Supplementary Materials: Near-Optimal Resilient Aggregation Rules for Distributed Learning Using 1-Center and 1-Mean Clustering with Outliers

Yuhao Yi[1*], **Ronghui You**[2*], **Hong Liu**[1], **Changxin Liu**[3], **Yuan Wang**[4†], **Jiancheng Lv**[1†]

[1]College of Computer Science, Sichuan University
[2]School of Statistics and Data Science, Nankai University
[3]School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology
[4]School of Robotics, Hunan University
yuhaoyi@scu.edu.cn, nijuyoumo@gmail.com, changxin@kth.se, yuanw@hnu.edu.cn, lvjiancheng@scu.edu.cn

Organization of the Appendix

Appendix A provides a proof of the 2-approximation of the MeanwO algroithm. Appendix B contains a proof of Theorem 2, which provides robust guarantees for CenterwO and MeanwO. Appendix C explains details for the Inner and Outer algorithm in the 2PRASHB framework. Appendix D contains details about the running environment of the experiments, description of the adversarial attacks, all experimental results, and an ablation study of the two-phase framework.

A Proof of the 2-Approximation of MeanwO

We recall the following proposition.

Propostion 1 (Propostion 3 in (Banerjee, Ostrovsky, and Rabani 2021))**.** *Let X be a finite set of points in $\mathbb{R}^d$, then*

$$\min_{y\in X}\sum_{x\in X}\|x-y\|^2 \le 2\sum_{x\in X}\|x-\mathrm{cm}(X)\|^2 . \tag{6}$$

Proof of Lemma 1. We recall the notation $K(c)$, which represents the set of $(n-f)$ data points in X closest to the point c (breaking ties arbitrarily). Let $\widehat{c}$ be the optimal cluster center for the problem of 1-mean clustering with outliers, and $\widehat{S} := \{i \mid x_i \in K(\widehat{c})\}$. We denote by

$$j \in \arg\min_{j\in\widehat{S}}\sum_{i\in\widehat{S}}\|x_i-x_j\|^2 ,$$

then by Proposition 1,

$$\sum_{i\in\widehat{S}}\|x_i-x_j\|^2 \le 2\sum_{i\in\widehat{S}}\left\|x_i-\overline{x}_{\widehat{S}}\right\|^2 . \tag{7}$$

By the definition of $K(x_j)$ we attain

$$\sum_{i\in K(x_j)}\|x_i-x_j\|^2 \le \sum_{i\in\widehat{S}}\|x_i-x_j\|^2 . \tag{8}$$

The MeanwO algorithm chooses a data point x_k as the cluster center (medoid) such that

$$x_k \in \arg\min_{x_i\in X}\sum_{\ell\in K(x_j)}\|x_\ell-x_i\|^2 . \tag{9}$$

Then we attain

$$\begin{aligned}
&\sum_{i\in K(x_k)}\|x_i-x_k\|^2 \\
&\le \sum_{i\in K(x_j)}\|x_i-x_j\|^2 && \text{by applying (9)} \\
&\le \sum_{i\in\widehat{S}}\|x_i-x_j\|^2 && \text{by applying (8)} \\
&\le 2\sum_{i\in\widehat{S}}\left\|x_i-\overline{x}_{\widehat{S}}\right\|^2 . && (10)
\end{aligned}$$

[*]These authors contributed equally.
[†]Corresponding author.

The last inequality is due to (7). Note that (10) is exactly the definition of 2-approximation. □

B Proof of Robust Properties of CenterwO and MeanwO

Proof of Theorem 2. Given a set of n vectors X in $\mathbb{R}^d$, let $B(c^*, r^*)$ be an optimal solution to the problem of 1-center clustering with outliers. The set of vectors covered by $B(c^*, r^*)$ is denoted as S^*. The CenterwO finds a ball $B(c', r')$ which contains $n-f$ vectors in X satisfying $r' \leq 2r^*$. We denote by S' the set of point indices based on which the returned vector is calculated, i.e. $S' := \{i \mid x_i \in K(c')\}$. For any subset $S \subseteq [n]$ we denote by $B(c_S, r_S)$ the minimum enclosing ball of the set of vectors $\{x_i\}_{i\in S}$. Then we attain

$$\begin{aligned}
&\|\mathrm{CenterwO}(X) - \overline{x}_S\| \\
&= \left\| \frac{1}{n-f} \sum_{i\in S'} x_i - \frac{1}{n-f} \sum_{i\in S} x_i \right\| \\
&= \left\| \frac{1}{n-f} \sum_{i\in S'\setminus S} x_i - \frac{1}{n-f} \sum_{i\in S\setminus S'} x_i \right\| \\
&\leq \frac{f}{n-f} \cdot \max_{i\in S\setminus S', j\in S'\setminus S} \|x_i - x_j\| \,. \qquad (11)
\end{aligned}$$

The inequality follows by the fact that $|S\setminus S'| = |S'\setminus S| = |S \cup S'| - |S| = |S \cup S'| - |S'| \leq n - (n-f) = f$.

