# A ADDITIONAL EXPERIMENT DETAILS AND RESULTS

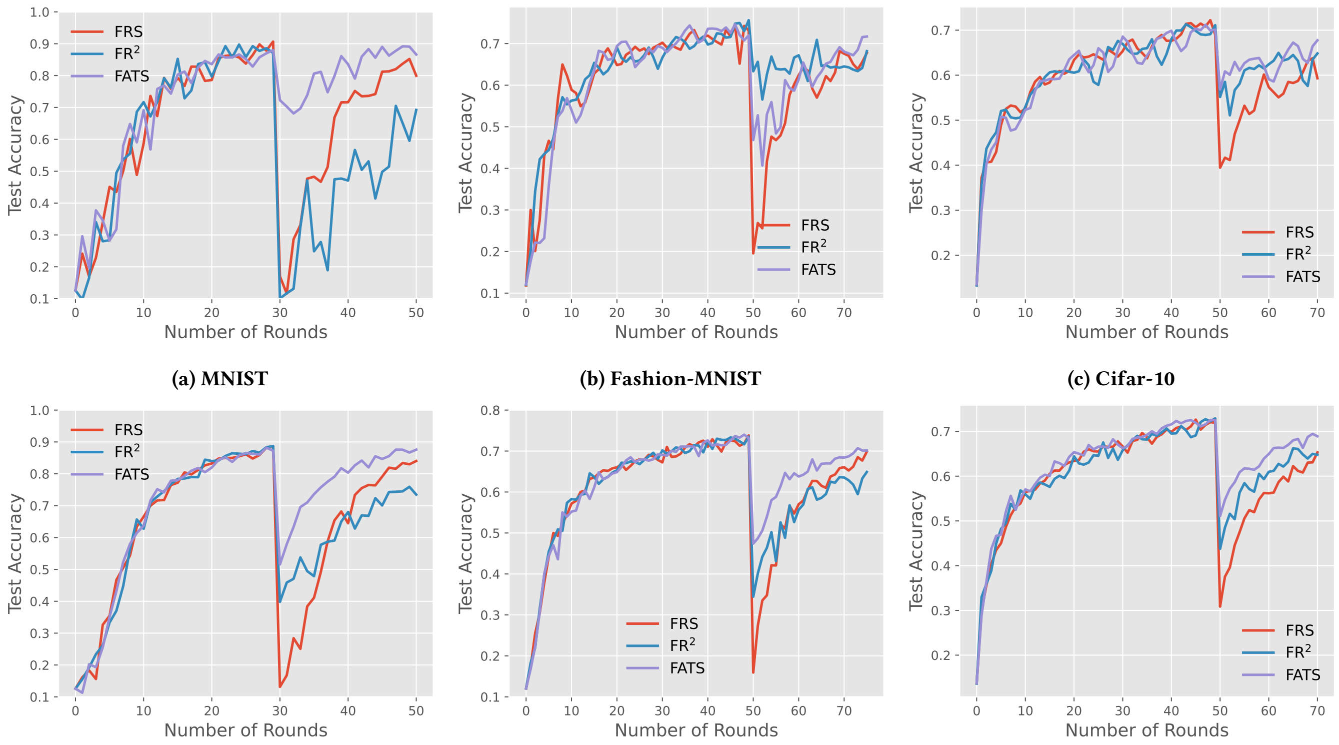
## A.1 Dataset Details for Experiments

**Table 2: Datasets**

| Dataset | # sample | M | K | R | E | b | Model |
|---|---|---|---|---|---|---|---|
| MNIST | 60,000 | 300 | 5 | 30 | 10 | 10 | CNN |
| FashionM | 60,000 | 300 | 5 | 50 | 10 | 10 | CNN |
| Cifar-10 | 60,000 | 600 | 5 | 50 | 10 | 10 | VGG16 |
| Cifar-100 | 60,000 | 600 | 10 | 50 | 10 | 10 | VGG16 |
| FEMNIST | 811,586 | 3,556 | 5 | 350 | 20 | 10 | CNN |
| Shakes | 3,678,451 | 660 | 20 | 30 | 100 | 60 | LSTM |

## A.2 Additional Results for Performance Evaluation

We present the learning and unlearning performance evaluation results over MNIST, Fashion-MNIST, and Cifar-10 in Figure 5.



