entropy loss is used for training the network. We used Adam optimizer, batch size of 64 and a learning rate of 0.02 for evaluating the algorithms.
– **Spam Filtering**: This is a binary classification problem, where the objective is to determine whether an email message is spam or not. The Spambase dataset contains a total of 4601 instances, with each instance representing a single email message, with an appropriate label of 0 (not spam) or 1 (spam). Each instance is described by a feature vector of length 57, which contains numerical attributes derived from

Table C.2: Network architecture for digital classification task.

#	Layer(type)	Details
1	Conv2d	output channels: 32, kernel size:3, stride:1
2	ReLU	-
3	MaxPool2d	kernel size:2
4	Conv2d	output channels: 64, kernel size:3, stride:1
5	ReLU	-
6	MaxPool2d	kernel size:2
7	Conv2d	output channels: 64, kernel size:3, stride:1
8	ReLU	-
9	MaxPool2d	kernel size:2
10	Linear	number of neurons: 128
11	ReLU	-
12	Linear	number of neurons: 10

the content of the messages. The data is divided among the 10 workers in a non-uniform fashion in the sense that the distribution varies across the workers, even though all the local datasets contain data samples belonging to both classes. We used a connected neural network as the model which is described in Table. C.3. Batch-size used here is 20, and the learning rate is 0.01. The network is trained using cross-entropy loss.

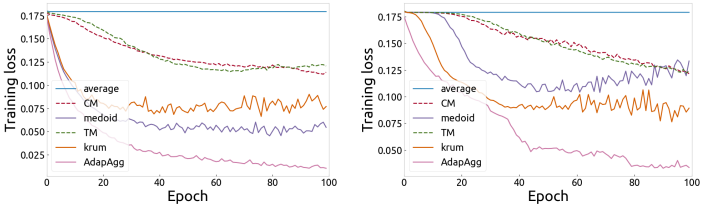
Table C.3: Network architecture for spam filtering task.

#	Layer(type)	Details
1	Linear	number of neurons: 20
2	ReLU	-
3	Linear	number of neurons: 2
4	Softmax	-

D Additional Empirical Results and Discussion

In Section 5, the algorithms are evaluated based on the test accuracy of the worst performing worker. Illustrated in Fig. 1, when 13 workers act adversarially during activity recognition task, the proposed adaptive aggregation method surpasses all other baseline methods across all attack types. Specifically, under the ALIE attack, all alternative methods experience complete failure. Likewise, in digit classification, as depicted in Fig. 2, the adaptive aggregation consistently excels under each attack type, unlike other methods which exhibit inconsistency. A parallel trend is observed in the spam detection task, illustrated in Fig. 3, where the proposed method consistently outperforms all other baselines.

The performance of the training of a ML algorithm can be represented by plotting the training loss. Here we present the training performance by plotting the highest training loss of the normal workers at each epoch. This is demonstrated in Fig.D.1, Fig.D.2, and Fig. D.3 for the activity recognition, digit classification, and spam detection tasks, respectively. One can observe a correlation between the trends in these plots and those depicted in the test accuracy plots.

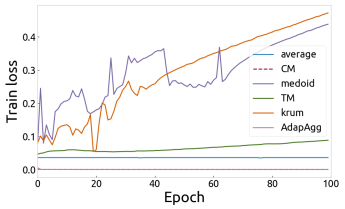


(a) sign-flip attack (b) arbitrary Byzantine attack

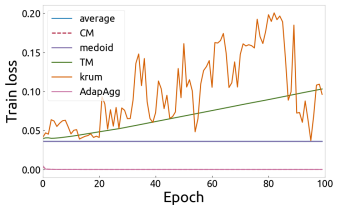
(c) Fall-of-empire attack (d) A-little-is-enough attack

Fig. D.1: Training loss for activity recognition task with 13 adversarial workers.

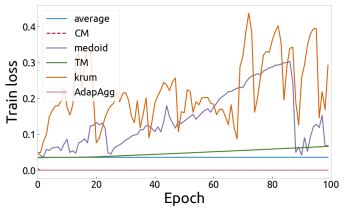
We further simulated an additional attack scenario for the activity recognition task, involving 6 adversarial workers. The training loss and testing accuracy achieved under various attack types are depicted in Fig.D.4 and Fig.D.5, respectively. It is evident that even with fewer adversarial workers, the proposed adaptive aggregation technique consistently outperforms all other baselines.



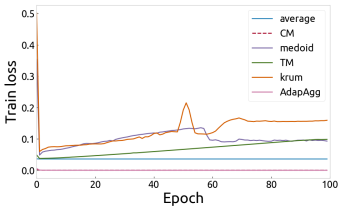
(a) sign-flip attack



(b) arbitrary Byzantine attack

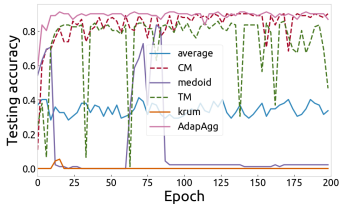


(c) Fall-of-empire attack

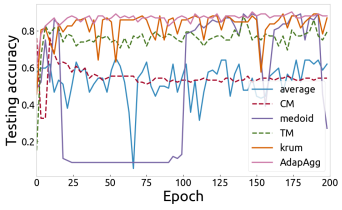


(d) A-little-is-enough attack

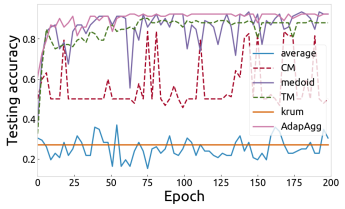
Fig. D.2: Training loss for digit classification task with 3 adversarial workers.