We assume $f < n/2$. Since $|S \cup S'| \leq n$, then $|S \cap S'| = |S| + |S'| - |S \cup S'| \geq 2(n-f) - n = n - 2f > 0$, indicating $|S \cap S'| \neq \varnothing$. Let $k \in S \cap S'$. By applying the triangle inequality, we arrive at

$$\begin{aligned}
&\max_{i\in S\setminus S', j\in S'\setminus S} \|x_i - x_j\| \\
&\leq \max_{i,k\in S} \|x_i - x_k\| + \max_{i,k\in S'} \|x_j - x_k\| \,. \qquad (12)
\end{aligned}$$

Because the CenteroW algorithm provides a 2-approximation to the optimum, we attain $r^* \leq r' \leq 2r^*$.

The Jung's theorem (Jung 1901) states that for a set of points X in a Euclidean space, let $B(c, r)$ be the minimum enclosing ball of X,

$$r \leq \left(\frac{d}{2d+1}\right)^{\frac{1}{2}} \cdot \max_{x_i, x_j \in X} \|x_i - x_j\| \,.$$

Then we attain

$$\sqrt{2}r \leq \max_{x_i, x_j \in X} \|x_i - x_j\| \leq 2r \,, \qquad (13)$$

in any d-dimensional space. The first inequality is attributed to the fact that $(2d+1)/d > 2$. The second inequality is attained by applying the triangle inequality.

Then we arrive at

$$\begin{aligned}
&\max_{i\in S\setminus S', j\in S'\setminus S} \|x_i - x_j\| \\
&\leq \max_{i,j\in S} \|x_i - x_j\| + 2r' \qquad \text{by applying (13)} \\
&\leq \max_{i,j\in S} \|x_i - x_j\| + 4r^* \qquad \text{since } r' \leq 2r^* \\
&\leq \max_{i,j\in S} \|x_i - x_j\| + 4r_S \qquad \text{optimality of } r^* \\
&\leq \max_{i,j\in S} \|x_i - x_j\| + 2\sqrt{2} \cdot \max_{i,j\in S} \|x_i - x_j\|
\end{aligned}$$

The last inequality follows by (13). Therefore

$$\max_{i\in S\setminus S', j\in S'\setminus S} \|x_i - x_j\| \leq (2\sqrt{2}+1) \max_{i,j\in S} \|x_i - x_j\| \,. \qquad (14)$$

Substituting (14) into (11) yields the $(f, (2\sqrt{2}+1)f/(n-f))$-resilient averaging guarantee for the CenterwO algorithm.

Since the CenterwO algorithm is deterministic, by squaring both sides of the result for (f, λ)-resilient averaging aggregator, we attain the result for $(\delta_{\max}, \zeta)$-ARAgg.

Then we proof the result for (f,κ)-robustness. We recall the first equality of (40) in (Allouah et al. 2023b)

$$\left(1-\frac{|S'\backslash S|}{n-f}\right)^2\|\overline{x}_{S'}-\overline{x}_S\|^2=\left\|\frac{1}{n-f}\sum_{i\in S'\backslash S}(x_i-\overline{x}_{S'})-\frac{1}{n-f}\sum_{i\in S\backslash S'}(x_i-\overline{x}_S)\right\|^2. \tag{15}$$

From (15) we attain

$$\begin{aligned}
&\left(1-\frac{f}{n-f}\right)^2\|\overline{x}_{S'}-\overline{x}_S\|^2\\
&\leq\left\|\frac{1}{n-f}\sum_{i\in S'\backslash S}(x_i-\overline{x}_{S'})-\frac{1}{n-f}\sum_{i\in S\backslash S'}(x_i-\overline{x}_S)\right\|^2\\
&\leq\frac{1}{(n-f)^2}\left(\sum_{i\in S'\backslash S}\|x_i-\overline{x}_{S'}\|+\sum_{i\in S\backslash S'}\|x_i-\overline{x}_S\|\right)^2\\
&\leq\frac{2f}{(n-f)^2}\left(\sum_{i\in S'\backslash S}\|x_i-\overline{x}_{S'}\|^2+\sum_{i\in S\backslash S'}\|x_i-\overline{x}_S\|^2\right)
\end{aligned}\tag{16}$$

The first inequality follows from (40) in (Allouah et al. 2023b) and $|S'\backslash S|\leq f$; the second inequality is attained by using a series of triangle inequalities; the third inequality follows by the Cauchy-Schwarz inequality, and the fact that $|S\backslash S'|=|S'\backslash S|\leq f$.

We observe that

$$\begin{aligned}
&\sum_{i\in S'\backslash S}\|x_i-\overline{x}_{S'}\|^2\\
&\leq f\cdot(r')^2 && \text{by applying (13)}\\
&\leq 4f\cdot(r^*)^2 && \text{by } r'\leq 2r^*\\
&\leq 4f\cdot(r_S)^2 && r^* \text{ is optimal}\\
&\leq 4f\max_{i\in S}\|x_i-\overline{x}_S\|^2 && (17)\\
&\leq 4f\sum_{i\in S}\|x_i-\overline{x}_S\|^2\,, && (18)
\end{aligned}$$

where the fourth inequality holds because r_S is the radius of the minimum ball enclosing $\{x_i\}_{i\in S}$. Therefore

$$\kappa=\left(1+\frac{f}{n-2f}\right)^2\cdot\frac{8f^2+2f}{n-f}\,,$$

which concludes the proof for properties of the CenterwO aggregation rule.