**(a) MNIST**

**(b) Fashion-MNIST**

**(c) Cifar-10**

**(d) MNIST**

**(e) Fashion-MNIST**

**(f) Cifar-10**

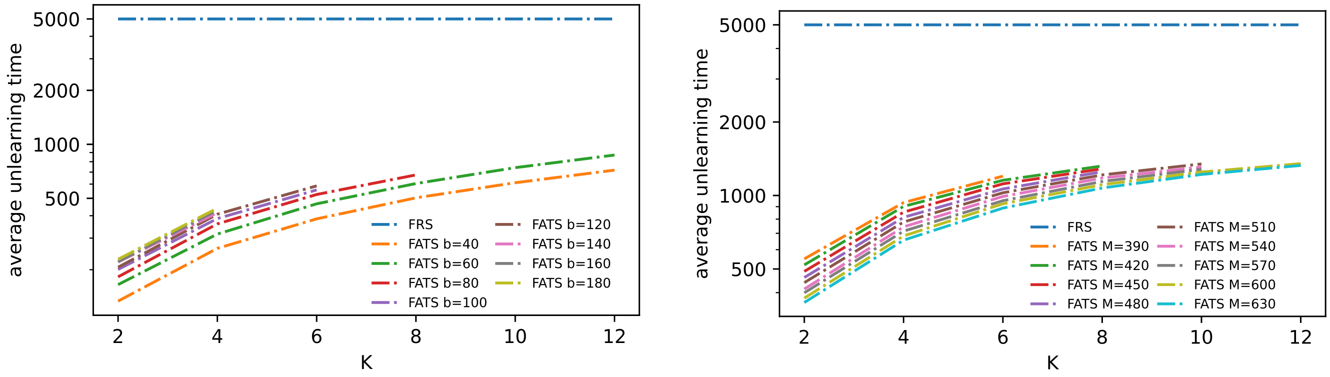
**Figure 5: Comparison of the test accuracy of different methods and their changes after conducting unlearning on MNIST, Fashion-MNIST, and Cifar-10. Top: sample-level unlearning. Bottom: client-level unlearning.**

## A.3 Additional Results for Unlearning Efficiency

We also tested the unlearning efficiency of FATS on the Shakespeare dataset. The results are shown in Figure 6 and convey the same observations made in Section 6.2.2.

## A.4 Additional Results for Utility v.s. Efficiency.

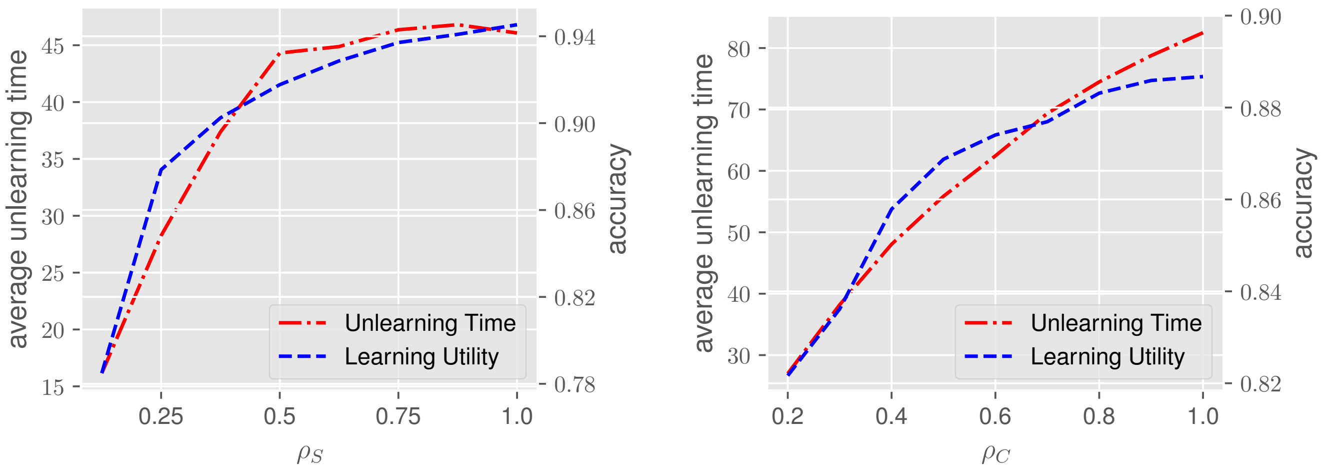
We showcase our additional results for the trade-off between learning utility and unlearning efficiency on MNIST in Figure 7.



