**Organization.** These appendices are organized as follows.

(A) In Appendix A, we prove Theorem 1.

(B) In Appendix B, we prove Theorem 2.

(C) In Appendix C, we show that batch partitioning is necessary to satisfy the multi-round privacy definition given in (5).

(D) In Appendix D, we provide the two components of Multi-RoundSecAgg which are Algorithm 1 and Algorithm 2.

(E) Appendix G provides additional experiments with various system parameters.

(F) Appendix E provides additional experiments on the MNIST dataset.

(G) Appendix F provides additional details and the hyperparameters of the experiments of Section 6 and Appendix E.

(H) In Appendix H, we theoretically show that the random selection strategy discussed in Remark 2 that aims to select $K$ available users at each round and the random selection strategy that selects the users in i.i.d fashion both have a multi-round privacy $T = 1$ with high probability. We also empirically demonstrate that the local models can be reconstructed accurately when random selection is used.

(I) Finally, in Appendix I, we consider the convergence rate of the general convex and the non-convex cases.

We list the notations in Table 2.

**Table 2:** Notations occurred in the paper.

| Notations | Description |
|---|---|
| $N$ | total number of users |
| $K$ | number of users selected at each iteration |
| $J$ | total number of iterations |
| $E$ | number of local iterations in each user |
| $d$ | dimension of model |
| $\boldsymbol{x}^{(t)}$ | global model at iteration $t$, $\boldsymbol{x}^{(t)} \in \mathbb{R}^d$ |
| $\boldsymbol{x}_i^{(t)}$ | local model of user $i$ at iteration $t$, $\boldsymbol{x}_i^{(t)} \in \mathbb{R}^d$ |
| $\mathbf{X}^{(t)}$ | concatenation of the weighted local models at iteration $t$, $\mathbf{X}^{(t)} \in \mathbb{R}^{N\times d}$ |
| $\boldsymbol{p}^{(t)}$ | participation vector at iteration $t$, $\boldsymbol{p}^{(t)} \in \{0,1\}^N$ |
| $\mathbf{P}^{(t)}$ | participation matrix, $\mathbf{P}^{(t)} \in \{0,1\}^{t\times N}$ |
| $T$ | multi-round privacy guarantee |
| $F$ | aggregation fairness gap |
| $C$ | average aggregation cardinality |
| $\mathbf{B}$ | privacy-preserving family, $\mathbf{B} \in \{0,1\}^{R_{BP}\times N}$ |
| $R_{BP}$ | the size of the privacy-preserving family of sets |
| $\mathcal{U}^{(t)}$ | set of available users at iteration $t$ |
| $p_i$ | dropout probability of user $i$ |
| $f_i^{(t)}$ | frequency of participation of user $i$ before round $t$ |

# A Theoretical Guarantees of Multi-RoundSecAgg: Proof of Theorem 1

In this appendix, we provide the proof of Theorem 1.

*Proof.* 1. First, we prove that Multi-RoundSecAgg ensures a multi-round privacy of $T$. We first partition the matrix $\mathbf{B}$ into $R \times T$ matrices as $\mathbf{B} = [\mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \cdots, \mathbf{B}^{(N/T)}]$ and the aggregated

models as $\mathbf{X} = [\mathbf{X}^{(1)\top}, \mathbf{X}^{(2)\top}, \cdots, \mathbf{X}^{(N/T)\top}]^\top$. We can then express any linear combination of the aggregated models $\mathbf{X}^\top \mathbf{B}^\top z$, where $z \in \mathbb{R}^R \setminus \{\mathbf{0}\}$, as follows

$$\mathbf{X}^\top \mathbf{B}^\top z = \sum_{i=1}^{N/T} \mathbf{X}^{(i)\top} \mathbf{B}^{(i)\top} z. \tag{9}$$

Denote the $j$-th column of $\mathbf{B}^{(i)}$ by $\mathbf{b}_j^{(i)}$ which is either a zero vector or all ones vector due to the batch partitioning structure. That is, $\mathbf{b}_j^{(i)} \in \{\mathbf{0}, \mathbf{1}\}$. Hence, $\mathbf{B}^{(i)\top} z \in \{\mathbf{0}, a_i.\mathbf{1}\}$ for some $a_i \in \mathbb{R} \setminus \{0\}$. Therefore, we have

$$\mathbf{X}^{(i)\top} \mathbf{B}^{(i)\top} z = \begin{cases} \mathbf{0} & \mathbf{B}^{(i)\top} z = 0, \\ a_i \sum\limits_{j=(i-1)T+1}^{iT} \boldsymbol{x}_j & \text{otherwise}, \end{cases} \tag{10}$$

$\forall i \in [N/T]$, which shows that Multi-RoundSecAgg achieves a multi-round privacy $T$.

2. Next, we prove that Multi-RoundSecAgg has an aggregation fairness gap $F = 0$.

   It is clear that the total number of times user $i$ is being selected up to time $J$ is the same as that of user $j$ who lies in the same batch as user $i$. This follows since all users in the same batch either participate together or they do not participate at all.

   It suffices to show that the *expected* number of selections of user $i$ up to time $J$ is the same as that of user $j$, where user $i$ and user $j$ are in different batches. The main observation is that our protocol is *symmetric*. Indeed, the only randomness in the system are the user availability randomness and the set selection randomness when there are multiple user sets satisfying the requirements. We note that for any realization of random variables such that the batch of user $i$ is selected at time $t$, there is a corresponding realization of random variables such that the batch of user $j$ is selected at time $t$ and all other selections remain exactly the same. Hence, $F_i = F_j$ for any $i \neq j$.

3. Finally, we characterize the average aggregation cardinality of Multi-RoundSecAgg. The average aggregation cardinality can be expressed as follows

$$\begin{aligned} C &= K\,(1 - \Pr[\text{No row of } \mathbf{B} \text{ is available}]) \\ &= K\left(1 - \Pr[\text{At least } \frac{N}{T} - \frac{K}{T} + 1 \text{ batches are not available}]\right) \\ &= K\left(1 - \sum_{i=N/T-K/T+1}^{N/T} \binom{N/T}{i} q^i (1-q)^{N/T-i}\right), \end{aligned} \tag{11}$$

   where $q$ is the probability that a certain batch is not available, which is given by $q = 1 - (1-p)^T$.

□

# B Convergence Analysis of Multi-RoundSecAgg : Proof of Theorem 2

The proof of Theorem 2 is divided into two parts. In the first part, we introduce a new sequence to represent the local updates in each user with respect to step index while we use the global round index $t$ for $\boldsymbol{x}^{(t)}$ in (2). We carefully define the sequence and the step index, and then provide the convergence analysis of the sequence. In the second part, we bridge the newly defined sequence and $\boldsymbol{x}^{(t)}$ in (2), and provide convergence analysis of $\boldsymbol{x}^{(t)}$.