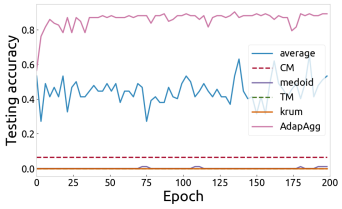
(a) sign-flip attack



(b) arbitrary Byzantine attack

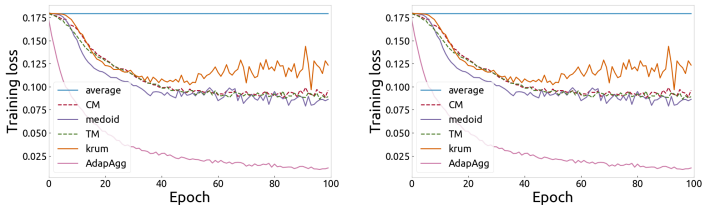


(c) Fall-of-empire attack



(d) A-little-is-enough attack

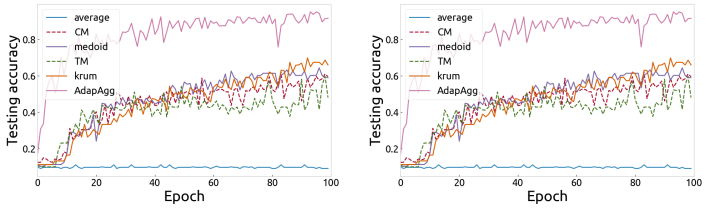
Fig. D.3: Training loss for spam detection task with 3 adversarial workers.



(a) sign-flip attack (b) arbitrary Byzantine attack

(c) Fall-of-empire attack (d) A-little-is-enough attack

Fig. D.4: Training loss for activity recognition task with 6 adversarial workers.

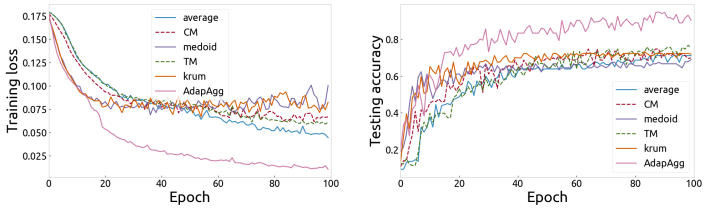


(a) sign-flip attack (b) arbitrary Byzantine attack

(c) Fall-of-empire attack (d) A-little-is-enough attack

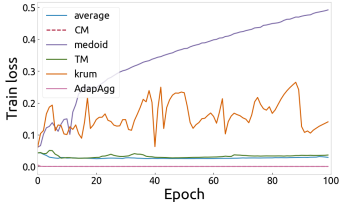
Fig. D.5: Test accuracy for activity recognition task with 6 adversarial workers.

No-attack scenario While aggregation methods are primarily designed to withstand various types of attacks, it's equally important that they maintain good performance under normal conditions when there is no attack. This ensures that the system operates effectively and efficiently in typical usage scenarios, contributing to its overall reliability and usability. Therefore, to illustrate the performance of the aggregation methods in the absence of attacks, we display the test accuracy and training loss for all three tasks when no worker acts adversarially. These plots are presented in Fig.D.6. The plots indicate that the proposed method adaptively aggregates parameters, ensuring sustained performance even in the absence of attacks. It performs comparably, if not better, than the other baselines.

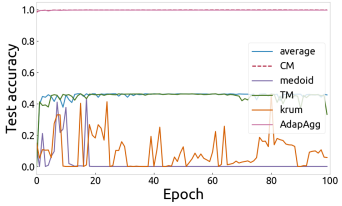


(a) Training loss

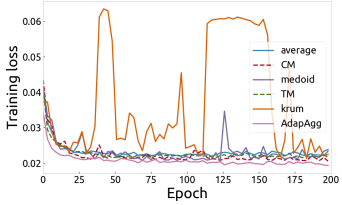
(b) Test accuracy



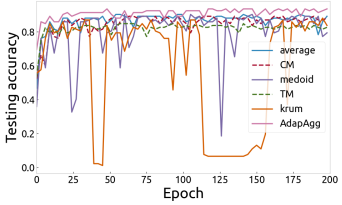
(c) Training loss



(d) Test accuracy



(e) Training loss



(f) Test accuracy

Fig. D.6: Training loss and test accuracy under no attack scenario: (a-b) Activity recognition task, (c-d) Digit classification task, (e-f) Spam detection task.

References

1. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM review **60**(2), 223–311 (2018)