**(a) Shakespeare: Sample Unlearning**

**(b) Shakespeare: Client Unlearning**

**Figure 6: Unlearning Efficiency of FATS compared with FRS on Shakespeare.**



**(a) MNIST: Sampling Unlearning**

**(b) MNIST: Client Unlearning**

**Figure 7: Impacts of stability parameters on learning utility and unlearning efficiency**

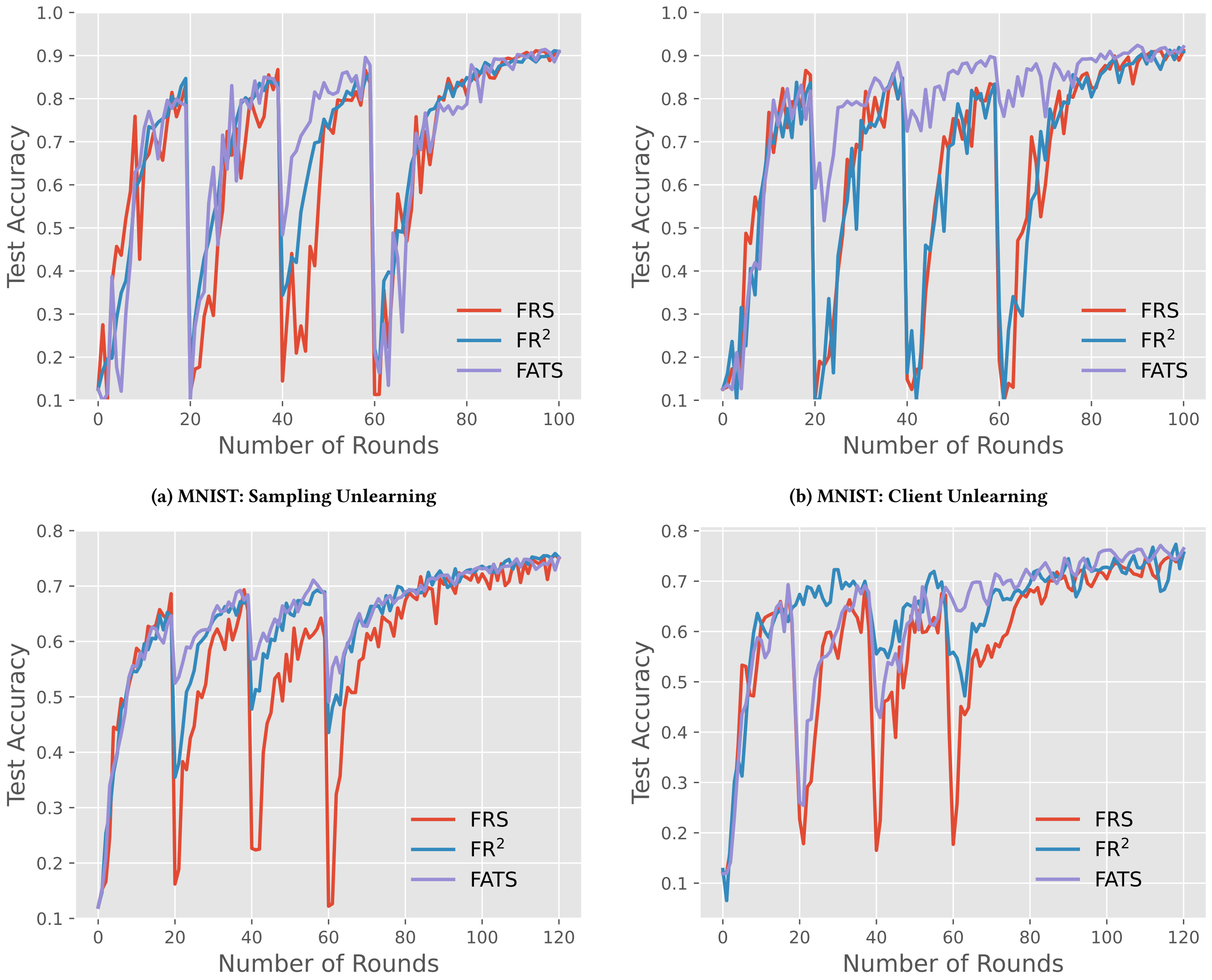
### A.5 Performance Evaluation over A Stream of Unlearning Requests

To provide a more comprehensive performance evaluation, we also investigate the streaming unlearning setting. We evaluate the performance of different methods on MNIST and FEMNIST datasets under streaming unlearning requests, where we sequentially issue multiple requests to remove data samples or clients from the training process. We use the same setting as described in Section 6.2.1. Figure 8 shows the results. The results are consistent with the ones presented in Section 6.2.1. As we can observe from Figure 8, our proposed method FATS achieves the highest test accuracy among all methods and maintains a stable performance under different unlearning scenarios. The baseline methods, $\mathsf{FR}^2$ and FRS, suffer from significant drops in test accuracy when unlearning requests are issued and thus take more time to recover from the unlearning. This indicates that FATS is more effective and efficient in handling streaming unlearning requests than the other methods.