**First Part (Convergence analysis of local model updates)**.

Let $\boldsymbol{w}_i^{(j)}$ be the local model updated by user $i$ at the $j$-th step. Note that this step index is different from the global round index $t$ in (2) as each user updates the local model by carrying out $E(\geq 1)$ local SGD steps before sending the results to the server. Let $\mathcal{I}_E$ be the set of global synchronization steps, i.e., $\mathcal{I}_E = \{nE | n = 0, 1, 2, \ldots\}$. Importantly, we define the step index $j$ such it increases from $nE$ to $nE + 1$ only when the server does not skip the selection, i.e., there are at least $K$ available users

at step $nE+1$ for $n \in \{0,1,2,\ldots\}$. We denote by $\mathcal{H}_{nE}$ the set selected by Multi-RoundSecAgg at step index $nE$ and from the definition, $|\mathcal{H}_{nE}| = K$ for all $n \in \{0,1,2,\ldots\}$. Then, the update equation can be described as

$$v_i^{j+1} = w_i^j - \eta^j \nabla L_i \left( w_i^j, \xi_i^j \right), \tag{12}$$

$$w_i^{j+1} = \begin{cases} v_i^{j+1} & \text{if} \quad j+1 \in \mathcal{I}_E \\ \frac{1}{K} \sum_{k \in \mathcal{H}_{j+1}} v_k^{j+1} & \text{if} \quad j+1 \notin \mathcal{I}_E \end{cases}, \tag{13}$$

where we introduce an additional variable $v_i^{j+1}$ to represent the immediate result of one step SGD from $w_i^j$. We can view $w_i^{j+1}$ as the model obtained after aggregation step (when $j+1$ is a global synchronization step). Motivated by [33, 23], we define two virtual sequences

$$\overline{v}^j = \frac{1}{N} \sum_{i=1}^N v_i^j, \tag{14}$$

$$\overline{w}^j = \frac{1}{N} \sum_{i=1}^N w_i^j. \tag{15}$$

We can interpret $\overline{v}^{j+1}$ as the result of single step SGD from $\overline{w}^j$. When $j \notin \mathcal{I}_E$, both $\overline{v}^j$ and $\overline{w}^j$ are not accessible. We also define $\overline{g}^j = \frac{1}{N} \sum_{i=1}^N \nabla L_i \left( w_i^j \right)$ and $g^j = \frac{1}{N} \sum_{i=1}^N \nabla L_i \left( w_i^j, \xi_i^j \right)$. Then, $\overline{v}^{j+1} = \overline{w}^j - \eta^j g^j$.

Now, we state our two key lemmas.

**Lemma 1** (Unbiased selection)**.** *When $j+1 \in \mathcal{I}_E$, the following is satisfied,*

$$\mathbb{E}_{\mathcal{H}_{j+1}} [\overline{w}^{j+1}] = \overline{v}^{j+1}. \tag{16}$$

*Proof.* Let $\mathcal{H}_{j+1} = \{i_1, \ldots, i_K\}$. Then, we have

$$\begin{aligned} \mathbb{E}_{\mathcal{H}_{j+1}} [\overline{w}^{j+1}] = \frac{1}{K} E_{\mathcal{H}_{j+1}} \left[ \sum_{k \in \mathcal{H}_{j+1}} v_k^{j+1} \right] = \frac{1}{K} E_{\mathcal{H}_{j+1}} \left[ \sum_{k=1}^K v_{i_k}^{j+1} \right] = E_{\mathcal{H}_{j+1}} [v_{i_k}^{j+1}] \\ = \sum_{k=1}^N \frac{1}{N} v_k^{j+1} = \overline{v}^{j+1} \end{aligned} \tag{17}$$

where (17) follows as $\Pr[i_k = j] = \frac{1}{N}$ for $i \in [N]$. This is because the sampling probability of each user is identical due to the symmetry in the construction and the fact that all users have the same dropout probability. □

Now, we provide the convergence analysis of the sequence $\overline{w}^j$ defined in (15). We have,

$$\begin{aligned} \|\overline{w}^{j+1} - w^*\|^2 &= \|\overline{w}^{j+1} - \overline{v}^{j+1} + \overline{v}^{j+1} - w^*\|^2 \\ &= \|\overline{w}^{j+1} - \overline{v}^{j+1}\|^2 + \|\overline{v}^{j+1} - w^*\|^2 + 2 \left( \overline{w}^{j+1} - \overline{v}^{j+1} \right)^\top \left( \overline{v}^{j+1} - w^* \right). \end{aligned} \tag{18}$$

When the expectation is taken over $\mathcal{H}_{j+1}$, the last term in (18) becomes zero due to Lemma 1. For the second term in (18), we have

$$\|\overline{v}^{j+1} - w^*\|^2 \leq (1 - \eta^j \mu) \|\overline{w}^j - w^*\|^2 + \alpha (\eta^j)^2, \tag{19}$$

where $\alpha = \frac{1}{N} \sum_{i=1}^N \sigma_i^2 + 6\rho\Gamma + 8(E-1)^2 G^2$ and (19) directly follows from Lemma 1, 2, 3 of [23]. The first term in (18) becomes zero if $j+1 \in \mathcal{I}_E$, and if $j+1 \notin \mathcal{I}_E$, from Lemma 5 of [23], it is bounded by

$$\mathbb{E}_{\mathcal{H}_{j+1}} \|\overline{w}^{j+1} - \overline{v}^{j+1}\|^2 \leq \beta (\eta^j)^2, \tag{20}$$

where $\beta = \frac{4(N-K)E^2G^2}{K(N-1)}$. By combining (18) to (20), we have

$$\mathbb{E} \|\overline{w}^{j+1} - w^*\|^2 \leq (1 - \eta^j \mu) \|\overline{w}^j - w^*\|^2 + (\alpha + \beta)(\eta^j)^2. \tag{21}$$

Then by utilizing the similar induction in [23], we can show that

$$\mathbb{E}\|\overline{\boldsymbol{w}}^{j+1} - \boldsymbol{w}^*\|^2 \leq \frac{1}{\gamma + t - 1}\left(\frac{4(\alpha+\beta)}{\mu^2} + \gamma\mathbb{E}\|\overline{\boldsymbol{w}}^0 - \boldsymbol{w}^*\|^2\right), \tag{22}$$

where $\gamma = \max\left\{\frac{8\rho}{\mu}, E\right\}$. By combining (22) with $\rho$-smoothness of the global loss function in (1), we have

$$\mathbb{E}[L(\overline{\boldsymbol{w}}^I)] - L^* \leq \frac{\rho}{\gamma + I - 1}\left(\frac{2(\alpha+\beta)}{\mu^2} + \frac{\gamma}{2}\mathbb{E}\|\overline{\boldsymbol{w}}^0 - \boldsymbol{x}^*\|^2\right). \tag{23}$$

**Second Part (Convergence analysis of global model)**.

Now, we bridge the sequence $\overline{\boldsymbol{w}}^T$ and $\boldsymbol{x}^{(t)}$ in (2) to provide the convergence analysis of $\boldsymbol{x}^{(t)}$. Since we define the step index $j$ such that $j$ increases from $nE$ to $nE+1$ only when the server does not skip the selection, we have

$$\mathbb{E}[L(\boldsymbol{x}^{(J)})] = \mathbb{E}[L(\overline{\boldsymbol{w}}^{(JE\phi)})] \tag{24}$$

where $\phi$ is the probability that there are at least $K$ available users at a certain synchronization step, and $\phi = \frac{C}{K}$ due to the fact that $C = K \cdot \Pr[\text{at least one row of } \mathbf{B} \text{ is available}] = K\phi$. By combining (23) and (24), we have that,

$$\mathbb{E}[L(\boldsymbol{x}^{(J)})] - L^* \leq \frac{\rho}{\gamma + \frac{C}{K}EJ - 1}\left(\frac{2(\alpha+\beta)}{\mu^2} + \frac{\gamma}{2}\mathbb{E}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2\right), \tag{25}$$

which completes the proof.

