Appendix for *Noise-Aware Algorithm for Heterogeneous differentially private federated learning*

A EXPERIMENTAL SETUP

In this section we provide more experimental details that are deferred in the main paper.

A.1 DATASETS AND MODELS

MNIST and FMNIST datasets: We consider a distributed setting with 20 users. In order to create a non-i.i.d. dataset, we follow a similar procedure as in [1]: first we split the data from each class into several shards. Then, each user is randomly assigned a number of shards of data. For example, in some experiments, in order to guarantees that no user receives data from more than 6 classes, we split each class of MNIST/FMNIST into 12 shards (i.e., a total of 120 shards for the whole dataset), and each user is randomly assigned 6 shards of data. By considering 20 users, this procedure guarantees that no user receives data from more than 6 classes and the data distribution of each user is different from each other. The local datasets are balanced–all users have the same amount of training samples. The local data is split into train and test sets with percentage of 80%, and 20%, respectively. In this way, each user has 2400 data points for training, and 600 for testing. We use a simple 2-layer CNN model with ReLU activation, the detail of which can be found in Table 3. To update the local models at each user using its local data, unless otherwise is stated, we apply gradient descent.

Table 3: CNN model for classification on MNIST/FMNIST datasets

Layer	Output Shape	# of Trainable Parameters	Activation	Hyper-parameters
Input	$(1, 28, 28)$	0		
Conv2d	$(16, 28, 28)$	416	ReLU	kernel size =5; strides=$(1, 1)$
MaxPool2d	$(16, 14, 14)$	0		pool size=$(2, 2)$
Conv2d	$(32, 14, 14)$	12,832	ReLU	kernel size =5; strides=$(1, 1)$
MaxPool2d	$(32, 7, 7)$	0		pool size=$(2, 2)$
Flatten	1568	0		
Dense	10	15,690	ReLU	
Total		28,938		

CIFAR-10 dataset: We consider a distributed setting with 20 users, and split the 50,000 training samples and the 10,000 test samples in the dataset among them. In order to create a non-i.i.d. dataset, we follow a similar procedure as in [1]: first we sort all data points according to their classes. Then, they are split into 100 shards, and each user is randomly assigned 5 shards of data. We use the residual neural network (ResNet) defined in [44], which is a large model with $11,181,642$ parameters. To update the local models at each user using its local data, we apply stochastic gradient descent (SGD). In the reported experimental results, all users participate in each communication round.

A.2 ALGORITHMS TO COMPARE AND TUNING HYPERPARAMETERS

Distribution	Parameter setting
Dist1	Gaussian distribution $\mathcal{N}(2.0, 1.0)$
Dist2	mixture of $\mathcal{N}(0.2, 0.01)$, $\mathcal{N}(1.0, 0.1)$ and $\mathcal{N}(5.0, 1.0)$ with weights $(0.2, 0.6, 0.2)$
Dist3	Uniform distribution $U[0.2, 5]$
Dist4	mixture of $\mathcal{N}(0.2, 0.01)$, $\mathcal{N}(0.5, 0.1)$ and $\mathcal{N}(2.0, 1.0)$ with weights $(0.2, 0.6, 0.2)$
Dist5	Uniform distribution $U[0.2, 2]$
Dist6	mixture of $\mathcal{N}(0.2, 0.01)$, $\mathcal{N}(0.5, 0.1)$ and $\mathcal{N}(1.0, 0.1)$ with weights $(0.3, 0.5, 0.2)$
Dist7	Uniform distribution $U[0.2, 1]$
Dist8	mixture of $\mathcal{N}(0.2, 0.01)$ and $\mathcal{N}(0.5, 0.1)$ with weights $(0.6, 0.4)$
Dist9	Uniform distribution $U[0.2, 0.5]$

Table 4: Distribution of privacy preferences.

Algorithm 2: WeiAvg [19]

Input: Initial parameter $\boldsymbol{\theta}^0$, Clients batch sizes $\{b_1, \ldots, b_n\}$, Clients dataset sizes $\{N_1, \ldots, N_n\}$, Clients noise scales $\{z_1, \ldots, z_n\}$, gradient norm bound c, local epochs $\{K_1, \ldots, K_n\}$, global round E, privacy parameter δ, number of model parameters p, privacy accountant **PA**.

Output: $\boldsymbol{\theta}_E, \{\epsilon_1, \ldots, \epsilon_n\}$

Initialize $\boldsymbol{\theta}_0$ randomly.
for $e \in [E]$ **do**
  sample a set of clients $\mathcal{S}^e \subseteq \{1, \ldots, n\}$
  for *each client* $i \in \mathcal{S}^e$ ***in parallel*** **do**
    $\Delta\tilde{\boldsymbol{\theta}}_i^e \leftarrow$**DPSGD**$(\boldsymbol{\theta}^e, b_i, N_i, K_i, z_i, c)$
    $\epsilon_i^e \leftarrow$ **PA**$(\frac{b_i}{N_i}, z_i, K_i, e)$
  for $i \in \mathcal{S}^e$ **do**
    $w_i^e \leftarrow \frac{\epsilon_i}{\sum_{j \in \mathcal{S}_e} \epsilon_j}$
  $\boldsymbol{\theta}^{e+1} \leftarrow \boldsymbol{\theta}^e + \sum_{i \in \mathcal{S}_e} w_i^e \Delta\tilde{\boldsymbol{\theta}}_i^e$

Output: $\boldsymbol{\theta}^E, \{\epsilon_1^E, \ldots, \epsilon_n^E\}$

Algorithm 3: Principal Component Pursuit by Alternating Directions [24]

Input: matrix M, shrinkage operator $\mathcal{S}_\tau[x] = \text{sgn}(x)\max(|x| - \tau, 0)$, singular value thresholding operator $\mathcal{D}_\tau(U\Sigma V^*) = U\mathcal{S}_\tau(\Sigma)V^*$

Initialize $S_0 = Y_0 = 0, \mu > 0$.
while *not converged* **do**
  compute $L_{k+1} = \mathcal{D}_{\mu^{-1}}(M - S_k - \mu^{-1}Y_k)$
  compute $S_{k+1} = \mathcal{S}_{\lambda\mu^{-1}}(M - L_{k+1} + \mu^{-1}Y_k)$
  compute $Y_{k+1} = Y_k + \mu(M - L_{k+1} - S_{k+1})$

Output: L, S

Table 5: The learning rates used for training with each algorithm on MNIST dataset

alg \ dist	Dist1	Dist2	Dist3	Dist4	Dist5	Dist6	Dist7	Dist8	Dist9
WeiAvg [19]	1e-2	1e-2	1e-2	5e-3	5e-3	1e-3	1e-3	1e-3	1e-3
PFA [19]	1e-2	1e-2	1e-2	1e-2	1e-2	2e-3	2e-3	2e-3	2e-3
DPFedAvg [38]	5e-3	1e-3	1e-3	1e-3	1e-3	5e-4	1e-3	1e-3	1e-3
minimum ϵ [19]	5e-4	5e-4	5e-4	5e-4	1e-3	1e-4	1e-3	5e-4	1e-3
Robust-HDP	1e-2	1e-2	1e-2	1e-2	5e-3	2e-3	2e-3	2e-3	2e-3

Table 6: The learning rates used for training with each algorithm on FMNIST dataset

alg \ dist	Dist1	Dist2	Dist3	Dist4	Dist5	Dist6	Dist7	Dist8	Dist9
WeiAvg [19]	5e-3	5e-3	5e-3	5e-3	2e-3	5e-4	5e-4	5e-4	5e-4
PFA [19]	5e-3	5e-3	5e-3	5e-3	5e-3	5e-3	1e-3	1e-3	1e-3
DPFedAvg [38]	2e-3	1e-3	1e-3	1e-3	1e-3	5e-4	5e-4	5e-4	5e-4
minimum ϵ [19]	1e-3	5e-4	5e-4	5e-4	5e-4	1e-4	5e-4	5e-4	5e-4
Robust-HDP	5e-3	5e-3	5e-3	5e-3	5e-3	1e-3	1e-3	1e-3	1e-3

Table 7: The learning rates used for training with each algorithm on CIFAR-10 dataset

alg \ dist	Dist1	Dist2	Dist3	Dist4	Dist5	Dist6	Dist7	Dist8	Dist9
WeiAvg [19]	2e-3	1e-3	1e-3	5e-4	5e-4	2e-4	2e-4	1e-4	1e-4
PFA [19]	5e-3	2e-3	2e-3	2e-3	1e-3	5e-4	5e-4	2e-4	2e-4
DPFedAvg [38]	2e-3	1e-3	1e-3	5e-4	5e-4	2e-4	2e-4	1e-4	1e-4
minimum ϵ [19]	2e-3	1e-3	1e-3	5e-4	5e-4	2e-4	2e-4	1e-4	1e-4
Robust-HDP	5e-3	2e-3	2e-3	2e-3	1e-3	5e-4	5e-4	2e-4	2e-4

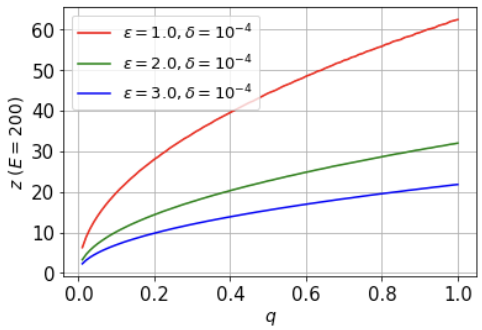

Figure 6: Plot of z v.s. q obtained from Moments Accountant [14] in a centralized setting with $E = 200$. Hence, z increases sub-linearly with q (or equivalemntly with b).