Now we investigate the (f,ξ)-robust averaging property of the CenterwO algorithm. We recall (40) in (Allouah et al. 2023b):

$$\left(1-\frac{|S'\backslash S|}{n-f}\right)^2\|\overline{x}'_S-\overline{x}_S\|^2\leq\frac{2f}{(n-f)^2}\sup_{\|v\|\leq1}\sum_{i\in S'\backslash S}|\langle v,x_i-\overline{x}_{S'}\rangle|^2+\frac{2f}{(n-f)^2}\sup_{\|v\|\leq1}\sum_{i\in S\backslash S'}|\langle v,x_i-\overline{x}_S\rangle|^2\,. \tag{19}$$

We analyze the supremum of the first term in (19):

$$\begin{aligned}
&\sup_{\|v\|\leq1}\sum_{i\in S'\backslash S}|\langle v,x_i-\overline{x}_{S'}\rangle|^2\\
&\leq\sum_{i\in S'\backslash S}\sup_{\|v_i\|\leq1}|\langle v_i,x_i-\overline{x}_{S'}\rangle|^2\\
&=\sum_{i\in S'\backslash S}\|x_i-\overline{x}_{S'}\|^2\\
&\leq 4f\max_{i\in S}\|x_i-\overline{x}_S\|^2 \qquad \text{by applying (17)}\\
&=4f\max_{i\in S}\sup_{\|v_i\|\leq1}|\langle v_i,x_i-\overline{x}_S\rangle|^2\,, && (20)
\end{aligned}$$

where the two equalities are attained when the vectors v_i and $\overline{x}_{S'}$ (resp. v_i and $\overline{x}_S$) have the same direction. We note that for all $i \in S$, there are

$$\begin{aligned}\sup_{\|v_i\|\le 1} |\langle v_i, x_i - \overline{x}_S\rangle|^2 &= \sup_{\|v_i\|\le 1} v_i^\top (x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top v_i \\ &\le \sup_{\|v_i\|\le 1} v_i^\top \left(M_S\right) v_i \,. \end{aligned} \tag{21}$$

The inequality follows by the Courant-Fischer min-max theorem since a positive semi-definite matrix is added to the rank-one semi-definite matrix in the middle. By combining (19), (20), and (21) we attain

$$\left(1 - \frac{|S'\backslash S|}{n-f}\right)^2 \left\|\overline{x}'_S - \overline{x}_S\right\|^2 \le \frac{8f^2 + 2f}{(n-f)^2} \lambda_{\max}(M_S)\,,$$

where the Courant-Fischer min-max theorem is also applied to the second term in (19). Note that $|S'\backslash S| \le f$, we attain the result for (f, ξ)-robust averaging of the CenterwO algorithm.

Next we prove the properties of the MeanwO aggregation rule. Let $\widehat{c}$ be the optimal cluster center for the problem of 1-mean clustering with outliers, and let $\widehat{r} := \max_{x\in K(\widehat{c})} \|x - \widehat{c}\|$. The MeanwO Algorithm finds a cluster center (medoid) $\widetilde{c}$ from the data points. We let $\widetilde{r} := \max_{x\in K(\widetilde{c})} \|x - \widetilde{c}\|$. We further denote by $\widetilde{S}$ the set of vector indices based on which the returned vector is calculated, i.e. $\widetilde{S} := \{i \mid x_i \in K(\widetilde{c})\}$. Similarly, let $\widehat{S} := \{i \mid x_i \in K(\widehat{c})\}$.

For the (f, κ)-robustness, we show that

$$\begin{aligned}&\left(1 - \frac{f}{n-f}\right)^2 \left\|\overline{x}_{\widetilde{S}} - \overline{x}_S\right\|^2 \\ &\le \frac{2f}{(n-f)^2} \left(\sum_{i\in\widetilde{S}\backslash S} \left\|x_i - \overline{x}_{\widetilde{S}}\right\|^2 + \sum_{i\in S\backslash\widetilde{S}} \|x_i - \overline{x}_S\|^2\right) \\ &\le \frac{2f}{(n-f)^2} \left(\sum_{i\in\widetilde{S}} \left\|x_i - \overline{x}_{\widetilde{S}}\right\|^2 + \sum_{i\in S} \|x_i - \overline{x}_S\|^2\right) \\ &\le \frac{2f}{(n-f)^2} \left(\sum_{i\in\widetilde{S}} \|x_i - \widetilde{c}\|^2 + \sum_{i\in S} \|x_i - \overline{x}_S\|^2\right) \\ &\le \frac{2f}{(n-f)^2} \left(2\cdot\sum_{i\in\widehat{S}} \left\|x_i - \overline{x}_{\widehat{S}}\right\|^2 + \sum_{i\in S} \|x_i - \overline{x}_S\|^2\right) \\ &\le \frac{6f}{(n-f)^2} \sum_{i\in S} \|x_i - \overline{x}_S\|^2 \,,\end{aligned}$$

where the second inequality is due to (16); the third inequality follows by the fact that the centroid (or center of mass) $\overline{x}_{\widetilde{S}}$ minimizes the sum of squared distances to the cluster center; the fourth inequality follows by the 2-approximation guarantee of the MeanwO algorithm; the last inequality is attained since $\widehat{S}$ minimizes $\sum_{i\in S} \|x_i - \overline{x}_S\|$.