# C Necessity of Batch Partitioning (BP)

In this appendix, we show that batch partitioning is necessary to satisfy the multi-round privacy guarantee of Equation (5) and our strategy is optimal in the sense that no other strategy can have more distinct user selection sets than our strategy.

*Proof.* Consider any scheme which selects sets from an $R \times N$ matrix $\mathbf{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N]^\top$. Denote the linear coefficients multiplying them by $z_i, i \in [R]$. Then, the $i$-th element of $\mathbf{V}^\top \boldsymbol{z}$ is given by

$$\{\mathbf{V}^\top \boldsymbol{z}\}_i = \sum_{j \in \mathrm{supp}(\boldsymbol{v}_i)} z_i. \tag{26}$$

We now claim that we can cluster the entries using equivalence of linear functions to groups, where each group must have a size of at least $T$ except for the group corresponding to the zero function. To show this, we choose each $z_i \overset{\text{i.i.d.}}{\sim} U[0,1]$, and the key observation is that if two entries have different linear functions then their final value after this assignment would be different with probability one. Since the scheme satisfies a multi-round privacy $T$, this implies that for each non-zero linear function of the form of Equation (26), there must be at least $T$ of them. If we group the entries according to the equivalence of linear functions, we get at most $N/T$ groups (ignoring the group of constant zero).

Then, we show that the total number of possible sets $R$ is upper-bounded by $\binom{N/T}{K/T}$. We observe that the total number of non-zero groups we can choose for each vector is at most $K/T$ due to the size of each group, so the total number of distinct vectors satisfying the weight requirement is at most

$$R \leq R_{\max} \overset{\text{def}}{=} \binom{D}{E}, \tag{27}$$

where $D \leq N/T$ is the total number of groups corresponding to the non-zero linear functions, and $E \leq K/T$ is the total number of groups we may select in each round. Next, we have

$$\begin{aligned} R_{\max} &= \binom{D}{E} \\ &\overset{(i)}{\leq} \binom{N/T}{E} \\ &\overset{(ii)}{\leq} \binom{N/T}{K/T} = R_{\mathrm{BP}}, \end{aligned} \tag{28}$$

where $(i)$ follows since $\binom{D}{E}$ is monotonically increasing w.r.t $D$, and $(ii)$ follows as $\binom{D}{E}$ is monotonically increasing w.r.t $E$ if $E \leq D/2$. □

# D The Two Components of Multi-RoundSecAgg : Algorithms 1 and 2

**Algorithm 1** Batch Partitioning Privacy-preserving Family Generation

**Input:** Number of users $N$, row weight $K$ and the desired multi-round privacy guarantee $T$.
**Output:** Privacy-preserving Family $\mathbf{B} \in \{0,1\}^{R_{\text{BP}} \times N}$, where $R_{\text{BP}} = \binom{N/T}{K/T}$
**Initialization:** $\mathbf{B} = \mathbf{0}_{R_{\text{BP}} \times N}$.

1: Partition index sets $\{1, 2, \ldots, N\}$ into $\frac{N}{T}$ sets, $\mathcal{G}_1, \ldots, \mathcal{G}_{\frac{N}{T}}$, where $|\mathcal{G}_i| = T$ for all $i \in [\frac{N}{T}]$.
2: Generate all possible sets each of which is union of $\frac{K}{T}$ sets out of $\frac{N}{T}$ sets $(\mathcal{G}_1, \ldots, \mathcal{G}_{\frac{N}{T}})$ without replacement. Denote the generated sets by $\mathcal{L}_1, \ldots, \mathcal{L}_{R_{\text{BP}}}$.
3: **for** $i = 1, 2, \ldots, R_{\text{BP}}$ **do**
4:  **for** $j = 1, 2, \ldots, N$ **do**
5:  **if** $j \in \mathcal{L}_i$ **then** $\{\boldsymbol{b}_i\}_j = 1$

**Algorithm 2** Available Batch Selection

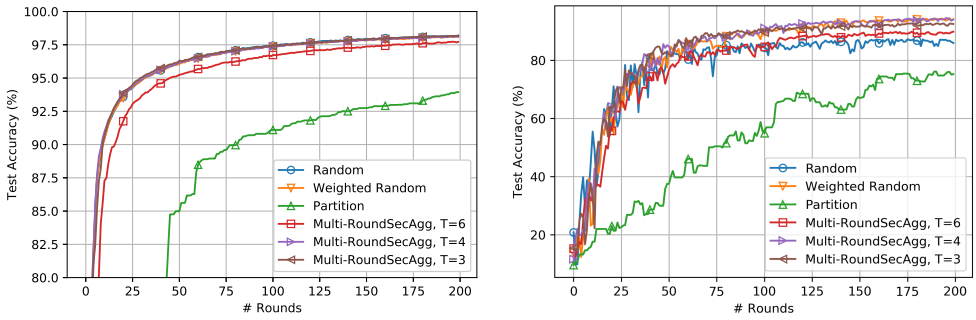
**Input:** A family of sets $\mathbf{B}$, set of available users $\mathcal{U}^{(t)}$, the frequency of participation vector $\boldsymbol{f}^{(t-1)}$, and the selection mode $\lambda$. ▷ $\lambda = 0$ when $p_i = p, \forall i \in [N]$ and 1 otherwise
**Output:** A participation vector $\boldsymbol{p}^{(t)}$.
**Initialization:** $\mathbf{B}^{(t)} = [\,]$, $\ell_{\min}^{(t-1)} := \arg\min_{i \in \mathcal{U}^{(t)}} f_i^{(t-1)}$.