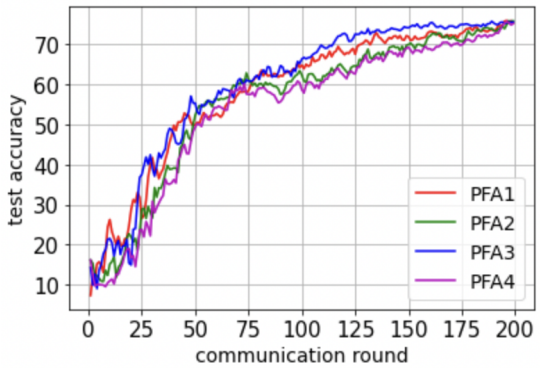

Figure 7: Comparison between the performance of PFA with different projections dimensions. The results are for Dist9. In this case, the total number of clients in the public cluster was 4. Therefore maximum dimension for projection was 4. We keep using 1 as the projection dimension (as in [19]).

B DERIVATIONS

B.1 COMPUTATION OF $\sigma^2_{i,\tilde{g}}$, WHEN GRADIENT CLIPPING IS INEFFECTIVE:

We know that the two sources of randomness (i.e. minibatch sampling and Gaussian noise) are independent, thus the variance is additive. Assuming that $E[\bar{g}_{ij}(\boldsymbol{\theta})]$ is the same for all j and is $G_i(\boldsymbol{\theta})$, we have:

$$\begin{aligned}
\sigma^2_{i,\tilde{g}} := \texttt{Var}(\tilde{g}_i(\boldsymbol{\theta})) &= \texttt{Var}\left(\frac{1}{b_i}\sum_{j\in\mathcal{B}_i^t}\bar{g}_{ij}(\boldsymbol{\theta})\right) + \frac{d\sigma^2_{i,\texttt{dp}}}{b_i^2} \\
&= \frac{1}{b_i^2}\left(\mathbb{E}\left[\left\|\sum_{j\in\mathcal{B}_i^t}\bar{g}_{ij}(\boldsymbol{\theta})\right\|^2\right] - \left\|\mathbb{E}\left[\sum_{j\in\mathcal{B}_i^t}\bar{g}_{ij}(\boldsymbol{\theta})\right]\right\|^2\right) + \frac{dc^2z^2(\epsilon_i,\delta_i,q_i,K_i,E)}{b_i^2} \\
&= \frac{1}{b_i^2}\left(\mathbb{E}\left[\left\|\sum_{j\in\mathcal{B}_i^t}\bar{g}_{ij}(\boldsymbol{\theta})\right\|^2\right] - \left\|\sum_{j\in\mathcal{B}_i^t}G_i(\boldsymbol{\theta})\right\|^2\right) + \frac{dc^2z^2(\epsilon_i,\delta_i,q_i,K_i,E)}{b_i^2} \\
&= \frac{1}{b_i^2}\left(\mathbb{E}\left[\left\|\sum_{j\in\mathcal{B}_i^t}\bar{g}_{ij}(\boldsymbol{\theta})\right\|^2\right] - b_i^2\left\|G_i(\boldsymbol{\theta})\right\|^2\right) + \frac{dc^2z^2(\epsilon_i,\delta_i,q_i,K_i,E)}{b_i^2} \quad (13) \\
& \quad (14)
\end{aligned}$$

We also have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j\in\mathcal{B}_i^t}\bar{g}_{ij}(\boldsymbol{\theta})\right\|^2\right] &= \sum_{j\in\mathcal{B}_i^t}\mathbb{E}\left[\left\|\bar{g}_{ij}(\boldsymbol{\theta})\right\|^2\right] + \sum_{m\neq n\in\mathcal{B}_i^t}2\mathbb{E}\left[[\bar{g}_{im}(\boldsymbol{\theta})]^\top[\bar{g}_{in}(\boldsymbol{\theta})]\right] \\
&= \sum_{j\in\mathcal{B}_i^t}\mathbb{E}\left[\left\|\bar{g}_{ij}(\boldsymbol{\theta})\right\|^2\right] + \sum_{m\neq n\in\mathcal{B}_i^t}2\mathbb{E}\left[\bar{g}_{im}(\boldsymbol{\theta})\right]^\top\mathbb{E}\left[\bar{g}_{in}(\boldsymbol{\theta})\right] \\
&= b_ic^2 + 2\binom{b_i}{2}\left\|G_i(\boldsymbol{\theta})\right\|^2, \quad (15)
\end{aligned}$$

where the last equation has used eq. (2) and that we clip the norm of sample gradients $\bar{g}_{ij}(\boldsymbol{\theta})$ with an "effective" clipping threshold c. We can now rewrite eq. 13 as:

$$\begin{aligned}
\sigma^2_{i,\tilde{g}} := \texttt{Var}(\tilde{g}_i(\boldsymbol{\theta})) &= \frac{1}{b_i^2}\left(\mathbb{E}\left[\left\|\sum_{j\in\mathcal{B}_i^t}\bar{g}_{ij}(\boldsymbol{\theta})\right\|^2\right] - b_i^2\left\|G_i(\boldsymbol{\theta})\right\|^2\right) + \frac{dc^2z^2(\epsilon_i,\delta_i,q_i,K_i,E)}{b_i^2} \quad (16) \\
&= \frac{1}{b_i^2}\left(b_ic^2 + \left(2\binom{b_i}{2} - b_i^2\right)\left\|G_i(\boldsymbol{\theta})\right\|^2\right) + \frac{dc^2z^2(\epsilon_i,\delta_i,q_i,K_i,E)}{b_i^2} \quad (17) \\
&= \frac{1}{b_i^2}\left(b_ic^2 - b_i\left\|G_i(\boldsymbol{\theta})\right\|^2\right) + \frac{dc^2z^2(\epsilon_i,\delta_i,q_i,K_i,E)}{b_i^2} \quad (18) \\
&= \frac{c^2 - \left\|G_i(\boldsymbol{\theta})\right\|^2}{b_i} + \frac{dc^2z^2(\epsilon_i,\delta_i,q_i,K_i,E)}{b_i^2} \quad (19) \\
&\approx \frac{dc^2z^2(\epsilon_i,\delta_i,q_i,K_i,E)}{b_i^2} \quad (20)
\end{aligned}$$

C ASSUMPTIONS AND LEMMAS

In this section, we formalize some definitions and lemmas, which we will use in our proofs.

Assumption 1 (Lipschitz continuity, β-smoothness and bounded gradient variance). *$\{f_i\}_{i=1}^n$ are L_0-Lipschitz continuous and β-smooth: $\forall\, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p, i : \|f_i(\boldsymbol{\theta}) - f_i(\boldsymbol{\theta}')\| \leq L_0\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ and $\|\nabla f_i(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}')\| \leq \beta\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$. Also, the stochastic gradient $g_i(\boldsymbol{\theta})$ is an unbiased estimate of $\nabla f_i(\boldsymbol{\theta})$ with bounded variance: $\forall \boldsymbol{\theta} \in \mathbb{R}^p : \mathbb{E}_{\mathcal{B}_i^t}[g_i(\boldsymbol{\theta})] = \nabla f_i(\boldsymbol{\theta}),\ \mathbb{E}_{\mathcal{B}_i^t}\left[\|g_i(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta})\|^2\right] \leq \sigma_{i,g}^2$. We also assume that for every $i, j \in [n]$, $f_i - f_j$ is σ-Lipschitz continuous: $\|\nabla f_i(\boldsymbol{\theta}) - \nabla f_j(\boldsymbol{\theta})\| \leq \sigma$.*

Assumption 2 (bounded sample gradients). *There exists a clipping threshold $\mathcal{C}$ such that for all i, j:*

$$\|g_{ij}(\boldsymbol{\theta})\|_2 := \|\nabla \ell(h(x_{ij}, \boldsymbol{\theta}), y_{ij})\|_2 \leq \mathcal{C} \tag{21}$$

Note that this condition always holds if ℓ (cross-entropy loss) is Lipschitz continuous or if h is bounded.