## B OMITTED PROOFS FOR UNLEARNING ANALYSIS

### B.1 Proof of Lemma 1

Proof of Lemma 1. We start with the sample-level TV-stability. We denote by $\mu_{(M,K,N,b)}$ the per-round sampling probability measure for $\mathcal{D}$, including both client and mini-batch sampling in each round. Let $\mu^{\otimes R}_{(M,K,N,b)}$ denote the product measure of $R$ copies of $\mu_{(M,K,N,b)}$. Analogously, we define $\mu_{(M,K,N-1,b)}$ and $\mu^{\otimes R}_{(M,K,N-1,b)}$ for $\mathcal{D}'$. We extend the $\sigma$-algebra for probability space with measure $\mu_{(M,K,N-1,b)}$ to obtain measure $\mu'_{(M,K,N,b)}$ such that it has a common probability space with $\mu_{(M,K,N,b)}$. To this end, we define $\mu'_{(M,K,N,b)}$ as follows: for



**(a) MNIST: Sampling Unlearning**

**(b) MNIST: Client Unlearning**

**(c) FEMNIST: Sample Unlearning**

**(d) FEMNIST: Client Unlearning**

**Figure 8: Comparison of the test accuracy of different methods and their changes under streaming unlearning requests.**

any feasible pair of $(\mathcal{P}^{[r]}, \mathcal{B}^{[r]})$,

$$\mu'_{(M,K,N,b)}(\mathcal{P}^{[r]}, \mathcal{B}^{[r]}) = \begin{cases} 0, & \text{if } X_u \in \mathcal{B}_{k_u}^{[r,i]} \text{ for some } i \in [E], \\ \mu_{(M,K,N-1,b)}(\mathcal{P}^{[r]}, \mathcal{B}^{[r]}), & \text{otherwise.} \end{cases}$$

Similarly, we extend the $\sigma$-algebra for the product space with measure $\mu^{\otimes R}_{(M,K,N-1,b)}$ to get $\mu'^{\otimes R}_{(M,K,N,b)}$.

We use $L(\cdot)$ to denote the output model of learning algorithm $\mathcal{L}(\cdot)$. Note that for fixed initial model $\theta^{(0)}$ and other parameters, $L(\mathcal{D})$ and $L(\mathcal{D}')$ are the same deterministic map from a sampling history $\mathcal{H} := [(\mathcal{P}^{[1]}, \mathcal{B}^{[1]}), (\mathcal{P}^{[2]}, \mathcal{B}^{[2]}), \ldots, (\mathcal{P}^{[R]}, \mathcal{B}^{[R]})]$ to an output model parameter. The discrepancy between $L(\mathcal{D})$ and $L(\mathcal{D}')$ arises from the distinct measures on the input space that is induced by the sampling histories. Therefore, the total variation distance between $L(\mathcal{D})$ and $L(\mathcal{D}')$ is equal to the total variation distance between the push-forward measures $L(\mathcal{D})\#\mu^{\otimes R}_{(M,K,N,b)}$ and $L(\mathcal{D}')\#\mu'^{\otimes R}_{(M,K,N,b)}$, which can be bounded above by the total variation distance between $\mu^{\otimes R}_{(M,K,N,b)}$ and $\mu'^{\otimes R}_{(M,K,N,b)}$ due to the equivalence of $L(\mathcal{D})$ and $L(\mathcal{D}')$ and the data processing inequality. Let $\overline{\mathcal{H}}$ be a sampling history that contains at least one pair $(\overline{\mathcal{P}}^{[r]}, \overline{\mathcal{B}}^{[r]})$ such that $k_u \in \overline{\mathcal{P}}^{[r]}$ and $X^{(N)}_{k_u} \in \overline{\mathcal{B}}^{[r]}_{k_u}$, which means that the target sample was involved in some round $r$. The total variation distance can be bounded by

$$\begin{aligned}
&\mathrm{TV}(L(\mathcal{D}), L(\mathcal{D}')) \leq \mathrm{TV}(\mu^{\otimes R}_{(M,K,N,b)}, \mu'^{\otimes R}_{(M,K,N,b)}) \\
&= \sup_{\mathcal{H} \text{ over all choices}} |\mu^{\otimes R}_{(M,K,N,b)}(\mathcal{H}) - \mu'^{\otimes R}_{(M,K,N,b)}(\mathcal{H})| \\
&= \mu^{\otimes R}_{(M,K,N,b)}(\overline{\mathcal{H}}) \leq R \cdot \mu_{(M,K,N,b)}(\overline{\mathcal{P}}^{[r]}, \overline{\mathcal{B}}^{[r]}) \\
&\leq R \cdot (\frac{K}{M} \cdot E \cdot \frac{b}{N}) \leq \frac{TKb}{MN},
\end{aligned}$$

where the first inequality is due to the data-processing inequality and the last inequality is due to $R \leq \frac{T}{E}$. By the definition of TV distance, we have $\mathrm{TV}(L(\mathcal{D}), L(\mathcal{D}')) \leq 1$ as a trivial bound. Noting that $K = \frac{\rho_C \cdot E \cdot M}{T}$ and $b = \frac{\rho_S \cdot N}{\rho_C \cdot E}$, we obtain the result that the output of FATS is $\min\{\rho_S, 1\}$ sample-level TV-stable.