1: **for** $i = 1, 2, \ldots, R_{\text{BP}}$ **do**
2:  **if** $\text{supp}(\boldsymbol{b}_i) \subseteq \mathcal{U}^{(t)}$ **then** $\mathbf{B}^{(t)} = [\mathbf{B}^{(\mathbf{t})\top}, \boldsymbol{b}_i]^\top$.
3: **if** $\mathbf{B}^{(t)} = [\,]$ **then**
4:  $\boldsymbol{b}_{r(t)}^{(t)} = \mathbf{0}$.
5: **else if** $\lambda = 0$ **then** ▷ Uniform selection
6:  Select a row from $\mathbf{B}^{(t)}$, $\boldsymbol{b}_{r(t)}^{(t)}$, uniformly at random.
7: **else** ▷ Fairness-aware selection
8:  Select a row from $\mathbf{B}^{(t)}$, $\boldsymbol{b}_{r(t)}^{(t)}$, uniformly at random from the rows that include $\ell_{\min}^{(t-1)}$.
9: $\boldsymbol{p}^{(t)} = \boldsymbol{b}_{r(t)}^{(t)}$.
10: Update $\boldsymbol{f}^{(t)} = \boldsymbol{f}^{(t-1)} + \boldsymbol{p}^{(t)}$

# E Additional Experiments: MNIST dataset

**MNIST.** To further investigate the performance of Multi-RoundSecAgg, we implement a simple CNN [24] with two $5 \times 5$ convolution layers, a fully connected layer with ReLU activation, and a final Softmax output layer. This standard model has 1,663,370 parameters and is sufficient for our needs, as our goal is to evaluate various schemes, not to achieve the best accuracy. We study the two settings for partitioning the MNIST dataset across the users.

- **IID Setting.** In this setting, the 60000 training samples are shuffled and partitioned uniformly across the $N = 120$ users, where each user receives 500 samples.
- **Non-IID Setting.** In this setting, we first sort the dataset by the digit labels, partition the sorted dataset into 120 shards of size 500, and assign each of the 120 users one shard. This is similar to the pathological non-IID partitioning setup proposed in [24], where our partition is an extreme case as each user has only one digit label while each user in [24] has two.

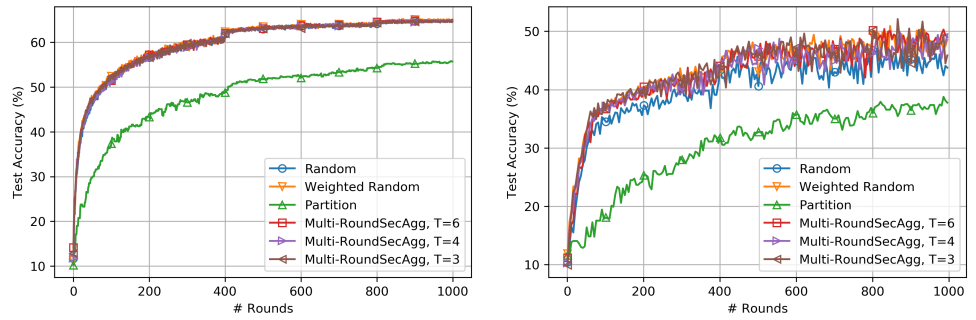



(a) IID data distribution. (b) Non-IID data distribution.

**Figure 7:** Training rounds versus test accuracy of CNN in [24] on the MNIST with $N = 120$ and $K = 12$.

**CIFAR-10** We also consider both IID and Non-IID distribution, and implement LeNet [22] for both setting. While the state-of-the-art models [19, 34] achieve 99% accuracy, LeNet is sufficient for our needs, as our goal is to evaluate various schemes, not to achieve the best accuracy.

- **IID Setting.** In this setting, the 50000 training samples are shuffled and partitioned uniformly across the $N = 120$ users, where each user receives 417 or 416 samples.
- **Non-IID dataset.** In this setting, we utilize the *data-sharing strategy* of [41], where the 50000 training samples are divided into a globally shared dataset $\mathcal{G}$ and private dataset $\mathcal{D}$. We set $|\mathcal{G}| = 200$ and $|\mathcal{D}| = 49800$. Then, we sort $\mathcal{D}$ by the labels, partition it into 120 shards of size 415, and assign each of the 120 users one shard. Each user has 200 samples of globally shared data and 415 samples of private dataset with one label.




(a) IID data distribution. (b) Non-IID data distribution.

**Figure 8:** Training rounds versus test accuracy of LeNet in [22] on the CIFAR-10 with $N = 120$ and $K = 12$.

We measure the test accuracy of the six schemes on the MNIST and CIFAR-10 dataset with the two distribution settings, the IID and the Non-IID. Our results are demonstrated in Figure 7 and Figure 8. We make the following key observations, which are similar to the observations on the CIFAR-100 dataset.

- In the IID setting, the Multi-RoundSecAgg schemes show comparable test accuracy to the random selection and random weighted selection schemes while the Multi-RoundSecAgg schemes provide better multi-round privacy guarantee $T$.
- In the non-IID setting, the Multi-RoundSecAgg schemes outperform the random selection scheme while showing comparable test accuracy to the weighted random selection scheme. This is because Multi-RoundSecAgg schemes have better aggregation fairness gaps as demonstrated in Figure 4(b), which results in better test accuracy in the non-IID setting.
- In both IID and non-IID settings, the user partitioning scheme has the worst test accuracy as its average aggregation cardinality is much smaller than the other schemes.

**Table 3:** Test accuracy of VGG11 in [29] on the CIFAR-100 dataset with $N = 120$ and $K = 12$.

| Scheme | IID Setting | Non-IID Setting |
|---|---|---|
| Random selection | 49.15% | 44.32% |
| Weighted random selection | 50.06% | 47.11% |
| User partition | 25.73% | 22.32% |
| Multi-RoundSecAgg, T=6 | 42.89% | 39.57% |
| Multi-RoundSecAgg, T=4 | 49.43% | 46.99% |
| Multi-RoundSecAgg, T=3 | 50.22% | 47.06% |

**Table 4:** Test accuracy of LeNet in [22] on the CIFAR-10 dataset with $N = 120$ and $K = 12$.

| Scheme | IID Setting | Non-IID Setting |
|---|---|---|
| Random selection | 64.64% | 45.20% |
| Weighted random selection | 65.06% | 47.89% |
| User partition | 55.70% | 37.74% |
| Multi-RoundSecAgg, T=6 | 65.01% | 46.35% |
| Multi-RoundSecAgg, T=4 | 64.95% | 47.00% |
| Multi-RoundSecAgg, T=3 | 64.80% | 47.21% |

**Table 5:** Test accuracy of the CNN in [24] on the MNIST dataset with $N = 120$ and $K = 12$.

| Scheme | IID Setting | Non-IID Setting |
|---|---|---|
| Random selection | 98.21% | 85.79% |
| Weighted random selection | 98.10% | 94.04% |
| User partition | 93.94% | 75.26% |
| Multi-RoundSecAgg, T=6 | 97.72% | 89.88% |
| Multi-RoundSecAgg, T=4 | 98.11% | 92.51% |
| Multi-RoundSecAgg, T=3 | 98.15% | 94.16% |

# F Experiment Details

In this section, we provide more details about the experiments of Section 6 and Appendix E.