Lemma 2 (Relaxed triangle inequality)**.** *Let $\{v_1, \ldots, v_n\}$ be n vectors in $\mathbb{R}^d$. Then, the followings is true:*

- $\|v_i + v_j\|^2 \leq (1 + a)\|v_i\|^2 + (1 + \frac{1}{a})\|v_j\|^2$ *(for any $a > 0$)*
- $\|\sum_i v_i\|^2 \leq n \sum_i \|v_i\|^2$

Proof. The proof for the first inequality is obtained from identity:

$$\|v_i + v_j\|^2 = (1 + a)\|v_i\|^2 + (1 + \frac{1}{a})\|v_j\|^2 - \|\sqrt{a}v_i + \frac{1}{\sqrt{a}}v_j\|^2 \tag{22}$$

The proof for the second inequality is achieved by using the fact that $h(x) = \|x\|^2$ is convex:

$$\|\frac{1}{n}\sum_i v_i\|^2 \leq \frac{1}{n}\sum_i \|v_i\|^2 \tag{23}$$

□

Lemma 3. *Let $\{v_1, \ldots, v_n\}$ be n random variables in $\mathbb{R}^d$, with $\mathbb{E}[v_i] = \mathcal{E}_i$ and $\mathbb{E}[\|v_i - \mathcal{E}_i\|^2] = \sigma_i^2$. Then, we have the following inequality:*

$$\mathbb{E}[\|\sum_{i=1}^n v_i\|^2] \leq \|\sum_{i=1}^n \mathcal{E}_i\|^2 + n\sum_{i=1}^n \sigma_i^2. \tag{24}$$

Proof. From the definition of variance, we have:

$$\mathbb{E}[\|\sum_{i=1}^n v_i\|^2] = \|\sum_{i=1}^n \mathcal{E}_i\|^2 + \mathbb{E}[\|\sum_{i=1}^n (v_i - \mathcal{E}_i)\|^2] \tag{25}$$

$$\leq \|\sum_{i=1}^n \mathcal{E}_i\|^2 + n\sum_{i=1}^n \mathbb{E}[\|v_i - \mathcal{E}_i\|^2] \tag{26}$$

$$= \|\sum_{i=1}^n \mathcal{E}_i\|^2 + n\sum_{i=1}^n \sigma_i^2, \tag{27}$$

$$\tag{28}$$

where the inequality is based on the lemma 2. □

D PROOFS

Lemma 1 (Precision of Robust-HDP). *Let $s_{i,j}$ in matrix $\mathbf{S}$ represent the true value of noise in the i-th element of $\Delta\tilde{\boldsymbol{\theta}}_j^e$ ($j \in \mathcal{S}^e$). Then, assume that $\mathbf{S}'$ is the matrix computed by Robust-HDP at the server with bounded elements $s_{i,j}'^2 \leq U$, where $\mathbb{E}[s'_{i,j}] = r s_{i,j}$, for some constant $r > 0$, and $\mathbb{E}[|s'_{i,j} - r s_{i,j}|^2] \leq \alpha_j^2$ (i.e., on average, Robust-HDP is able to estimate the true noise values $s_{i,j}$ up to a multiplicative factor r). Then:*

$$Pr(|\hat{\sigma}_j^2 - (r^2\sigma_j^2 + \alpha_j^2)| > \epsilon) \leq 2e^{\frac{-2p\epsilon^2}{U^2}}. \tag{9}$$

Proof. We have $\mathbb{E}[s_{i,j}'^2] = \texttt{Var}(s'_{i,j}) + (\mathbb{E}[s'_{i,j}])^2 = \alpha_j^2 + (r\mathbb{E}[s_{i,j}])^2 = \alpha_j^2 + r^2\sigma_j^2$. Hence, $\mathbb{E}[\frac{s_{i,j}'^2}{p}] = \frac{\alpha_j^2 + r^2\sigma_j^2}{p}$ (for all rows i). Also, from $s_{i,j}'^2 \leq U$, we have $\frac{s_{i,j}'^2}{p} \leq \frac{U}{p}$. Therefore, by applying Hoeffding's inequality to the sum $\hat{\sigma}_j^2 = \sum_{i=1}^p \frac{s_{ij}'^2}{p}$, we get:

$$\texttt{Pr}(|\hat{\sigma}_j^2 - (r^2\sigma_j^2 + \alpha_j^2)| > \epsilon) \leq 2e^{\frac{-2p\epsilon^2}{U^2}} \tag{29}$$

□

Property 1 (Parallel Composition [45]). *Assume each of the randomized mechanisms $M_i : \mathcal{D}_i \to \mathbb{R}$ for $i \in [n]$ satisfies (ϵ_i, δ_i)-DP and their domains $\mathcal{D}_i$ are disjoint subsets. Any function g of the form $g(M_1, \ldots, M_n)$ satisfies $(\max_i \epsilon_i, \max_i \delta_i)$-DP.*

Theorem 1. *For each client i , there exist constants c_1 and c_2 such that given its number of steps $E \cdot E_i$, for any $\epsilon < c_1 q_i^2 E \cdot E_i$, the output model of Robust-HDP satisfies $(\epsilon_i, \delta_i)-DP$ with respect to $\mathcal{D}_i$ for any $\delta_i > 0$ if $z_i > c_2 \frac{q_i\sqrt{E\cdot E_i \log\frac{1}{\delta_i}}}{\epsilon_i}$, where z_i is the noise scale used by the client i for DPSGD. The algorithm also satisfies $(\epsilon_{max}, \delta_{max})$-DP, where $(\epsilon_{max}, \delta_{max}) = \big(\max(\{\epsilon_i\}_{i=1}^n), \max(\{\delta_i\}_{i=1}^n)\big)$.*

Proof. The proof for the first part follows the proof of DPSGD algorithm [14]. Also, in Robust-HDP, each client i runs DPSGD locally to achieve (ϵ_i, δ_i)-DP independently. Hence, it satisfies heterogeneous DP with the set of preferences $\{(\epsilon_i, \delta_i)\}_{i=1}^n$. Also, the clients datasets $\{\mathcal{D}_i\}_{i=1}^n$ are disjoint. Hence, as Robust-HDP runs RPCA on the clients models updates, it satisfies $\big(\max(\{\epsilon_i\}_{i=1}^n), \max(\{\delta_i\}_{i=1}^n)\big)$-DP, according to parallel composability property above. □

Theorem 2 (Robust-HDP). *Assume that Assumption 1 holds, and for every i, learning rate η_l satisfies: $\eta_l \leq \frac{1}{6\beta E_i}$ and $\eta_l \leq \frac{1}{12\beta\sqrt{(1+\sum_{i=1}^n E_i)\left(\sum_{i=1}^n E_i^4\right)}}$. Then, we have:*

$$\min_{0\leq e\leq E-1} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}^e)\|^2] \leq \frac{12}{(11E_l^{min} - 7)}\left(\frac{f(\boldsymbol{\theta}^0) - f^*}{E\eta_l} + \frac{\sum_{e=0}^{E-1}(\Psi_\sigma^e + \Psi_p^e)}{E}\right), \tag{10}$$

where $E_l^{min} = \min_i E_i$ and

$$\Psi_\sigma^e = 6\beta^2\eta_l^2(1 + \sum_{i=1}^n E_i)\left(2\sum_{i=1}^n E_i^4\sigma^2 + \frac{1}{3}\sum_{i=1}^n E_i^3\sigma_{i,\tilde{g}}^2\right) + \beta\eta_l\sum_{i=1}^n E_i^2\sigma_{i,\tilde{g}}^2$$

$$\Psi_p^e = \frac{8L_0^2}{3}\left(n\sum_{i=1}^n E_i^2\mathbb{E}[(w_i^e - \lambda_i)^2] + \|\boldsymbol{\lambda}\|^2\sum_{i=1}^n \mathbb{E}[(E_i - \mu_w^e)^2]\right), \text{where } \mu_w^e = \sum_{i=1}^n w_i^e E_i. \tag{11}$$