From Propostion 8 and Proposition 9 in (Allouah et al. 2023a) we arrive at the results for (f, λ)-resilience and (δ_{max}, ζ)-agnostic robustness of the MeanwO algorithm.

The result for the (f, ξ)-robust averaging property follows straightforwardly from the fact that $\sum_{i\in S} \|x_i - \overline{x}_S\|^2$ is the trace (sum of eigenvalues) of the matrix M_S, which is positive semi-definite and has at most $\min\{n-f, d\}$ non-zero eigenvalues. □

C Details for the Two-Phase Algorithm

Now we describe the Inner and Outer algorithm used in Algorithm 3. The Inner algorithm should be resilient to siege attacks. In our realization, the Inner algorithm directly calls the 1-Center/Mean Clustering with Outliers algorithm shown in Algorithm 2. The Outer algorithm should be designed to defend against sneak attacks. In our implementation, the Outer algorithm returns the average of the $(n - f)$ vectors *outside* of an approximate minimum 1-center/mean cluster with f in-cluster vectors. The pseudo code of the Outer algorithm is shown in Algorithm 4.

D Details for Experiments

D.1 Simulation Environment

Both the server and workers are simulated on a cloud virtual machine equipped with a 32-core Intel Xeon Gold 6278@2.6G CPU, 128GB of memory, and a 16GB Quadro RTX 5000 GPU.

Algorithm 4: Outer

Input: a set of n vectors X in $\mathbb{R}^d$, and an integer $f < \frac{n}{2}$.
Output: the mass center of the $n - f$ vectors outside of an approximate minimum 1-center/mean cluster with f in-cluster vectors.

for $i = 1, \ldots, n$ **do**
  Find $N(x_i)$, the f closest vectors in X to the vector x_i (including x_i), breaking ties arbitrarily;
  Let $\text{dist}_i = \text{cost}(x_i, N(x_i))$; /*see (5) for definition of 1-center/mean cost*/
end for
Let $j = \arg\min_i \text{dist}_i$;
return $\frac{1}{n-f} \sum_{x \in (X \setminus N(x_j))} x$;

D.2 Attacks

Here we give the detailed description of adversarial attacks in the experiments.

1) **LF**: the label flipping attack, where the labels of the images are replaced by labels described by a deterministic permutation in corrupted workers;

2) **SF**: the sign flipping attack, where each corrupted worker sends the negative of its true update vector;

3) **Gauss**: the Gauss attack, where random Gaussian vectors replace the update vectors with the same vector norm;

4) **Omn**: the omniscient attack (Blanchard et al. 2017), where all the corrupted workers send the average of all update vectors without corruptions minus the average vector of corrupted workers multiplied by $2n/f$;

5) **Empire**: the fall of empire attack (Xie, Koyejo, and Gupta 2020), where the update vectors or corrupted workers are set to the average of the update vectors without corruptions multiplied by -0.1;

6) **SV**: the scaled variance (Baruch, Baruch, and Goldberg 2019; Allen-Zhu et al. 2021) attack, where the corrupted workers set their update vectors to the mean of all workers, shifted by 20 times the standard deviation in each coordinate.

7) **PGA**: In particular, we customize more sophisticated attacks, the PGA algorithm, to attack various aggregation rules (Shejwalkar et al. 2022). PGA algorithm leverages STAT-OPT attacks to generate a malicious update, and all Byzantine workers send the same malicious update to the server. As the proposed 2PRASHB algorithm could easily filter out all malicious updates when they are at the same position in multi-dimensional space, we eliminate the data-based stochastic gradient ascent (SGA) in PGA to improve the running efficiency. STAT-OPT computes the average updates ∇^b from benign workers, and computes a static malicious direction $w = -\text{sign}(\nabla^b)$. Moreover, STAT-OPT attacks tailor themselves to the target aggregation rule (Agg) by searching a suboptimal γ so that the final malicious update $\nabla' = \nabla^b - \gamma * w$ could circumvent the target Agg.

In order to be consistent with the experimental settings of the paper by Shejwalkar et al. (2022), we only conduct experiments on extremely non-iid dataset drawn from FEMNIST. Corresponding results are presented in Table 3, clearly showing the robustness of Cent2P and Mean2P.

D.3 Architecture of Client Model

For the image classification task on FEMNIST dataset, we construct a Convolutional Neural Network (CNN) consisting of two convolutional layers. Each convolutional layer has a kernel size of (5×5), and we use 32 and 64 kernels, respectively. After each convolutional layer, we apply a ReLU non-linear activation function followed by a Max-pooling layer with a (2×2) kernel size. We incorporate a fully-connected layer for classification. To train the model, we employ the Cross Entropy loss function and the Stochastic Gradient Descent (SGD) optimizer.

D.4 Additional Experimental Evaluations

Table 2 and Table 3 show the full results on FEMNIST dataset for 3 levels of adversarial rates: 0.1, 0.2, and 0.4. Each cell shows the average and standard deviation of testing accuracy in 5 simulations. The results clearly show the consistent resilience of our methods.