Next, we move on to the client-level TV-stability. Unlike the sample-level case where a re-computation is triggered only when the target sample was involved, in client-level unlearning, a re-computation would be triggered if the target client was involved. Thus, for client-level unlearning, we focus on the client sampling history, i.e., $\mathcal{H}_c = \{\mathcal{P}^{[1]}, \mathcal{P}^{[2]}, \dots, \mathcal{P}^{[R]}\}$. Let $C$ be the original set of all clients and $C' := C \setminus \{k_u\}$. We denote by $\nu_{(M,K)}$ the per-round sampling probability measure for $C$. Then $\nu^{\otimes R}_{(M,K)}$ denotes the product measure of $R$ copies of $\nu_{(M,K)}$. We similarly define $\nu_{(M-1,K)}$ and $\nu^{\otimes R}_{(M-1,K)}$ for $C'$. We extend the $\sigma$-algebra for the probability space with measure $\nu_{(M-1,K)}$ to get $\nu'_{(M,K)}$ so that it has a common probability space with $\nu_{(M,K)}$. We define $\nu'_{(M,K)}$ as follows: for any possible client multiset $\mathcal{P}^{[r]}$,

$$\nu'_{(M,K)}(\mathcal{P}^{[r]}) = \begin{cases} 0 & \text{, if } k_u \in \mathcal{P}^{[r]} \\ \nu_{(M,K)}(\mathcal{P}^{[r]}) & \text{, otherwise} \end{cases}.$$

We similarly extend the $\sigma$-algebra for the product space with measure $\nu^{\otimes R}_{(M-1,K)}$ to get $\nu'^{\otimes R}_{(M,K)}$. Let $\overline{\mathcal{H}_c}$ be a client sampling history that contains at least one $\overline{\mathcal{P}}^{[r]}$ for $r \in [R]$ such that $k_u \in \overline{\mathcal{P}}^{[r]}$. Following a similar argument as in the sample-level case and using the data processing inequality, we can bound the total variation distance between $L(\mathcal{D}(C))$ and $L(\mathcal{D}(C'))$ by

$$\begin{aligned}
&\mathrm{TV}(L(\mathcal{D}(C)), L(\mathcal{D}(C'))) \leq \mathrm{TV}(\nu^{\otimes R}_{(M,K)}, \nu'^{\otimes R}_{(M,K)}) \\
&= \sup_{\mathcal{H}_c \text{ over all choices}} |\nu^{\otimes R}_{(M,K)}(\mathcal{H}_c) - \nu'^{\otimes R}_{(M,K)}(\mathcal{H}_c)| \\
&= \nu^{\otimes R}_{(M,K)}(\overline{\mathcal{H}_c}) \leq R \cdot \nu_{(M,K)}(\overline{\mathcal{P}}^{[r]}) \leq R \cdot \frac{K}{M} \leq \frac{TK}{EM},
\end{aligned}$$

where the first inequality is due to the data-processing inequality and the last inequality is due to $R \leq \frac{T}{E}$. By definition of TV distance, we have $\mathrm{TV}(L(\mathcal{D}(C)), L(\mathcal{D}(C'))) \leq 1$ as a trivial bound. Noticing that $K = \frac{\rho_C \cdot E \cdot M}{T}$, we obtain the result that the output of FATS is $\min\{\rho_C, 1\}$ client-level TV-stable. □

## B.2 Proof of Theorem 1

Proof of Theorem 1. We adopt the same notation as in the proof of Lemma 1 and we first consider the one deletion case. For sample-level unlearning, we essentially substitute some mini-batches drawn by client $k_u$ in each round that $k_u$ participated in. To simplify the presentation, we use $\mathsf{SU}_i : (\mathcal{D}_{k_u})^b \to (\mathcal{D}_{k_u} \setminus \{X_u\})^b$ to denote one iteration of FATS-SU for handling a mini-batch drawn by $k_u$, and $\mathsf{SU}_r : (\mathcal{D}_{k_u})^{b \times E} \to (\mathcal{D}_{k_u} \setminus \{X_u\})^{b \times E}$ to denote all such operations in one communication round. For any non-empty $\mathcal{B}^{[r]}_{k_u} \in (\mathcal{D}_{k_u})^{b \times E}$ generated by $\mu_{(M,K,N,b)}$, we observe that the mapping $\mathsf{SU}_r$ acts *component-wise* and *symmetrically*, which means that

$$\mathsf{SU}_r(\mathcal{B}^{[r]}_{k_u}) = \left(\mathsf{SU}_i(\mathcal{B}^{[r,1]}_{k_u}), \mathsf{SU}_i(\mathcal{B}^{[r,2]}_{k_u}), \dots, \mathsf{SU}_i(\mathcal{B}^{[r,E]}_{k_u})\right).$$

Define $\mu^{\mathsf{SU}_r}_{(M,K,N,b)} := \mathsf{SU}_r \# \mu_{(M,K,N,b)}$. We first show in Claim 1 that $\mathsf{SU}_r$ transports $\mu_{(M,K,N,b)}$ to $\mu_{(M,K,N-1,b)}$, i.e., $\mu^{\mathsf{SU}_r}_{(M,K,N,b)} = \mu_{(M,K,N-1,b)}$. The proof of Claim 1 is in Appendix B.3.