Proof. From our assumption 1 and that we use cross-entropy loss, we can conclude that Assumption 2 also holds for some $\mathcal{C}$. When we use a ineffective clipping threshold $\mathcal{C}$, we have:

$$\tilde{g}_i(\boldsymbol{\theta}^t) = \frac{\sum_{j\in\mathcal{B}_i^t} g_{ij}(\boldsymbol{\theta}^t)}{b_i} + \mathcal{N}(0, \frac{\sigma_{i,\text{dp}}^2}{b_i^2}I_p) = g_i(\boldsymbol{\theta}^t) + \mathcal{N}(0, \frac{\sigma_{i,\text{dp}}^2}{b_i^2}I_p) \tag{30}$$

Therefore:

$$\mathbb{E}_e[\tilde{g}_i(\boldsymbol{\theta})] = \mathbb{E}_e[g_i(\boldsymbol{\theta})] = \nabla f_i(\boldsymbol{\theta}) \tag{31}$$

$$\texttt{Var}(\tilde{g}_i(\boldsymbol{\theta})) = \texttt{Var}(g_i(\boldsymbol{\theta})) + \frac{p\sigma_{i,\text{dp}}^2}{b_i^2} \leq \sigma_{i,\tilde{g}}^2 := \sigma_{i,g}^2 + \frac{p\sigma_{i,\text{dp}}^2}{b_i^2}. \tag{32}$$

i.e., the assumption of having unbiased gradient with bounded variance still holds (with a larger bound $\sigma_{i,\tilde{g}}^2$, due to adding DP noise). Consistent with the previous notations, we assume that the set of participating clients in round e are $\mathcal{S}^e$, and for every client $i \notin \mathcal{S}^e$, we set $w_i^e = 0$. Using this, we can write the model parameter at the end of round e as:

$$\boldsymbol{\theta}^{e+1} = \sum_{i=1}^{n} w_i^e \boldsymbol{\theta}_{i,E_i}^e, \tag{33}$$

where $\{E_i\}_{i=1}^n$ is the heterogeneous number of gradient steps of clients (depending on their dataset size and batch size). From $\boldsymbol{\theta}_{i,k}^e = \boldsymbol{\theta}_{i,k-1}^e - \eta_l \tilde{g}_i(\boldsymbol{\theta}_{i,k-1}^e)$, we can rewrite the equation above as:

$$\boldsymbol{\theta}^{e+1} = \boldsymbol{\theta}^e - \eta_l \sum_{i \in \mathcal{S}^e} w_i^e \sum_{k=1}^{E_i} \tilde{g}_i(\boldsymbol{\theta}_{i,k-1}^e) = \boldsymbol{\theta}^e - \eta_l \sum_{i=1}^{n} w_i^e \sum_{k=1}^{E_i} \tilde{g}_i(\boldsymbol{\theta}_{i,k-1}^e). \tag{34}$$

Note that the second equality holds because we assumed above that if client i is not participating in round e (i.e., $i \notin \mathcal{S}^e$), we set $w_i^e = 0$. From β-smoothness of $\{f_i\}_{i=1}^n$, and consequently β-smoothness of f, we have:

$$\begin{aligned} f(\boldsymbol{\theta}^{e+1}) &\leq f(\boldsymbol{\theta}^e) + \langle \nabla f(\boldsymbol{\theta}^e), \boldsymbol{\theta}^{e+1} - \boldsymbol{\theta}^e \rangle + \frac{\beta}{2} \|\boldsymbol{\theta}^{e+1} - \boldsymbol{\theta}^e\|^2 \\ &= f(\boldsymbol{\theta}^e) - \eta_l \big\langle \nabla f(\boldsymbol{\theta}^e), \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) \big\rangle + \frac{\beta \eta_l^2}{2} \big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) \big\|^2 \end{aligned} \tag{35}$$

Now, we use identity $\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) = \nabla f(\boldsymbol{\theta}^e) + \tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)$ to rewrite the equation above as:

$$\begin{aligned} f(\boldsymbol{\theta}^{e+1}) \leq f(\boldsymbol{\theta}^e) &- \eta_l \big\langle \nabla f(\boldsymbol{\theta}^e), \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \nabla f(\boldsymbol{\theta}^e) \big\rangle - \eta_l \big\langle \nabla f(\boldsymbol{\theta}^e), \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \big\rangle \\ &+ \frac{\beta \eta_l^2}{2} \big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) + \sum_{i=1}^{n} w_i^e E_i \nabla f(\boldsymbol{\theta}^e) \big\|^2 \end{aligned} \tag{36}$$

Hence,

$$\begin{aligned} f(\boldsymbol{\theta}^{e+1}) \leq f(\boldsymbol{\theta}^e) &- \eta_l \big\langle \nabla f(\boldsymbol{\theta}^e), \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \nabla f(\boldsymbol{\theta}^e) \big\rangle - \eta_l \big\langle \nabla f(\boldsymbol{\theta}^e), \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \big\rangle \\ &+ \frac{\beta \eta_l^2}{2} \big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \|^2 + \frac{\beta \eta_l^2}{2} (\underbrace{\sum_{i=1}^{n} w_i^e E_i}_{\bar{E}^e})^2 \big\| \nabla f(\boldsymbol{\theta}^e) \big\|^2 \\ &+ \beta \eta_l^2 \big\langle \sum_{i=1}^{n} w_i^e E_i \nabla f(\boldsymbol{\theta}^e), \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \big\rangle. \end{aligned} \tag{37}$$

Note that we denote $\sum_{i=1}^n w_i^e E_i$ with $\bar{E}^e$ from now on. With doing some algebra we get to:

$$f(\boldsymbol{\theta}^{e+1}) \leq f(\boldsymbol{\theta}^e) - \eta_l \bar{E}^e (1 - \frac{\beta}{2}\eta_l \bar{E}^e) \|\nabla f(\boldsymbol{\theta}^e)\|^2 \\ - \eta_l (1 - \beta \eta_l \bar{E}^e) \big\langle \nabla f(\boldsymbol{\theta}^e), \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i - 1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \big\rangle \\ + \frac{\beta \eta_l^2}{2} \Big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i - 1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \Big\|^2. \tag{38}$$

By taking expectation from both side and using Cauchy-Schwarz inequality, we have:

$$\mathbb{E}\big[f(\boldsymbol{\theta}^{e+1})\big] \leq \mathbb{E}\big[f(\boldsymbol{\theta}^e)\big] - \eta_l \bar{E}^e (1 - \frac{\beta \eta_l}{2} \bar{E}^e) \mathbb{E}\big[\|\nabla f(\boldsymbol{\theta}^e)\|^2\big] \\ + \eta_l (1 - \beta \eta_l \bar{E}^e) \mathbb{E}\Big[\|\nabla f(\boldsymbol{\theta}^e)\| \times \Big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i - 1} \big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \Big\| \Big] \\ + \frac{\beta \eta_l^2}{2} \mathbb{E}\Big[\Big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i - 1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \Big\|^2 \Big]. \tag{39}$$

Now, we use the inequality $ab \leq \frac{1}{2}(a^2 + b^2)$ for the second line to get:

$$\mathbb{E}\big[f(\boldsymbol{\theta}^{e+1})\big] \leq \mathbb{E}\big[f(\boldsymbol{\theta}^e)\big] + \underbrace{\big(\frac{1}{2}\eta_l (1 - \beta \eta_l \bar{E}^e) - \eta_l \bar{E}^e (1 - \frac{\beta \eta_l}{2} \bar{E}^e)\big)}_{\leq -\eta_l \frac{11\bar{E}^e - 6}{12}} \mathbb{E}\big[\|\nabla f(\boldsymbol{\theta}^e)\|^2\big] \\ + \frac{1}{2}\eta_l (1 - \beta \eta_l \bar{E}^e) \mathbb{E}\Big[\Big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i - 1} \big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \Big\|^2 \Big] \\ + \frac{\beta \eta_l^2}{2} \mathbb{E}\Big[\Big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i - 1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \Big\|^2 \Big], \tag{40}$$

where the constant inequality in the first line is achieved from our assumption that $\eta_l \leq \frac{1}{6\beta E_i}$ (and consequently: $\eta_l \leq \frac{1}{6\beta \bar{E}^e}$):

$$\begin{aligned} \frac{1}{2}\eta_l (1 - \beta \eta_l \bar{E}^e) - \eta_l \bar{E}^e (1 - \frac{\beta \eta_l}{2} \bar{E}^e) &\leq -\eta_l \left(\bar{E}^e - \frac{1}{2} - \frac{\beta \eta_l}{2} \bar{E}^{e^2} + \frac{1}{2} \beta \eta_l \bar{E}^e \right) \\ &\leq -\eta_l \left(\frac{11\bar{E}^e - 6}{12} + \frac{\beta \eta_l \bar{E}^e}{2} \right) \\ &\leq -\eta_l \frac{11\bar{E}^e - 6}{12}. \end{aligned} \tag{41}$$