CIFAR-10: To further show the robustness of the proposed aggregation frameworks, we also run experiments on the CIFAR-10 dataset, another typical image classification benchmark. The CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class (Krizhevsky 2009). We use a small dataset of 35 clients uniformly sampled from the CIFAR-10 dataset, and each client contains 300 train samples and 60 test samples. As presented in Table 4, the proposed algorithms show consistent advantages against all baselines.

Figure 3 shows the performance comparison on homogeneous datasets during the training processes. Error bars show the standard deviations. Our methods are among the ones with best performance under all attacks. For homogeneous datasets, CenterwO and MeanwO are the only rules which resist the Omn attack. We have discussed the results for heterogeneous datasets in the main paper. Figure 4 shows the results with error bars.

Table 2: Performance comparison on the uniform sampling (homogeneous) datasets.

Rate	Aggregation	LF	SF	Gauss	Omn	Empire	SV	Worst
0.1	Avg	0.77 ± 0.02	0.73 ± 0.06	0.81 ± 0.01	0.01 ± 0.00	0.81 ± 0.01	0.75 ± 0.03	0.01
	GM	0.80 ± 0.01	0.79 ± 0.00	0.80 ± 0.01	0.19 ± 0.27	0.80 ± 0.01	0.66 ± 0.07	0.19
	CClip	0.70 ± 0.01	0.68 ± 0.01	0.70 ± 0.01	0.60 ± 0.04	0.67 ± 0.01	0.68 ± 0.03	0.60
	CWM	0.60 ± 0.02	0.60 ± 0.02	0.55 ± 0.06	0.32 ± 0.14	0.55 ± 0.04	0.56 ± 0.03	0.32
	CWTM	0.63 ± 0.01	0.60 ± 0.03	0.66 ± 0.01	0.23 ± 0.08	0.57 ± 0.02	0.28 ± 0.05	0.23
	Krum	0.51 ± 0.02	0.51 ± 0.02	0.50 ± 0.02	0.52 ± 0.02	0.35 ± 0.06	0.05 ± 0.00	0.05
	Cent2P	0.78 ± 0.02	0.78 ± 0.01	0.80 ± 0.01	0.80 ± 0.00	0.80 ± 0.01	0.76 ± 0.01	0.76
	Mean2P	0.78 ± 0.02	0.79 ± 0.01	0.81 ± 0.01	0.79 ± 0.01	0.80 ± 0.01	0.76 ± 0.01	0.76
0.2	Avg	0.74 ± 0.02	0.53 ± 0.25	0.80 ± 0.01	0.00 ± 0.00	0.81 ± 0.00	0.62 ± 0.02	0.00
	GM	0.78 ± 0.01	0.76 ± 0.01	0.80 ± 0.00	0.05 ± 0.02	0.79 ± 0.01	0.41 ± 0.09	0.05
	CClip	0.69 ± 0.01	0.64 ± 0.02	0.68 ± 0.01	0.49 ± 0.05	0.57 ± 0.04	0.58 ± 0.02	0.49
	CWM	0.60 ± 0.01	0.56 ± 0.03	0.59 ± 0.04	0.05 ± 0.02	0.49 ± 0.06	0.44 ± 0.07	0.05
	CWTM	0.56 ± 0.03	0.55 ± 0.03	0.56 ± 0.04	0.02 ± 0.01	0.38 ± 0.03	0.10 ± 0.01	0.02
	Krum	0.53 ± 0.02	0.44 ± 0.05	0.49 ± 0.02	0.52 ± 0.01	0.30 ± 0.05	0.05 ± 0.00	0.05
	Cent2P	0.78 ± 0.02	0.76 ± 0.01	0.79 ± 0.01	0.80 ± 0.01	0.79 ± 0.00	0.73 ± 0.02	0.73
	Mean2P	0.77 ± 0.01	0.76 ± 0.02	0.80 ± 0.01	0.79 ± 0.01	0.79 ± 0.01	0.75 ± 0.01	0.75
0.4	Avg	0.56 ± 0.03	0.04 ± 0.01	0.79 ± 0.00	0.00 ± 0.00	0.79 ± 0.01	0.32 ± 0.04	0.00
	GM	0.64 ± 0.03	0.46 ± 0.16	0.79 ± 0.00	0.00 ± 0.00	0.74 ± 0.00	0.07 ± 0.02	0.00
	CClip	0.58 ± 0.04	0.45 ± 0.06	0.64 ± 0.01	0.20 ± 0.02	0.26 ± 0.04	0.26 ± 0.06	0.20
	CWM	0.50 ± 0.05	0.45 ± 0.05	0.54 ± 0.04	0.02 ± 0.01	0.07 ± 0.02	0.05 ± 0.00	0.02
	CWTM	0.51 ± 0.02	0.39 ± 0.03	0.55 ± 0.04	0.02 ± 0.01	0.05 ± 0.01	0.05 ± 0.00	0.02
	Krum	0.53 ± 0.03	0.36 ± 0.04	0.47 ± 0.02	0.10 ± 0.06	0.01 ± 0.01	0.05 ± 0.00	0.01
	Cent2P	0.74 ± 0.02	0.62 ± 0.03	0.76 ± 0.00	0.79 ± 0.00	0.74 ± 0.01	0.76 ± 0.02	0.62
	Mean2P	0.73 ± 0.03	0.61 ± 0.01	0.76 ± 0.01	0.79 ± 0.00	0.73 ± 0.01	0.75 ± 0.02	0.61