**Claim 1.** For any possible pair $(\mathcal{P}^{[r]}, \mathcal{B}^{[r]})$, it holds that

$$\mu^{\mathsf{SU}_r}_{(M,K,N,b)}(\mathcal{P}^{[r]}, \mathcal{B}^{[r]}) = \mu_{(M,K,N-1,b)}(\mathcal{P}^{[r]}, \mathcal{B}^{[r]}).$$

We proceed to describe the coupling constructed by the unlearning operations. Suppose we first get a sampling history

$$\mathcal{H} = ((\mathcal{P}^{[1]}, \mathcal{B}^{[1]}), (\mathcal{P}^{[2]}, \mathcal{B}^{[2]}), \dots, (\mathcal{P}^{[R]}, \mathcal{B}^{[R]}))$$

according to the measure $\mu^{\otimes R}_{(M,K,N,b)}$. Furthermore, we consider another sampling history $\mathcal{H}'$ whose each element, denoted as $(\mathcal{P}'^{[r]}, \mathcal{B}'^{[r]})$, is defined as follows. If $X_u$ is not contained in any mini-batch in $\mathcal{B}^{[r]}_{k_u}$, then $(\mathcal{P}'^{[r]}, \mathcal{B}'^{[r]}) = (\mathcal{P}^{[r]}, \mathcal{B}^{[r]})$. Otherwise, $(\mathcal{P}'^{[r]}, \mathcal{B}'^{[r]})$ is sampled from $\mu_{(M,K,N-1,b)}$. This way, we obtain a pair of coupled sampling histories $\mathcal{H}$ and $\mathcal{H}'$. Following directly from Claim 1, we have the following claim for the coupled history.

**Claim 2.** For any possible sampling history $H$, it holds that $\mathbb{P}(\mathcal{H} = H) = \mu^{\otimes R}_{(M,K,N,b)}$ and $\mathbb{P}(\mathcal{H}' = H) = \mu^{\otimes R}_{(M,K,N-1,b)}$.

Claim 2 states that our unlearning operations in FATS-SU indeed transport the per-round sample and client sampling probability measure $\mu^{\otimes R}_{(M,K,N,b)}$ for the original dataset $\mathcal{D}$ to the measure $\mu^{\otimes R}_{(M,K,N-1,b)}$ for the updated dataset $\mathcal{D}'$, which guarantees sample-level exact unlearning of FATS-SU. Next, we show in Claim 3 that the probability of disagreement under the above coupling, i.e., the probability that a re-computation occurs, is bounded by sample-level TV-stability $\rho_S$. The proof of Claim 3 is in Appendix B.4.

**Claim 3.** For the $\rho_S$ sample-level TV-stable FATS, under the coupling of $(\mathcal{H}, \mathcal{H}')$, we have $\mathbb{P}_{(\mathcal{H},\mathcal{H}')}(\mathcal{H} \neq \mathcal{H}') \leq \rho_S$.

Finally, from Remark 1, for $w$ sample unlearning requests, the re-compute probability is at most $w \cdot \rho_S$.

The proof for the client-level unlearning case follows a similar argument as sample-level unlearning. The main difference is that we only focus on client sampling, just as we did for Lemma 1. □

## B.3 Proof of Claim 1

Proof of Claim 1. We concentrate on the mini-batch sampling measure at client $k_u$, since the unlearning operation only affects the mini-batches sampled by client $k_u$. Specifically, we denote by $\xi_{(N,b)}$ the probability measure of $k_u$'s sampling a mini-batch of size $b$ from $\mathcal{D}_{k_u}$ and by $\xi^{\texttt{dul}}_{(N,b)} := \texttt{dul}\#\xi_{(N,b)}$. Similarly, we denote by $\xi_{(N-1,b)}$ the probability measure for sampling a $b$ size mini-batch from $\mathcal{D}_{k_u} \setminus \{X_u\}$. Let $B \in (\mathcal{D}_{k_u})^b$ be any possible mini-batch. We show $\xi^{\texttt{dul}}_{(N,b)}(B) = \xi_{(N-1,b)}(B)$ for each local iteration by two cases.