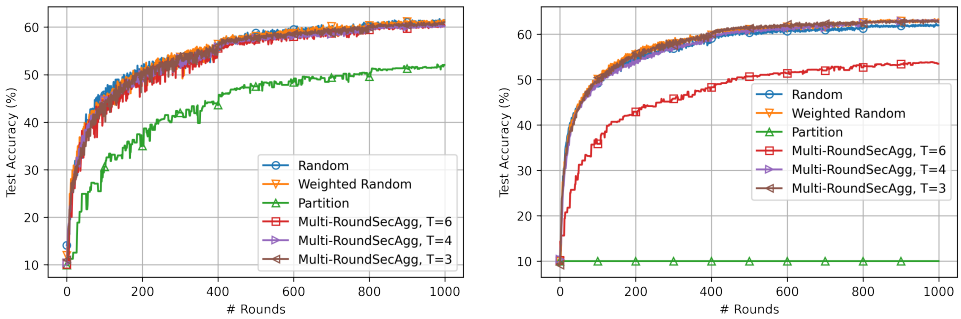
We summarize the test accuracy of CIFAR-100, CIFAR-10, and MNIST dataset in Table 3, Table 4 and Table 5, respectively. For all datasets, we run experiments five times with different random seeds and present the average value of the test accuracy in Table 4 and Table 5.

**Hyperparameters and computing resources.** For a fair comparison between 6 schemes, we find the best learning rate from $\{0.1, 0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001\}$. Given the choice of the best learning rate $\eta$, $\eta$ is decayed to $0.4\eta$ every 400 and 800 rounds to train the LeNet on the CIFAR-10 dataset or train VGG11 on the CIFAR-100 dataset while $\eta$ is not decayed in the CNN on the MNIST dataset. To train the LeNet on the CIFAR-10 dataset or train VGG11 on the CIFAR-100 dataset, we use the mini-batch size of 50 and $E = 1$ local epoch for both IID and Non-IID settings. To train the CNN on the MNIST dataset, we use the mini-batch size of 100 and $E = 1$ local epoch for both IID and Non-IID settings. All experiments are conducted with users equipped with 3.4 GHz 4 cores i-7 Intel CPU and NVIDIA Geforce 1080, and the users communicate amongst each other through Ethernet to transfer the model parameters.

# G Additional Experiments: Ablation Study

In this Appendix, we further investigate the performance of Multi-RoundSecAggwith various settings of the system design parameters, the number of total users($N$), the number of selected users per round ($K$), and target multi-round privacy guarantee($T$). We use the same dropout model as Section 6, i.e., considering heterogeneous environments where users have different dropout probability among

$\{0.1, 0.2, 0.3, 0.4, 0.5\}$. We implement LeNet [22] for image classification for CIFAR-10 with IID distribution.




(a) $(N, K) = (240, 12)$ (b) $(N, K) = (120, 24)$

**Figure 9:** Training rounds versus test accuracy of LeNet [22] on the CIFAR-10 with various system parameters $(N, K, T)$.

Figure 9(a) and Figure 9(b) show the performance comparison with $(N, K) = (240, 12)$ and $(N, K) = (120, 24)$, respectively. Similar to Section 6 and Appendix E, we can observe that Multi-RoundSecAgg schemes show comparable test accuracy to the random and weighted random selection schemes while the Multi-RoundSecAgg provide better multi-round privacy guarantee $T$, and the user partitioning scheme has the worst test accuracy as its average aggregation cardinality is much smaller than the other schemes. In particular, when $(N, K) = (120, 24)$, the user partition scheme fails to train the model as the probability that all partitions are not available at each round becomes almost one.

# H Multi-round Privacy Analysis of the Conventional Random User Selection Strategies

In this appendix, we first theoretically study the multi-round privacy of two random user selection strategies, and show that they have a very weak multi-round privacy of $T = 1$ with high probability (for the case where $p_i = p, \forall i \in [N]$). Furthermore, we also provide additional experiments showing that the server can reconstruct the local updates of all users with high accuracy when a random selection strategy is used. In the theoretical analysis, to simplify the problem, we assume that the model of the users have converged and don't change from one round to the next. However, in the experiments, we empirically evaluate the error in approximating the individual models of the users (via least-squares error estimation), and show that the server can approximate individual updates with very small error.

## H.1 Theoretical Analysis of the Random Selection Strategies

We start by our theoretical results, where we consider the following two random selection schemes.

1. **$K$-uniform Random Selection**. In this scheme, at round $t$, $K$ users are selected uniformly at random from the set of available users $\mathcal{U}^{(t)}$ if $|\mathcal{U}^{(t)}| \geq K$. Otherwise, the server skips this round.
2. **I.I.D Random Selection**. In this scheme, at round $t$, each user is selected with probability $\frac{K}{N(1-p)}$ independently from the other available users, where $K < N(1-p)$. Hence, the expected number of selected users at each round is $K$ user.

For both schemes, we show that the server can reconstruct all individual models after $N$ rounds in the worst-case scenario (assuming that the models do not change over $N$ rounds). Specifically, we show that the participation matrices in both schemes have full rank with high probability after $N$ rounds. This, in turn, implies that the server can reconstruct all local models after $N$ rounds with high probability in both schemes. We provide our results formally next in Theorem 3.

**Theorem 3.** *(Random selection schemes have a multi-round privacy guarantee $T = 1$).*

1. *Consider the $K$-uniform random selection scheme, where* $\min(K, N-K) \geq cN$. *In this scheme, the server can reconstruct all individual models of the $N$ users after $N$ rounds with probability at least*

$$1 - 2e^{-c'N}, \tag{29}$$

*for some constant $c' > 0$ that depends on $c$.*

2. *Consider the i.i.d random selection scheme, where the users are selected according to Bern($\frac{K}{N(1-p)}$) distribution and let $t = K/N$. In this scheme, the server can reconstruct the individual models of the $N$ users after $N$ rounds with probability at least*

$$1 - 2N(1-t)^N - (1+o_N(1))N(N-1)(t^2+(1-t)^2)^N, \tag{30}$$

*which converges to 1 exponentially fast if $t \in (0, 1/2)$ is a fixed constant.*

*Proof.* We first note that if the participation matrix has full rank after $N$ rounds, then the server can reconstruct the model of each individual user. Hence, we analyze the probability of the $N \times N$ participation matrix being full rank. We now consider each scheme separately.

1. In the $K$-uniform random selection scheme, the probability that the participation matrix after $N$ rounds $\mathbf{P}^{(N)}$ has full rank is lower-bounded as follows [36], when $\min(K, N-K) \geq cN$,

$$\Pr[\mathbf{P}^{(N)} \text{ has full rank}] \geq 1 - 2e^{-c'N},$$

for some constant $c' > 0$ that depends on $c$. Hence, it follows that the server can reconstruct all individual models with probability at least $1 - 2e^{-c'N}$.