D.5 Ablation Experiments

Table 5 presents a performance comparison between RASHB (Cent1P/Mean1P) and 2PRASHB (Cent2P/Mean2P) on the uniform sampling datasets, with an adversarial rate of 0.2. Across LF, SF, Gauss, and Empire attacks, both RASHB and 2PRASHB yield comparable outcomes. However, when subjected to Omn and SV attacks, 2PRASHB significantly outperforms PRASHB. Particularly noteworthy is the accuracy achieved under Omn attacks, with Cent2P and Mean2P achieving accuracies of 0.80 and 0.79, respectively, whereas Cent1P and Mean1P only attain 0.12 and 0.01. These results effectively showcase the superiority of 2PRASHB.

References

Allen-Zhu, Z.; Ebrahimianghazani, F.; Li, J.; and Alistarh, D. 2021. Byzantine-Resilient Non-Convex Stochastic Gradient Descent. In *International Conference on Learning Representations (ICLR'21)*.

Allouah, Y.; Farhadkhani, S.; Guerraoui, R.; Gupta, N.; Pinot, R.; and Stephan, J. 2023a. Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS'23)*, volume 206 of *Proceedings of Machine Learning Research*, 1232–1300. PMLR.

Allouah, Y.; Guerraoui, R.; Gupta, N.; Pinot, R.; and Stephan, J. 2023b. On the Privacy-Robustness-Utility Trilemma in Distributed Learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, volume 202 of *Proceedings of Machine Learning Research*. PMLR.

Banerjee, S.; Ostrovsky, R.; and Rabani, Y. 2021. Min-Sum Clustering (With Outliers). In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, (APPROX/RANDOM'21)*, volume 207 of *LIPIcs*, 16:1–16:16. Seattle, Washington, USA (Virtual Conference): Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Baruch, G.; Baruch, M.; and Goldberg, Y. 2019. A Little Is Enough: Circumventing Defenses For Distributed Learning. In *Advances in Neural Information Processing Systems (NeurIPS'19)*, volume 32. Curran Associates, Inc.

Blanchard, P.; El Mhamdi, E. M.; Guerraoui, R.; and Stainer, J. 2017. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Advances in Neural Information Processing Systems (NeurIPS'17)*, volume 30. Curran Associates, Inc.

Jung, H. 1901. Ueber die kleinste Kugel, die eine räumliche Figur einschliesst. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1901(123): 241–257.

Table 3: Performance comparison on the nonuniform sampling (heterogeneous) datasets.