Therefore,

$$\mathbb{E}\big[f(\boldsymbol{\theta}^{e+1})\big] \leq \mathbb{E}\big[f(\boldsymbol{\theta}^e)\big] - \eta_l \frac{11\bar{E}^e - 6}{12} \mathbb{E}\big[\|\nabla f(\boldsymbol{\theta}^e)\|^2\big] \\ + \frac{1}{2}\eta_l (1 - \beta \eta_l \bar{E}^e) \mathbb{E}\Big[\Big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i - 1} \big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \Big\|^2 \Big] \\ + \frac{\beta \eta_l^2}{2} \mathbb{E}\Big[\Big\| \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i - 1} \big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big) \Big\|^2 \Big]. \tag{42}$$

Now, we use the relaxed triangle inequality $\|a+b\|^2 \leq 2(\|a\|^2+\|b\|^2)$ for the last line above:

$$\begin{aligned}\mathbb{E}\big[f(\boldsymbol{\theta}^{e+1})\big] \leq \mathbb{E}\big[f(\boldsymbol{\theta}^{e})\big] - \eta_l \frac{11\bar{E}^e - 6}{12}\mathbb{E}\big[\|\nabla f(\boldsymbol{\theta}^e)\|^2\big] \\ + \frac{1}{2}\eta_l(1-\beta\eta_l\bar{E}^e)\underbrace{\mathbb{E}\Big[\Big\|\sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big)\Big\|^2\Big]}_{\mathcal{B}} \\ + \beta\eta_l^2\underbrace{\mathbb{E}\Big[\Big\|\sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f_i(\boldsymbol{\theta}_{i,k}^e)\big)\Big\|^2\Big]}_{\mathcal{A}} + \beta\eta_l^2\underbrace{\mathbb{E}\Big[\Big\|\sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big)\Big\|^2\Big]}_{\mathcal{B}}\end{aligned} \tag{43}$$

Now, we bound each of the terms $\mathcal{A}$ and $\mathcal{B}$ separately:

$$\begin{aligned}\mathcal{A} &\leq \mathbb{E}\Big[\Big(\sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\big\|\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f_i(\boldsymbol{\theta}_{i,k}^e)\big\|\Big)^2\Big] \leq \mathbb{E}\Big[\sum_{i=1}^{n}(w_i^e)^2 \times \sum_{i=1}^{n}\Big(\sum_{k=0}^{E_i-1}\big\|\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f_i(\boldsymbol{\theta}_{i,k}^e)\big\|\Big)^2\Big] \\ &= \mathbb{E}\Big[\|\mathbf{w}^e\|^2\sum_{i=1}^{n}\Big(\sum_{k=0}^{E_i-1}\big\|\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f_i(\boldsymbol{\theta}_{i,k}^e)\big\|\Big)^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{n}\Big(\sum_{k=0}^{E_i-1}\big\|\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f_i(\boldsymbol{\theta}_{i,k}^e)\big\|\Big)^2\Big] \\ &\leq \sum_{i=1}^{n} E_i \sum_{k=0}^{E_i-1}\mathbb{E}\Big[\Big\|\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f_i(\boldsymbol{\theta}_{i,k}^e)\Big\|^2\Big] \leq \sum_{i=1}^{n} E_i^2\sigma_{i,\tilde{g}}^2,\end{aligned} \tag{44}$$

where in the first inequality, we used triangle inequality and in the second and third inequalities, we used Cauchy-Schwarz inequality.

Similarly, we can bound $\mathcal{B}$:

$$\begin{aligned}\mathcal{B} &= \mathbb{E}\Big[\Big\|\sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e) - \nabla f(\boldsymbol{\theta}^e)\big)\Big\|^2\Big] = \mathbb{E}\Big[\Big\|\sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e) - \sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\nabla f(\boldsymbol{\theta}^e)\Big\|^2\Big] \\ &= \mathbb{E}\Big[\Big\|\sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e) - \big(\sum_{i=1}^{n} w_i^e E_i\big)\nabla f(\boldsymbol{\theta}^e)\Big\|^2\Big] = \mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^{n} w_i^e \sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)\Big) - \mu_w^e\nabla f(\boldsymbol{\theta}^e)\Big\|^2\Big],\end{aligned} \tag{45}$$

where μ_w^e is the weighted average of number of local steps $\{E_i\}$, obtained from weights $\{w_i^e\}$. Let us define $\Delta_i^e := w_i^e - \lambda_i$ to be the difference between the aggregation weight of client i in round e (w_i^e) and its corresponding aggregation weights in the global objective function $f(\boldsymbol{\theta})$ (λ_i). With this definition, we have:

$$\begin{aligned}\mathcal{B} &= \mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^{n}\Delta_i^e\sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)\Big) + \Big(\sum_{i=1}^{n}\lambda_i\sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)\Big) - \Big(\sum_{i=1}^{n}\lambda_i\mu_w^e\nabla f_i(\boldsymbol{\theta}^e)\Big)\Big\|^2\Big] \\ &\leq \underbrace{2\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}\Delta_i^e\sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)\Big\|^2\Big]}_{\mathcal{C}} + \underbrace{2\mathbb{E}\Big[\Big\|\Big(\sum_{i=1}^{n}\lambda_i\sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)\Big) - \Big(\sum_{i=1}^{n}\lambda_i\mu_w^e\nabla f_i(\boldsymbol{\theta}^e)\Big)\Big\|^2\Big]}_{\mathcal{D}}.\end{aligned} \tag{46}$$

Now, we bound each of the terms $\mathcal{C}$ and $\mathcal{D}$, separately:

$$\begin{aligned}
\mathcal{C} &= 2\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}\Delta_i^e\sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)\Big\|^2\Big]\\
&\leq 4\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}\Delta_i^e\sum_{k=0}^{E_i-1}\big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e)-\nabla f_i(\boldsymbol{\theta}^e)\big)\Big\|^2\Big]+4\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}E_i\Delta_i^e\nabla f_i(\boldsymbol{\theta}^e)\Big\|^2\Big]\\
&\leq 4(\sum_{i=1}^{n}E_i)\beta^2\sum_{i=1}^{n}\sum_{k=0}^{E_i-1}|\Delta_i^e|^2\|\boldsymbol{\theta}_{i,k}^e-\boldsymbol{\theta}^e\|^2+4nL_0^2\sum_{i=1}^{n}E_i^2|\Delta_i^e|^2\\
&\leq 4\beta^2(\sum_{i=1}^{n}E_i)\sum_{i=1}^{n}\sum_{k=0}^{E_i-1}\|\boldsymbol{\theta}_{i,k}^e-\boldsymbol{\theta}^e\|^2+4nL_0^2\sum_{i=1}^{n}E_i^2\mathbb{E}[|\Delta_i^e|^2]. \qquad (47)
\end{aligned}$$

Similarly:

$$\begin{aligned}
\mathcal{D} &= 2\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}\lambda_i\Big(\sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)-\mu_w^e\nabla f_i(\boldsymbol{\theta}^e)\Big)\Big\|^2\Big]\\
&\leq 2\|\boldsymbol{\lambda}\|^2\sum_{i=1}^{n}\mathbb{E}\Big[\Big\|\sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)-\mu_w^e\nabla f_i(\boldsymbol{\theta}^e)\Big\|^2\Big]\\
&\leq 2\|\boldsymbol{\lambda}\|^2\sum_{i=1}^{n}\mathbb{E}\Big[\Big\|\sum_{k=0}^{E_i-1}\big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e)-\nabla f_i(\boldsymbol{\theta}^e)\big)+\big(E_i-\mu_w^e\big)\nabla f_i(\boldsymbol{\theta}^e)\Big\|^2\Big]\\
&\leq 4\|\boldsymbol{\lambda}\|^2\sum_{i=1}^{n}\mathbb{E}\Big[\Big\|\sum_{k=0}^{E_i-1}\nabla f_i(\boldsymbol{\theta}_{i,k}^e)-\nabla f_i(\boldsymbol{\theta}^e)\Big\|^2+\mathbb{E}[\big(E_i-\mu_w^e\big)^2]\underbrace{\big\|\nabla f_i(\boldsymbol{\theta}^e)\big\|^2}_{\leq L_0^2}\Big]\\
&\leq 4\beta^2\|\boldsymbol{\lambda}\|^2\sum_{i=1}^{n}E_i\sum_{k=0}^{E_i-1}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{i,k}^e-\boldsymbol{\theta}^e\big\|^2\big]+4L_0^2\|\boldsymbol{\lambda}\|^2\sum_{i=1}^{n}\mathbb{E}[(E_i-\mu_w^e)^2]. \qquad (48)
\end{aligned}$$