2. In the i.i.d random selection scheme, the probability that the participation matrix after $N$ rounds $\mathbf{P}^{(N)}$ has full rank is lower-bounded as follows [14]

$$\Pr[\mathbf{P}^{(N)} \text{ has full rank}] \geq 1 - 2N(1-t)^N - (1+o_N(1))N(N-1)(t^2+(1-t)^2)^N,$$

which converges to 1 exponentially fast if $t = K/N \in (0, 1/2)$ is a fixed constant. Hence, it follows that the probability the server can reconstruct all individual models is lower-bounded by the same probability.

□

**Remark 12.** Our experimental results in Section 6 also show that the multi-round privacy guarantee of the $K$-uniform random selection scheme goes to 1 after almost $N$ rounds as shown in Fig. 4(a).

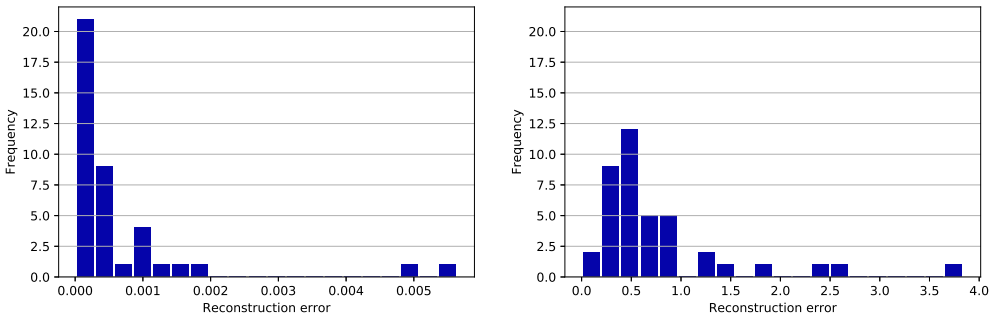
### H.2 Experimental Results

We now empirically evaluate the error in approximating the individual gradients of the users (via least-squares error estimation), and show that the server can approximate individual gradients of all users with a very small error when $K$-uniform random selection is used. To do so, we implement a reconstruction algorithm utilizing the least-squares method, and measure the $L_2$ distance between the true gradients and reconstructed gradients. We consider a FL setting with $N = 40$ users, where the server aims to choose $K = 8$ users at every round, to train the LeNet in [22] on the CIFAR-10 dataset with Non-IID setting, which is the same as the setting in Appendix E.

Let $\delta_i^{(t)}$ be the gradient of user $i$ at round $t$, i.e., $\delta_i^{(t)} = \boldsymbol{x}_i^{(t)} - \boldsymbol{x}^{(t)}$, and $\delta^{(t)}$ be the global update at round $t$, i.e., $\delta^{(t)} = \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)} = {\boldsymbol{\Delta}^{(t)}_{\text{individual}}}^{\top} \boldsymbol{p}^{(t)}$ where $\boldsymbol{\Delta}^{(t)}_{\text{individual}} = \left[w_1\delta_1^{(t)}, \ldots, w_N\delta_N^{(t)}\right]^{\top} \in \mathbb{R}^{N\times d}$. After a sufficiently large number of rounds $t_0$, the global model at the server converges and does not change much across the rounds, which results in that local updates also do not change much across the rounds. Then, we have

$$\boldsymbol{\Delta}^{(t_0;t_1)}_{\text{global}} = \mathbf{P}^{(t_0;t_1)} \boldsymbol{\Delta}^{(t_0)}_{\text{individual}} + \mathbf{Z}, \tag{31}$$

where $\boldsymbol{\Delta}^{(t_0;t_1)}_{\text{global}}$ denotes the concatenate of the global updates from round $t_0$ to round $t_1 - 1$, i.e., $\boldsymbol{\Delta}^{(t_0;t_1)}_{\text{global}} = \left[\delta^{(t_0)}, \ldots \delta^{(t_1-1)}\right]^{\top} \in \mathbb{R}^{(t_1-t_0)\times d}$ for $t_1 > t_0$, $\mathbf{P}^{(t_0;t_1)} \in \{0,1\}^{(t_1-t_0)\times N}$ is the participation




(a) $K(=8)$-uniform random selection ($T=1$). (b) Multi-RoundSecAgg ($T=2$).

**Figure 10:** Histogram of the reconstruction error defined in (33) when the $K(=8)$-uniform random selection or Multi-RoundSecAgg ($T=2$) scheme is used to train the LeNet on the CIFAR-10 dataset. The average reconstruction errors of $K(=8)$-uniform random selection and Multi-RoundSecAgg ($T=2$) are $6.715\times10^{-3}$ and 0.7829, respectively, which implies that the server can reconstruct all local updates when $K(=8)$-uniform random selection is used while the server cannot reconstruct the local updates when Multi-RoundSecAgg ($T=2$) is used.

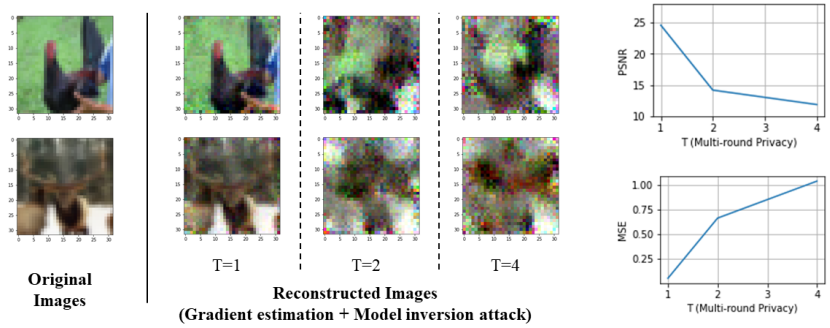



**Figure 11:** Comparison of the reconstructed images using the model inversion attack [12] with different value of multi-round privacy guarantee $T$ (left) and measurement of similarity between the reconstructed images and the original images, where PSNR $=\infty$ and MSE $=0$ for two identical images (right).

matrix from round $t_0$ to round $t_1-1$, and $\mathbf{Z}$ denotes the perturbation (or noise) incurred by the local updates across the rounds.

The server can then estimate $\mathbf{\Delta}_{\text{individual}}^{(t_0)}$ by utilizing the least-squares method as follows

$$\hat{\mathbf{\Delta}}_{\text{individual}}^{(t_0)} = \left(\mathbf{P}^{(t_0;t_1)\top}\mathbf{P}^{(t_0;t_1)}\right)^{-1}\mathbf{P}^{(t_0;t_1)\top}\mathbf{\Delta}_{\text{global}}^{(t_0;t_1)}, \tag{32}$$

and we measure the reconstruction error as follows

$$e_i^{(t_0)} = \frac{\|\delta_i^{(t_0)} - \hat{\delta}_i^{(t_0)}\|_2^2}{\|\delta_i^{(t_0)}\|_2^2}, \tag{33}$$

where $\hat{\delta}_i^{(t_0)}$ denotes the reconstructed gradient of user $i$, which corresponds to $i$-th row of $\hat{\mathbf{\Delta}}_{\text{individual}}^{(t_0)}$ in (32). On the other hand, in Multi-RoundSecAgg with multi-round privacy guarantee $T=2$, the server cannot estimate the individual gradients by utilizing (32) because $\mathbf{P}^{(t_0;t_1)}$ is not full rank hence the inverse of $\mathbf{P}^{(t_0;t_1)\top}\mathbf{P}^{(t_0;t_1)}$ does not exist. The best that the server can do is to estimate $\sum_{i\in\mathcal{G}_j}\delta_i^{(t_0)}$, where $\mathcal{G}_j$ is the index set of the users in the $j$-th batch. The server can then estimate $\delta_i^{(t_0)}$ by dividing the estimate of $\sum_{i\in\mathcal{G}_j}\delta_i^{(t_0)}$ by $T$, where $i\in\mathcal{G}_j$.