Rate	Aggregation	LF	SF	Gauss	Omn	Empire	SV	PGA	Worst
0.1	Avg	0.81 ± 0.01	0.30 ± 0.24	0.88 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.82 ± 0.01	0.03 ± 0.04	0.00
	GM	0.83 ± 0.01	0.60 ± 0.22	0.88 ± 0.01	0.09 ± 0.04	0.86 ± 0.01	0.76 ± 0.02	0.00 ± 0.00	0.00
	CClip	0.72 ± 0.01	0.51 ± 0.11	0.74 ± 0.01	0.50 ± 0.03	0.68 ± 0.01	0.68 ± 0.01	0.43 ± 0.02	0.43
	CWM	0.58 ± 0.05	0.55 ± 0.08	0.59 ± 0.06	0.11 ± 0.04	0.23 ± 0.06	0.52 ± 0.03	0.12 ± 0.03	0.11
	CWTM	0.61 ± 0.03	0.57 ± 0.06	0.70 ± 0.02	0.11 ± 0.00	0.25 ± 0.04	0.18 ± 0.05	0.04 ± 0.03	0.04
	Krum	0.29 ± 0.03	0.25 ± 0.04	0.28 ± 0.02	0.30 ± 0.05	0.15 ± 0.07	0.00 ± 0.00	0.19 ± 0.05	0.00
	Cent2P	0.84 ± 0.01	0.83 ± 0.01	0.87 ± 0.01	0.88 ± 0.01	0.87 ± 0.01	0.78 ± 0.03	0.87 ± 0.00	0.78
	Mean2P	0.82 ± 0.02	0.83 ± 0.02	0.87 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	0.83 ± 0.02	0.87 ± 0.01	0.82
0.2	Avg	0.72 ± 0.01	0.10 ± 0.02	0.87 ± 0.02	0.00 ± 0.00	0.87 ± 0.01	0.73 ± 0.03	0.03 ± 0.04	0.00
	GM	0.76 ± 0.05	0.10 ± 0.03	0.86 ± 0.01	0.03 ± 0.02	0.84 ± 0.02	0.39 ± 0.03	0.00 ± 0.00	0.00
	CClip	0.66 ± 0.04	0.25 ± 0.14	0.72 ± 0.02	0.34 ± 0.02	0.49 ± 0.03	0.44 ± 0.04	0.39 ± 0.01	0.25
	CWM	0.49 ± 0.03	0.50 ± 0.10	0.60 ± 0.05	0.02 ± 0.02	0.04 ± 0.01	0.33 ± 0.04	0.01 ± 0.01	0.01
	CWTM	0.49 ± 0.04	0.48 ± 0.06	0.63 ± 0.04	0.02 ± 0.01	0.06 ± 0.02	0.04 ± 0.01	0.04 ± 0.03	0.02
	Krum	0.27 ± 0.04	0.22 ± 0.03	0.26 ± 0.02	0.26 ± 0.03	0.16 ± 0.07	0.01 ± 0.01	0.18 ± 0.03	0.01
	Cent2P	0.75 ± 0.07	0.69 ± 0.06	0.86 ± 0.01	0.86 ± 0.02	0.84 ± 0.02	0.78 ± 0.02	0.84 ± 0.01	0.69
	Mean2P	0.76 ± 0.05	0.72 ± 0.04	0.85 ± 0.01	0.85 ± 0.02	0.83 ± 0.03	0.80 ± 0.01	0.85 ± 0.01	0.72
0.4	Avg	0.57 ± 0.04	0.07 ± 0.02	0.79 ± 0.05	0.00 ± 0.00	0.79 ± 0.05	0.37 ± 0.03	0.03 ± 0.04	0.00
	GM	0.55 ± 0.06	0.07 ± 0.03	0.78 ± 0.05	0.00 ± 0.00	0.57 ± 0.08	0.06 ± 0.05	0.03 ± 0.04	0.00
	CClip	0.44 ± 0.05	0.23 ± 0.03	0.63 ± 0.05	0.12 ± 0.01	0.16 ± 0.02	0.19 ± 0.03	0.28 ± 0.08	0.12
	CWM	0.37 ± 0.05	0.23 ± 0.05	0.57 ± 0.05	0.00 ± 0.00	0.01 ± 0.02	0.02 ± 0.02	0.00 ± 0.01	0.00
	CWTM	0.37 ± 0.05	0.20 ± 0.05	0.57 ± 0.04	0.02 ± 0.02	0.03 ± 0.03	0.00 ± 0.00	0.05 ± 0.03	0.00
	Krum	0.14 ± 0.07	0.14 ± 0.06	0.21 ± 0.04	0.13 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.08 ± 0.02	0.00
	Cent2P	0.44 ± 0.10	0.30 ± 0.26	0.74 ± 0.06	0.76 ± 0.05	0.54 ± 0.17	0.77 ± 0.05	0.77 ± 0.03	0.30
	Mean2P	0.43 ± 0.12	0.35 ± 0.22	0.74 ± 0.06	0.75 ± 0.04	0.55 ± 0.16	0.78 ± 0.05	0.77 ± 0.03	0.35

Table 4: Performance comparison on CIFAR-10 dataset with the uniform sampling at an adversarial rate of 0.2.

Aggregation	LF	SF	Gauss	Omn	Empire	SV	Worst
FedAvg	0.58	0.33	0.57	0.10	0.56	0.25	0.10
GM	0.60	0.44	0.58	0.10	0.59	0.21	0.10
CClip	0.55	0.48	0.52	0.37	0.50	0.29	0.29
CWM	0.47	0.39	0.46	0.09	0.35	0.16	0.09
CWTM	0.49	0.43	0.48	0.12	0.42	0.18	0.12
Krum	0.18	0.15	0.21	0.21	0.10	0.10	0.10
Cent2P	0.59	0.49	0.56	0.57	0.57	0.46	0.46
Mean2P	0.59	0.47	0.56	0.56	0.57	0.49	0.47

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Shejwalkar, V.; Houmansadr, A.; Kairouz, P.; and Ramage, D. 2022. Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning. In *Proc. IEEE Symposium on Security and Privacy (SP'22)*, 1354–1371. IEEE.

Xie, C.; Koyejo, O.; and Gupta, I. 2020. Fall of Empires: Breaking Byzantine-tolerant SGD by Inner Product Manipulation. In Adams, R. P.; and Gogate, V., eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference (UAI'20)*, volume 115 of *Proceedings of Machine Learning Research*, 261–270. PMLR.