Hence, by plugging the bounds above into eq. (46), we get:

$$\mathcal{B}\leq 4\beta^2(1+\sum_{i=1}^{n}E_i)\Big(\sum_{i=1}^{n}E_i\sum_{k=0}^{E_i-1}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{i,k}^e-\boldsymbol{\theta}^e\big\|^2\big]\Big)+4L_0^2\Big(n\sum_{i=1}^{n}E_i^2|\Delta_i^e|^2+\|\boldsymbol{\lambda}\|^2\sum_{i=1}^{n}\mathbb{E}[(E_i-\mu_w^e)^2]\Big) \qquad (49)$$

By plugging the bounds above on $\mathcal{A}$ and $\mathcal{B}$ into eq. (43), we get:

$$\begin{aligned}
\mathbb{E}\big[f(\boldsymbol{\theta}^{e+1})\big] &\leq \mathbb{E}\big[f(\boldsymbol{\theta}^e)\big]-\eta_l\frac{11\bar{E}^e-6}{12}\mathbb{E}\big[\|\nabla f(\boldsymbol{\theta}^e)\|^2\big]\\
&\quad+\beta\eta_l^2\underbrace{\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}w_i^e\sum_{k=0}^{E_i-1}\big(\tilde{g}_i(\boldsymbol{\theta}_{i,k}^e)-\nabla f_i(\boldsymbol{\theta}_{i,k}^e)\big)\Big\|^2\Big]}_{\mathcal{A}}\\
&\quad+\underbrace{(\beta\eta_l^2+\frac{1}{2}\eta_l(1-\beta\eta_l\bar{E}^e))}_{<\frac{2}{3}\eta_l}\underbrace{\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}w_i^e\sum_{k=0}^{E_i-1}\big(\nabla f_i(\boldsymbol{\theta}_{i,k}^e)-\nabla f(\boldsymbol{\theta}^e)\big)\Big\|^2\Big]}_{\mathcal{B}}, \qquad (50)
\end{aligned}$$

where from the assumption $\eta_l\leq\frac{1}{6\beta E_i}$, we get to $\frac{\beta\eta_l^2}{2}\leq\frac{\eta_l}{12}$. Hence:

$$\beta\eta_l^2+\frac{1}{2}\eta_l(1-\beta\eta_l\bar{E}^e)=\beta\eta_l^2(1-\frac{\bar{E}^e}{2})+\frac{\eta_l}{2}\leq\frac{\beta\eta_l^2}{2}+\frac{\eta_l}{2}\leq\frac{\eta_l}{12}+\frac{\eta_l}{2}<\frac{2\eta_l}{3}. \qquad (51)$$

Therefore, we have:

$$\mathbb{E}\big[f(\boldsymbol{\theta}^{e+1})\big] \leq \mathbb{E}\big[f(\boldsymbol{\theta}^{e})\big] - \eta_l \frac{11\bar{E}^e - 6}{12}\mathbb{E}\big[\|\nabla f(\boldsymbol{\theta}^e)\|^2\big] + \beta\eta_l^2 \sum_{i=1}^{n} E_i^2 \sigma_{i,\tilde{g}}^2$$
$$+ \left(\frac{8}{3}\beta^2\eta_l(1+\sum_{i=1}^{n}E_i)\left(\sum_{i=1}^{n}E_i\sum_{k=0}^{E_i-1}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{i,k}^e - \boldsymbol{\theta}^e\big\|^2\big]\right)\right)$$
$$+ \left(\frac{8}{3}L_0^2\eta_l\left(n\sum_{i=1}^{n}E_i^2\mathbb{E}[|\Delta_i^e|^2] + \|\boldsymbol{\lambda}\|^2\sum_{i=1}^{n}\mathbb{E}[(E_i-\mu_w^e)^2]\right)\right). \tag{52}$$

We now have the following lemma to bound local drift of clients during each communication round e:

Lemma 4 (**Bounded local drifts**). *Suppose Assumption 1 holds. The local drift happening at client i during communication round e is bounded:*

$$\xi_i^e := \sum_{k=0}^{E_i-1}\mathbb{E}\big[\big\|\boldsymbol{\theta}_{i,k}^e - \boldsymbol{\theta}^e\big\|^2\big] \leq (\texttt{cte}-2)E_i^2\eta_l^2\left(\sigma_{i,\tilde{g}^2} + 6E_i\sigma^2 + 6E_i\mathbb{E}[\|\nabla f(\boldsymbol{\theta}^e)\|^2]\right), \tag{53}$$

where `cte` *is the mathematical constant e.*

Proof. From $\boldsymbol{\theta}_{i,0}^e = \boldsymbol{\theta}^e$, we only need to focus on $E_i \geq 2$. We have:

$$\mathbb{E}\|\boldsymbol{\theta}_{i,k}^e - \boldsymbol{\theta}^e\|^2 = \mathbb{E}[\|\boldsymbol{\theta}_{i,k-1}^e - \eta_l\tilde{g}_i(\boldsymbol{\theta}_{i,k-1}^e) - \boldsymbol{\theta}^e\|^2]$$
$$\leq \mathbb{E}[\|\boldsymbol{\theta}_{i,k-1}^e - \eta_l\nabla f_i(\boldsymbol{\theta}_{i,k-1}^e) - \boldsymbol{\theta}^e\|^2] + \eta_l^2\sigma_{i,\tilde{g}}^2 \tag{54}$$

where the inequality comes from lemma 3. The first term on the right side of the above inequality can be bounded as:

$$\mathbb{E}[\|\boldsymbol{\theta}_{i,k-1}^e - \eta_l\nabla f_i(\boldsymbol{\theta}_{i,k-1}^e) - \boldsymbol{\theta}^e\|^2] \leq \left(1+\frac{1}{2E_i-1}\right)\mathbb{E}[\|\boldsymbol{\theta}_{i,k-1}^e - \boldsymbol{\theta}^e\|^2] + 2E_i\eta_l^2\mathbb{E}\|[\nabla f_i(\boldsymbol{\theta}_{i,k-1}^e)\|^2], \tag{55}$$

where we have used lemma 2. Now, we bound the last term in the above inequality. We have:

$$\nabla f_i(\boldsymbol{\theta}_{i,k-1}^e) = (\nabla f_i(\boldsymbol{\theta}_{i,k-1}^e) - \nabla f_i(\boldsymbol{\theta}^e)) + (\nabla f_i(\boldsymbol{\theta}^e) - \nabla f(\boldsymbol{\theta}^e)) + \nabla f(\boldsymbol{\theta}^e), \tag{56}$$

By using relaxed triangle inequality (lemma 2) and Assumption 1, we get:

$$\|\nabla f_i(\boldsymbol{\theta}_{i,k-1}^e)\|^2 = 3\|\nabla f_i(\boldsymbol{\theta}_{i,k-1}^e) - \nabla f_i(\boldsymbol{\theta}^e)\|^2 + 3\|\nabla f_i(\boldsymbol{\theta}^e) - \nabla f(\boldsymbol{\theta}^e)\|^2 + 3\|\nabla f(\boldsymbol{\theta}^e)\|^2$$
$$\leq 3\beta^2\|\boldsymbol{\theta}_{i,k-1}^e - \boldsymbol{\theta}^e\|^2 + 3\sigma^2 + 3\|\nabla f(\boldsymbol{\theta})\|^2. \tag{57}$$

Now, we can rewrite eq. (55) and then eq. (54):

$$\mathbb{E}\|\boldsymbol{\theta}_{i,k}^e - \boldsymbol{\theta}^e\|^2 \leq \underbrace{\left(1+\frac{1}{2E_i-1}+6E_i\beta^2\eta_l^2\right)}_{\leq 1+\frac{1}{E_i}}\mathbb{E}[\|\boldsymbol{\theta}_{i,k-1}^e - \boldsymbol{\theta}^e\|^2] + \eta_l^2(6E_i\sigma^2+\sigma_{i,\tilde{g}}^2) + 6E_i\eta_l^2\mathbb{E}\|\nabla f(\boldsymbol{\theta}^e)\|^2$$
$$\leq (1+\frac{1}{E_i})\mathbb{E}[\|\boldsymbol{\theta}_{i,k-1}^e - \boldsymbol{\theta}^e\|^2] + \eta_l^2(6E_i\sigma^2+\sigma_{i,\tilde{g}}^2) + 6E_i\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{\theta}^e)\|^2] \tag{58}$$