Figure 10(a) and Figure 10(b) show the histogram of the reconstruction error of the individual gradients when the $K$-uniform random selection scheme and Multi-RoundSecAgg ($T=2$) scheme are used, respectively. We set $t_0=1460$ and $t_1=1500$ in this experiment. We observe that the $K$-uniform

random selection scheme has much smaller average reconstruction error $\frac{1}{N}\sum_{i=1}^{N} e_i^{(t_0)} = 6.715 \times 10^{-3}$ than the average reconstruction error of Multi-RoundSecAgg ($T = 2$), which implies that the server can reconstruct all local gradients as the $K$-uniform random selection scheme has a multi-round privacy guarantee $T = 1$.

Finally, the server can reconstruct the training images by applying model inversion attack [12] to the reconstructed gradient $\hat{\delta}_i^{(t_0)}$. Figure 11 the reconstructed images of random selection scheme ($T = 1$) and Multi-RoundSecAgg ($T = 2, 4$). We measure the reconstruction performance using peak signal-to-noise ratio (PSNR) and mean square error (MSE). Large PSNR and small MSE indicate more similarity between the reconstructed and original images, and hence we can observe that random selection scheme ($T = 1$) leaks much more information about the original image than Multi-RoundSecAgg ($T = 2, 4$).

# I General Convex and Non-Convex Convergence Rates

In this appendix, we discuss extending the convergence proof of [18] to our setting. Since we closely follow [18], we mainly here focus on the differences. The main difference between the two settings is that the number of participating users in [18] is fixed as $|\mathcal{S}^{(t)}| = K$ across all rounds as *dropouts* are not considered, whereas it may change in our case between the rounds based on the availability of the users. Specifically, the server in our case aims to choose $K$ users at each round. If this is not possible due to the dropouts, the server skips this round. That is, $|\mathcal{S}^{(t)}| \in \{0, K\}$ and $\mathbb{E}[|\mathcal{S}^{(t)}|] = C$, where we recall that $C$ is the average aggregation cardinality that depends on the desired multi-round privacy guarantee $T$.

Next, we recall the setting and the assumptions of [18]. In [18], the problem is formalized as minimizing a global loss function as follows

$$\min_{\boldsymbol{x}} L(\boldsymbol{x}) \text{ s.t. } L(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} L_i(\boldsymbol{x}), \tag{34}$$

where $L$ is bounded from below by $L^*$, the loss function of user $i$ $L_i$ is $\rho$-smooth and $g_i(\boldsymbol{x}) = \nabla L_i(\boldsymbol{x}, \zeta_i)$ is an unbiased stochastic gradient of $L_i$ with variance bounded by $\sigma^2$. Furthermore, following the assumptions of [18], we consider the following assumptions.

**Assumption 1**. $(G, B)$-Bounded Gradient Dissimilarity (BGD). There exists constants $G \geq 0$ and $B \geq 1$ such that

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla L_i(\boldsymbol{x})\|^2 \leq G^2 + B^2 \|\nabla L_i(\boldsymbol{x})\|^2, \forall \boldsymbol{x}, \tag{35}$$

and when $\{L_i\}$ are convex, this assumption can be relaxed as

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla L_i(\boldsymbol{x})\|^2 \leq G^2 + 2\rho B^2 (L(\boldsymbol{x}) - L^*), \forall \boldsymbol{x}. \tag{36}$$

**Assumption 2**. $\delta$-Bounded Hessian Dissimilarity (BHD).

$$\|\nabla^2 L_i(\boldsymbol{x}) - \nabla^2 L(\boldsymbol{x})\| \leq \delta, \forall \boldsymbol{x}, \tag{37}$$

and $L_i$ is $\delta$-weakly.

**Assumption 3**. $L_i$is $\mu$-convex for $\mu \geq 0$ and satisfies

$$\langle \nabla L_i(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq -\left( L_i(\boldsymbol{x}) - L_i(\boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \right), \text{for any } i, \boldsymbol{x}, \boldsymbol{y}, \tag{38}$$

where $\mu$ can be 0 (general convex case).

**Assumption 4**. $g_i(\boldsymbol{x}) = \nabla L_i(\boldsymbol{x}; \zeta_i)$ is an unbiased stochastic gradient of $L_i$ with bounded variance. That is, we have

$$\mathbb{E}\left[\|g_i(\boldsymbol{x}) - \nabla L_i(\boldsymbol{x})\|\right] \leq \sigma^2, \text{for any } i, \boldsymbol{x}. \tag{39}$$

**Assumption 5**. $\{L_i\}$ are $\rho$-smooth and satisfy

$$\|\nabla L_i(\boldsymbol{x}) - \nabla L_i(\boldsymbol{x})\| \leq \rho \|\boldsymbol{x} - \boldsymbol{y}\|, \text{for any } i, \boldsymbol{x}, \boldsymbol{y}. \tag{40}$$

It is also worth noting that when $\{L_i\}$ are convex and $\boldsymbol{x}^*$ this assumption implies that

$$\frac{1}{2\rho N}\sum_{i=1}^{N}\|\nabla L_i(\boldsymbol{x}) - L_i(\boldsymbol{x}^*)\| \leq L(\boldsymbol{x}) - L^*, \tag{41}$$

and when $L_i$ is twice differentiable this assumption implies that $\|\nabla^2 L_i(\boldsymbol{x})\| \leq \rho$ for any $\boldsymbol{x}$.

We now recall the global and the local updates of the setting considered in [18]. $\boldsymbol{x}^{(t)}$ is the global model after round $t$ and $\boldsymbol{x}_{i,e}^{(t)}$ is the local model of user $i$ in round $t$ and local step $e$. In round $t$, the server selects a subset of users $\mathcal{S}^{(t)}$. Each user then copies the global model $\boldsymbol{x}_{i,0}^{(t)} = \boldsymbol{x}^{(t-1)}$ and performs $E$ local update steps as follows

$$\boldsymbol{x}_{i,e}^{(t)} \coloneqq \boldsymbol{x}_{i,e-1}^{(t)} - \eta_l g_i(\boldsymbol{x}_{i,e-1}^{(t)}), \tag{42}$$

where $\eta_l$ is the local step size. The users then send their updates and the server updates the global model as follows

$$\boldsymbol{x}^{(t)} \coloneqq \boldsymbol{x}^{(t-1)} + \frac{\eta_g}{K}\sum_{i\in\mathcal{S}}(\boldsymbol{x}_{i,E}^{(t)} - \boldsymbol{x}^{(t-1)}), \tag{43}$$

where $\eta_g$ is the global step size. Finally, the output is given by

$$\overline{\boldsymbol{x}}^{(J)} = \overline{\boldsymbol{x}}^{(t-1)} \text{ with probability } \frac{\theta_t}{\sum_\tau \theta_\tau} \text{for } t \in \{1, \cdots, J+1\} \tag{44}$$

for some weights $\{\theta_t\}$, where $J$ is the total number of rounds. Next, we restate the convergence Theorem of [18].