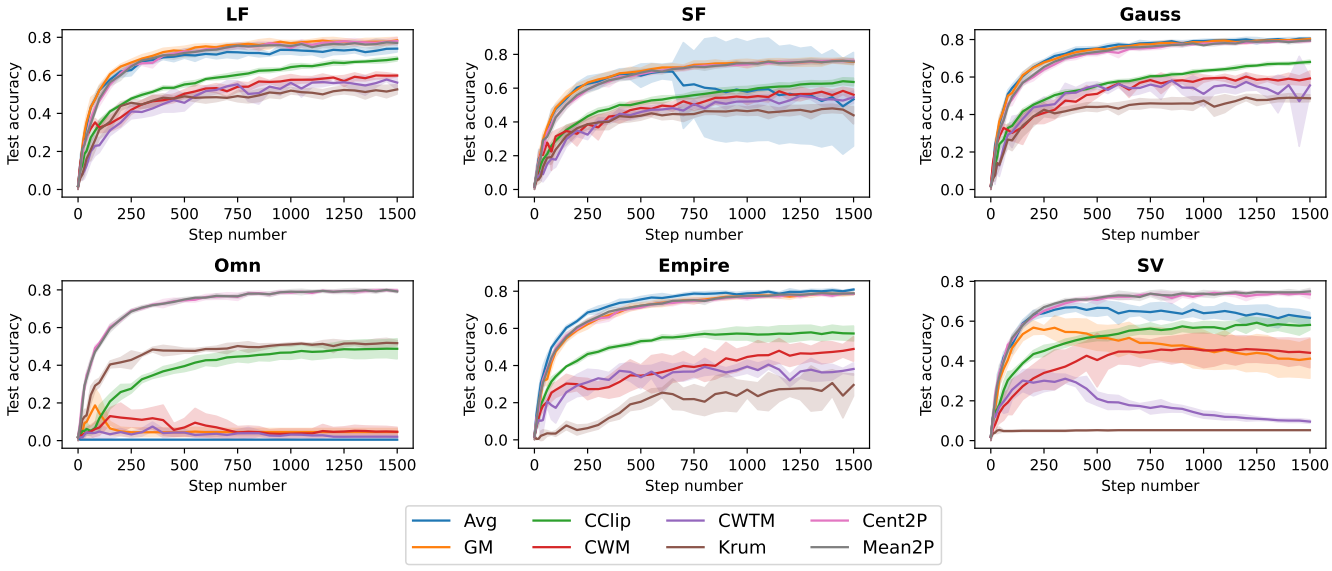

Figure 3: Performance comparison on the homogeneous datasets at an adversarial rate of 0.2.

Table 5: Performance comparison of RASHB (Cent1P/Mean1P) and 2PRASHB(Cent2P/Mean2P) on the uniform sampling datasets at an adversarial rate of 0.2.

Aggregation	LF	SF	Gauss	Omn	Empire	SV	Worst
Cent1P	0.78 ± 0.02	0.76 ± 0.02	0.79 ± 0.01	0.12 ± 0.14	0.75 ± 0.01	0.54 ± 0.01	0.12
Mean1P	0.77 ± 0.02	0.76 ± 0.01	0.78 ± 0.01	0.01 ± 0.00	0.77 ± 0.02	0.57 ± 0.02	0.01
Cent2P	0.78 ± 0.02	0.76 ± 0.01	0.79 ± 0.01	0.80 ± 0.01	0.79 ± 0.00	0.73 ± 0.02	0.73
Mean2P	0.77 ± 0.01	0.76 ± 0.02	0.80 ± 0.01	0.79 ± 0.01	0.79 ± 0.01	0.75 ± 0.01	0.75

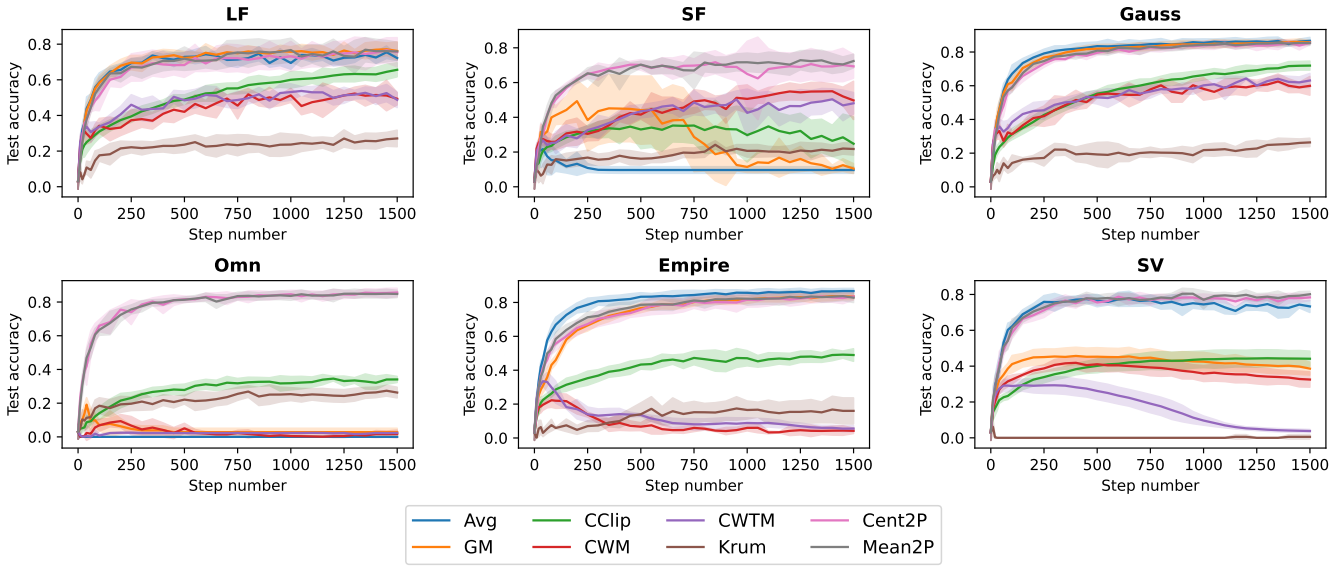

Figure 4: Performance comparison on the heterogeneous datasets at an adversarial rate of 0.2.