From the inequality above and that $\mathbb{E}\|\boldsymbol{\theta}_{i,k}^e - \boldsymbol{\theta}^e\|^2 = 0$, we have:

$$\mathbb{E}\|\boldsymbol{\theta}_{i,1}^e - \boldsymbol{\theta}^e\|^2 \leq \gamma \tag{59}$$
$$\mathbb{E}\|\boldsymbol{\theta}_{i,2}^e - \boldsymbol{\theta}^e\|^2 \leq (1+\frac{1}{E_i})\gamma + \gamma \tag{60}$$
$$\mathbb{E}\|\boldsymbol{\theta}_{i,3}^e - \boldsymbol{\theta}^e\|^2 \leq (1+\frac{1}{E_i})^2\gamma + (1+\frac{1}{E_i})\gamma + \gamma \tag{61}$$
$$\dots$$
$$\mathbb{E}\|\boldsymbol{\theta}_{i,k}^e - \boldsymbol{\theta}^e\|^2 \leq (1+\frac{1}{E_i})^{(k-1)}\gamma + \ldots + (1+\frac{1}{E_i})^2\gamma + (1+\frac{1}{E_i})\gamma + \gamma, \tag{62}$$
$$\tag{63}$$

where $\gamma = \eta_l^2(6E_i\sigma^2 + \sigma_{i,\tilde{g}}^2) + 6E_i\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{\theta}^e)\|^2]$. By using $1 + q + \cdots + q^{n-1} = \frac{q^n-1}{q-1}$, we get:

$$\mathbb{E}\|\boldsymbol{\theta}_{i,k}^e - \boldsymbol{\theta}^e\|^2 \leq E_i\bigg(\big(1+\frac{1}{E_i}\big)^k - 1\bigg)\big(\eta_l^2(6E_i\sigma^2 + \sigma_{i,\tilde{g}}^2) + 6E_i\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{\theta}^e)\|^2]\big). \tag{64}$$

Therefore, we have:

$$\begin{aligned}\sum_{k=0}^{E_i-1}\mathbb{E}\|\boldsymbol{\theta}_{i,k}^e - \boldsymbol{\theta}^e\|^2 &\leq E_i^2\bigg(\underbrace{\big(1+\frac{1}{E_i}\big)^{E_i}}_{\leq \texttt{cte}} - 2\bigg)\big(\eta_l^2(6E_i\sigma^2 + \sigma_{i,\tilde{g}}^2) + 6E_i\eta_l^2\mathbb{E}[\|\nabla f(\boldsymbol{\theta}^e)\|^2]\big)\\ &\leq (\texttt{cte}-2)E_i^2\eta_l^2\big(6E_i\sigma^2 + \sigma_{i,\tilde{g}}^2 + 6E_i\mathbb{E}[\|\nabla f(\boldsymbol{\theta}^e)\|^2]\big),\end{aligned} \tag{65}$$

where $E_i \geq 2$ and $\texttt{cte}$ above is the mathematical constant e. □

We can now plug the bound on local drifts into eq. (52) and get:

$$\begin{aligned}\mathbb{E}\big[f(\boldsymbol{\theta}^{e+1})\big] \leq \mathbb{E}\big[f(\boldsymbol{\theta}^e)\big] &- \eta_l\bigg(\underbrace{\frac{11\bar{E}^e - 6}{12} - 12\beta^2\eta_l^2(1+\sum_{i=1}^n E_i)\big(\sum_{i=1}^n E_i^4\big)}_{\geq \frac{11\bar{E}^e-7}{12}}\bigg)\mathbb{E}\big[\|\nabla f(\boldsymbol{\theta}^e)\|^2\big]\\ &+ 6\beta^2\eta_l^3(1+\sum_{i=1}^n E_i)\bigg(2\sum_{i=1}^n E_i^4\sigma^2 + \frac{1}{3}\sum_{i=1}^n E_i^3\sigma_{i,\tilde{g}}^2\bigg) + \beta\eta_l^2\sum_{i=1}^n E_i^2\sigma_{i,\tilde{g}}^2\\ &+ \frac{8}{3}L_0^2\eta_l\bigg(n\sum_{i=1}^n E_i^2\mathbb{E}[(w_i^e - \lambda_i)^2] + \|\boldsymbol{\lambda}\|^2\sum_{i=1}^n\mathbb{E}[(E_i - \mu_w^e)^2]\bigg),\end{aligned} \tag{66}$$

where we have used the second condition on η_l in the first line to bound the multiplicative factor.

Hence, we have:

$$\begin{aligned}&\mathbb{E}\big[f(\boldsymbol{\theta}^{e+1})\big] \leq\\ &\mathbb{E}\big[f(\boldsymbol{\theta}^e)\big] - \eta_l\bigg(\frac{11\bar{E}^e - 7}{12}\bigg)\mathbb{E}\big[\|\nabla f(\boldsymbol{\theta}^e)\|^2\big]\\ &+ \underbrace{\eta_l\bigg(6\beta^2\eta_l^2(1+\sum_{i=1}^n E_i)\bigg(2\sum_{i=1}^n E_i^4\sigma^2 + \frac{1}{3}\sum_{i=1}^n E_i^3\sigma_{i,\tilde{g}}^2\bigg) + \beta\eta_l\sum_{i=1}^n E_i^2\sigma_{i,\tilde{g}}^2\bigg)}_{\Psi_\sigma^e}\\ &+ \underbrace{\eta_l\,\frac{8L-0^2}{3}\bigg(n\sum_{i=1}^n E_i^2\mathbb{E}[(w_i^e - \lambda_i)^2] + \|\boldsymbol{\lambda}\|^2\sum_{i=1}^n\mathbb{E}[(E_i - \mu_w^e)^2]\bigg)}_{\Psi_p^e}\end{aligned} \tag{67}$$

Therefore:

$$\eta_l(\frac{11\bar{E}^e - 7}{12})\mathbb{E}\|\nabla f(\boldsymbol{\theta}^e)\|^2 \leq \mathbb{E}[f(\boldsymbol{\theta}^e) - f(\boldsymbol{\theta}^{e+1})] + (\Psi_\sigma^e + \Psi_p^e)\eta_l. \tag{68}$$

We can now replace $\bar{E}^e$, which is a weighted average of $\{E_i\}_{i=1}^n$, with $E_l^{\min} = \min_i\{E_i\}_{i=1}^n$, and the inequality still holds:

$$\eta_l(\frac{11E_l^{\min} - 7}{12})\mathbb{E}\|\nabla f(\boldsymbol{\theta}^e)\|^2 \leq \mathbb{E}[f(\boldsymbol{\theta}^e) - f(\boldsymbol{\theta}^{e+1})] + (\Psi_p^e + \Psi_\sigma^e)\eta_l. \tag{69}$$

By summing both sides of the above inequality over $e = 0, \ldots, E-1$ and dividing by E, we get:

$$\min_{0\leq e\leq E-1}\mathbb{E}[\|\nabla f(\boldsymbol{\theta}^e)\|^2] \leq \frac{12}{(11E_l^{\min} - 7)}\bigg((\frac{f(\boldsymbol{\theta}^0) - f^*}{E\eta_l}) + \frac{\sum_{e=0}^{E-1}(\Psi_\sigma^e + \Psi_p^e)}{E}\bigg), \tag{70}$$

which completes the proof. □

E DETAILED RESULTS

E.1 UTILITY COMPARISON

Table 8: Comparison of different algorithms (on MNIST, $E = 200$). FedAvg achieves 98.6%.

alg \ distr	Dist1	Dist2	Dist3	Dist4	Dist5	Dist6	Dist7	Dist8	Dist9
WeiAvg [19]	90.08	88.29	89.74	88.20	84.94	81.60	84.43	78.71	81.38
PFA [19]	87.78	89.70	86.23	88.67	85.65	80.25	81.32	79.34	75.41
DPFedAvg [38]	84.24	82.84	83.50	80.43	83.02	74.02	**85.71**	70.58	80.49
minimum ϵ [19]	77.80	74.86	74.86	71.75	68.42	9.61	77.62	56.10	68.44
Robust-HDP	**90.09**	**90.71**	**89.78**	**89.38**	**87.52**	**85.13**	84.03	**81.19**	**81.52**

Table 9: Comparison of different algorithms (on FMNIST, $E = 200$). FedAvg achieves 90.28%.

alg \ distr	Dist1	Dist2	Dist3	Dist4	Dist5	Dist6	Dist7	Dist8	Dist9
WeiAvg [19]	**77.65**	**78.30**	**75.92**	**77.10**	72.38	64.15	66.80	66.86	64.79
PFA [19]	69.95	75.49	62.17	74.51	61.18	71.27	63.93	60.77	54.66
DPFedAvg [38]	74.12	71.68	71.97	68.10	70.20	62.46	64.15	65.87	65.50
minimum ϵ [19]	73.15	64.26	64.26	62.60	64.35	28.66	65.13	58.44	66.36
Robust-HDP	75.13	76.25	75.04	76.19	**73.80**	**71.30**	**66.85**	**68.32**	**66.96**