**Theorem.** *Suppose that $\{L_i\}$ satisfy Assumptions 1, 4 and 5. Then for each of the following cases there exists weights $\{\theta_t\}$ and local step sizes $\eta_l$ such that for any global step size $\eta_g \geq 1$, we have*

- ***Strongly convex.*** *If $\{L_i\}$ satisfy Assumption 3 for $\mu > 0$, $\eta_l \leq \frac{1}{8(1+B^2)\rho E\eta_g}$, $J \geq \frac{8(1+B^2)\rho}{\mu}$, then*

$$\mathbb{E}\left[L(\overline{\boldsymbol{x}}^J)\right] - L(\boldsymbol{x}^*) \leq \tilde{O}\left(\frac{M^2}{\mu JEK} + \frac{\rho G^2}{\mu^2 J^2} + \mu D^2 \exp\left(-\frac{\mu}{16(1+B^2)\rho}J\right)\right). \tag{45}$$

- ***General convex.*** *If $\{L_i\}$ satisfy Assumption 3 for $\mu = 0$, $\eta_l \leq \frac{1}{8(1+B^2)\rho E\eta_g}$, $J \geq 1$, then*

$$\mathbb{E}\left[L(\overline{\boldsymbol{x}}^J)\right] - L(\boldsymbol{x}^*) \leq O\left(\frac{MD}{\sqrt{JEK}} + \frac{D^{4/3}(\rho G^2)^{1/3}}{(J+1)^{2/3}} + \frac{B^2\rho D^2}{J}\right). \tag{46}$$

- ***Non-convex.*** *If $\{L_i\}$ satisfy Assumption 1 and $\eta_l \leq \frac{1}{8(1+B^2)\rho E\eta_g}$, then*

$$\mathbb{E}\left[\|\nabla L(\overline{\boldsymbol{x}}^J)\|^2\right] \leq O\left(\frac{\rho M\sqrt{F}}{\sqrt{JEK}} + \frac{F^{2/3}(\rho G^2)^{1/3}}{(J+1)^{2/3}} + \frac{B^2\rho F}{J}\right). \tag{47}$$

*where $M^2 = \sigma^2(1 + K/\eta_g^2) + E(1-K/N)G^2$, $D = \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|$ and $F = L(\boldsymbol{x}^0) - L^*$.*

As we discussed, the main difference between our setting and the setting of [18] is that the number of selected users in our case in each round is a random variable $|\mathcal{S}^{(t)}| \in \{0, K\}$ with mean equal to the average aggregation cardinality $C$. We now show the effect of this difference on the key lemma of [18]. We first recall this key lemma from [18] and then derive a simple corollary that extends this lemma to our setting.

**Lemma.** *(Separating Mean and Variance)[Lemma 4 in [18]]. Let $\{\boldsymbol{E}_1, \boldsymbol{E}_2, \cdots, \boldsymbol{E}_w\}$ be random vectors in $\mathbb{R}^d$, which may not be independent. We consider the following two cases.*

- *When $\mathbb{E}[\boldsymbol{E}_i] = \boldsymbol{\zeta}_i$ and $\mathbb{E}[\|\boldsymbol{E}_i - \boldsymbol{\zeta}_i\|^2] \leq \sigma^2$, then we have*

$$\mathbb{E}[\|\sum_{i=1}^{w}\boldsymbol{E}_i\|^2] \leq \|\sum_{i=1}^{w}\boldsymbol{\zeta}_i\|^2 + w^2\sigma^2. \tag{48}$$

- *When* $\mathbb{E}[\boldsymbol{E}_i|\boldsymbol{E}_{i-1},\cdots,\boldsymbol{E}_1]=\boldsymbol{\zeta}_i$ *and* $\mathbb{E}[\|\boldsymbol{E}_i-\boldsymbol{\zeta}_i\|^2]\leq\sigma^2$*, then we have*

$$\mathbb{E}[\|\sum_{i=1}^{w}\boldsymbol{E}_i\|^2]\leq 2\|\sum_{i=1}^{w}\boldsymbol{\zeta}_i\|^2+2w\sigma^2. \tag{49}$$

In this key lemma, $w$ is constant as the setting [18] assumes the number of participating users is fixed as $K$ in every round. That is, this lemma is applied with $w=K$. In our case, however, the number of participating users is a random variable. Hence, we consider this case in the following corollary.

**Corollary 3.1.** *Let* $\{\boldsymbol{E}_1,\boldsymbol{E}_2,\cdots,\boldsymbol{E}_W\}$ *be random vectors in* $\mathbb{R}^d$*, which may not be independent and* $W\in\{0,w\}$ *is a random variable that is independent of* $E_i$ *and* $\mathbb{E}[W]=\mu_W$*. We consider the following two cases.*

- *When* $\mathbb{E}[\boldsymbol{E}_i]=\boldsymbol{\zeta}_i$ *and* $\mathbb{E}[\|\boldsymbol{E}_i-\boldsymbol{\zeta}_i\|^2]\leq\sigma^2$*, then we have*

$$\mathbb{E}[\|\sum_{i=1}^{W}\boldsymbol{E}_i\|^2]\leq\frac{\mu_W}{w}\|\sum_{i=1}^{w}\boldsymbol{\zeta}_i\|^2+w\mu_W\sigma^2. \tag{50}$$

- *When* $\mathbb{E}[\boldsymbol{E}_i|\boldsymbol{E}_{i-1},\cdots,\boldsymbol{E}_1]=\boldsymbol{\zeta}_i$ *and* $\mathbb{E}[\|\boldsymbol{E}_i-\boldsymbol{\zeta}_i\|^2]\leq\sigma^2$*, then we have*

$$\mathbb{E}[\|\sum_{i=1}^{W}\boldsymbol{E}_i\|^2]\leq 2\frac{\mu_W}{w}\|\sum_{i=1}^{w}\boldsymbol{\zeta}_i\|^2+2\mu_W\sigma^2. \tag{51}$$

This is the main difference between our setting and the setting considered in [18] and the rest of the proof follows similarly. Similar to Theorem 2, we can see that the average aggregation cardinality $C$ controls the convergence rate and hence there is a trade-off between the multi-round privacy $T$ and the convergence rate.