*Case 1.* If the verification fails, i.e., a re-computation is triggered, the algorithm will re-sample $B \sim \xi_{(N-1,b)}$. Thus, $\xi^{\texttt{dul}}_{(N,b)}(B) = \xi_{(N-1,b)}(B)$ holds trivially for this case.

*Case 2.* If the verification succeeds. Then the deleted data $X_u$ is not contained in $B$. The measure $\xi^{\texttt{dul}}_{(N,b)}$ is therefore just the probability under the original sampling measure $\xi_{(N,b)}$ conditioned on the event that $X_u \notin B$. Therefore,

$$\xi^{\texttt{dul}}_{(N,b)}(B) = \xi_{(N,b)}(B | X_u \notin B) = \frac{\xi_{(N,b)}(B \cap \{X_u \notin B\})}{\xi_{(N,b)}(\{X_u \notin B\})},$$

where, by direct computation, $\xi_{(N,b)}(\{X_u \notin B\}) = 1 - \xi_{(N,b)}(\{X_u \in B\}) = 1 - \frac{\binom{N-1}{b-1}}{\binom{N}{b}} = 1 - \frac{b}{N}$. To calculate the numerator, we consider two possible choices of $B$. First, if $X_u \in B$, then $\xi_{(N,b)}(B \cap \{X_u \notin B\}) = 0$, which gives us that $\xi^{\texttt{dul}}_{(N,b)}(B) = 0 = \xi_{(N-1,b)}(B)$. For the choice that $X_u \notin B$,

$$\begin{aligned} \xi^{\texttt{dul}}_{(N,b)}(B) &= \frac{\xi_{(N-1,b)}(B)}{\xi_{(N-1,b)}(\{X_u \notin B\})} = \frac{1/\binom{N}{b}}{1 - \frac{b}{N}} \\ &= \frac{N}{N-b} \frac{(N-b)!b!}{N!} = \frac{(N-b-1)!b!}{(N-1)!} \\ &= \frac{1}{\binom{N-1}{b}} = \xi_{(N-1,b)}(B). \end{aligned}$$

Therefore, $\xi^{\texttt{dul}}_{(N,b)}(B) = \xi_{(N-1,b)}(B)$ also holds for Case 2. □

## B.4 Proof of Claim 3

Proof of Claim 3.

$$\begin{aligned} &\mathbb{P}_{(\mathcal{H},\mathcal{H}')}(\mathcal{H} \neq \mathcal{H}') \\ &= \mathbb{P}_{(\mathcal{H},\mathcal{H}')}(\exists r \in [R] \text{ s.t. } (\mathcal{P}[r], \mathcal{B}[r]) \neq (\mathcal{P}'^{[r]}, \mathcal{B}'^{[r]})) \\ &= \mathbb{P}_{\mathcal{H}}(\exists r \in [R], i \in [E] \text{ s.t. } X_u \in \mathcal{B}^{[r,i]}_{k_u}) \leq \frac{TKb}{MN} = \rho_S. \end{aligned}$$

□

# C OMITTED PROOFS FOR CONVERGENCE ANALYSIS

## C.1 Proof of Lemma 2

Proof of Lemma 2. Our algorithm involves two sources of stochasticity. The first one arises from the stochastic gradient, while the second one stems from the client sampling. To differentiate them, we adopt the notation $\mathbb{E}_{\mathcal{B}^{(t)}} := \mathbb{E}_{\mathcal{B}_1^{(t)},\dots,\mathcal{B}_M^{(t)}}$ to indicate the expectation with respect to the randomness of mini-batch sampling in iteration $t$, and use $\mathbb{E}_{\mathcal{P}^{(t)}}$ to eliminate the randomness of client sampling in iteration $t$. Let $\mathbb{E}_t$ be the expectation with respect to all of the randomness until time $t$, then $\mathbb{E}_t = \mathbb{E}_{\mathcal{B}^{(t)}}\mathbb{E}_{\mathcal{P}^{(t)}}\mathbb{E}_{t-1}$. We drop the time indicator $t$ and simply use $\mathbb{E}$ to denote the expectation with respect to all the relevant randomness when it is clear from the context.