E.2 NOISE VARIANCE AFTER AGGREGATION

Table 10: Comparison of different heterogeneous DPFL algorithms (on MNIST with $E = 200$) in terms of the average noise power (eq. (7) and eq. (3)) in each parameter normalized by their corresponding used learning rate ($\frac{\sum_{i=1}^{n} w_i^{e2} \sigma_i^2}{p \eta_l^2}$) in the aggregated model update ($\sum_{i=1}^{n} w_i^e \Delta \tilde{\boldsymbol{\theta}}_i^e$). Due to the projection used in PFA, computing its noise after aggregation was not possible.

alg \ dist	Dist1	Dist2	Dist3	Dist4	Dist5	Dist6	Dist7	Dist8	Dist9
WeiAvg [19]	1.02	1.89	0.92	3.22	4.58	28.29	9.85	48.15	34.91
DPFedAvg [38]	1.27	16.94	16.28	26.87	25.64	70.71	18.50	85.70	43.20
minimum ϵ [19]	4.68	103.91	103.91	127.18	103.91	1868.45	74.41	241.37	87.15
Robust-HDP	0.267	0.473	0.074	0.642	0.385	7.616	2.252	13.855	5.937
Optimum	0.267	0.473	0.074	0.641	0.385	7.601	2.251	13.812	5.927

E.3 PRECISION OF Robust-HDP

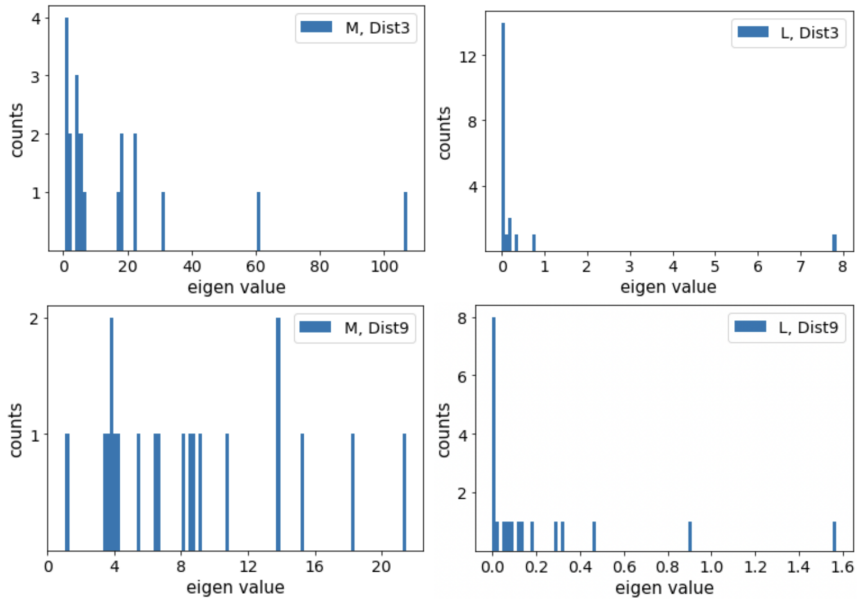

Figure 8: Comparison of the eigen values of matrices $\mathbf{M}$ (left) and $\mathbf{L}$ (right) on MNIST dataset. The concentration of eigen values in the right figures around 0 shows that the matrix L returned by Robust PCA, is indeed low rank, while M is not, which is due to the noise existing in clients model updates.

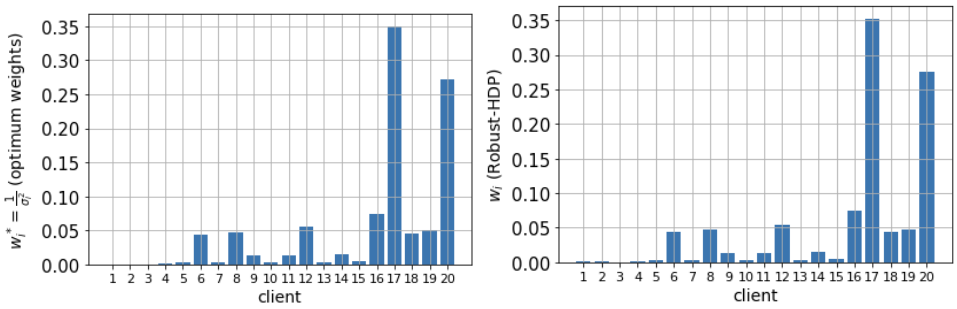

Figure 9: Weight assignment of optimum strategy (eq. (8)) (left) and Robust-HDP (right) for CIFAR10 and Dist2.

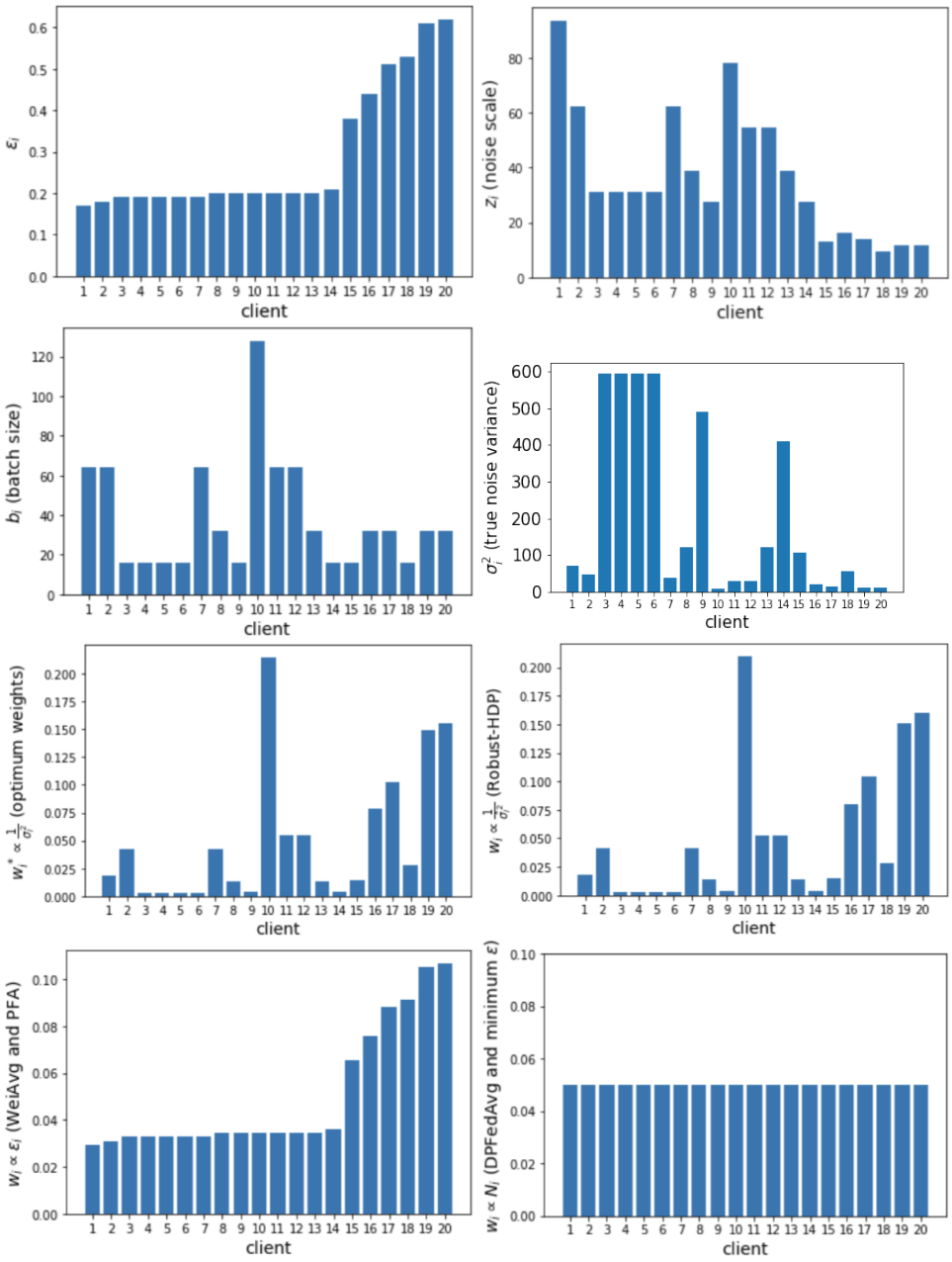

Figure 10: Comparison of weight assignments for Dist8 and MNIST dataset.