Using the update rule $\theta_k^{(t)} \leftarrow \theta_k^{(t-1)} - \eta \cdot \tilde{g}_k^{(t)}$ and the $L$-smoothness of the loss function $f$ (Assumption 1), we derive the following inequality:

$$F(\theta^{(t)}) - F(\theta^{(t-1)}) \leq -\eta\langle\nabla F(\theta^{(t-1)}), \tilde{g}^{(t)}\rangle + \frac{\eta^2 L}{2}\|\tilde{g}^{(t)}\|_2^2.$$

Taking expectation on both sides, we obtain

$$\mathbb{E}[F(\theta^{(t)}) - F(\theta^{(t-1)})] \leq -\eta\mathbb{E}[\langle\nabla F(\theta^{(t-1)}), \tilde{g}^{(t)}\rangle] + \frac{\eta^2 L}{2}\mathbb{E}[\|\tilde{g}^{(t)}\|_2^2].$$

Averaging over all iterations $t = 1, 2, \dots, T$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\theta^{(t)}) - F(\theta^{(t-1)})] \leq -\frac{\eta}{T}\sum_{t=1}^{T}\mathbb{E}[\langle\nabla F(\theta^{(t-1)}), \tilde{g}^{(t)}\rangle] + \frac{\eta^2 L}{2T}\sum_{t=1}^{T}\mathbb{E}[\|\tilde{g}^{(t)}\|_2^2]. \tag{11}$$

To further bound the terms on the right hand side of the above inequality, we need the following results.

**Claim 4.** The expected inner product between the stochastic and the full-batch gradients in iteration $t$ can be bounded as follows:

$$-\eta\mathbb{E}[\langle\nabla F(\theta^{(t-1)}), \tilde{g}^{(t)}\rangle] \leq -\frac{\eta}{2}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] - \frac{\eta}{2}\mathbb{E}[\|\frac{1}{M}\sum_{k=1}^{M} g_k^{(t)}\|_2^2] + \frac{\eta L^2}{2M}\sum_{k=1}^{M}\mathbb{E}[\|\theta^{(t-1)} - \theta_k^{(t-1)}\|_2^2].$$

**Claim 5.** The averaged distance of local models from their virtual average during the learning process holds for

$$\frac{1}{TM}\sum_{t=1}^{T}\sum_{k=1}^{M}\mathbb{E}[\|\theta^{(t-1)} - \theta_k^{(t-1)}\|_2^2] \leq \frac{E(K+1)\eta^2 G^2}{Kb} + \frac{2\eta^2\lambda E(E-1)}{T}\sum_{t=1}^{T}\mathbb{E}[\|\frac{1}{M}\sum_{k=1}^{M} g_k^{(t)}\|_2^2]$$

**Claim 6.** The expected squared norm of the averaged stochastic gradient of iteration $t$ can be bounded as follows:

$$\mathbb{E}[\|\tilde{g}^{(t)}\|_2^2] \leq \frac{G^2}{Kb} + \lambda\mathbb{E}[\|\frac{1}{M}\sum_{k=1}^{M} g_k^{(t)}\|_2^2]$$

We continue the proof by utilizing Claim 4, claim 5 and claim 6 to further bound (11) as follows:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\theta^{(t)}) - F(\theta^{(t-1)})] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}(-\eta\mathbb{E}[\langle\nabla F(\theta^{(t-1)}), \tilde{g}^{(t)}\rangle]) + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta^2 L}{2}\mathbb{E}[\|\tilde{g}^{(t)}\|_2^2] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\Big(-\frac{\eta}{2}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] - \frac{\eta}{2}\mathbb{E}[\|\frac{1}{M}\sum_{k=1}^{M} g_k^{(t)}\|_2^2] + \frac{\eta L^2}{2M}\sum_{k=1}^{M}\mathbb{E}[\|\theta^{(t-1)} - \theta_k^{(t-1)}\|_2^2]\Big) + \frac{1}{T}\sum_{t=1}^{T}(\frac{\eta^2 L}{2}(\frac{G^2}{Kb} + \lambda\mathbb{E}[\|\frac{1}{M}\sum_{k=1}^{M} g_k^{(t)}\|_2^2])) \\
&= -\frac{\eta}{2T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] + (-\frac{\eta}{2} + \eta^3 L^2\lambda E(E-1) + \frac{\eta^2\lambda L}{2})\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\frac{1}{M}\sum_{k=1}^{M} g_k^{(t)}\|_2^2] + \frac{\eta^3 L^2 G^2 E(K+1)}{2Kb} + \frac{\eta^2 L G^2}{2Kb} \\
&\leq -\frac{\eta}{2T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] + \frac{\eta^3 L^2 G^2 E(K+1)}{2Kb} + \frac{\eta^2 L G^2}{2Kb},
\end{aligned}$$

where the last inequality is due to the condition that $-\frac{\eta}{2} + \eta^3 L^2\lambda E(E-1) + \frac{\eta^2\lambda L}{2} < 0$.

By rearranging, we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] \leq \frac{2(F(\theta^{(0)}) - F^*)}{\eta T} + \frac{\eta^2 L^2 G^2 E(K+1)}{Kb} + \frac{\eta L G^2}{Kb}.$$

Finally, by noting that $K = \frac{\rho_C E M}{T}$ and $b = \frac{\rho_S N}{\rho_C E}$, we can conclude the proof. □

## C.2 Proof of Claim 4

Proof of Claim 4.

$$
\begin{aligned}
& -\eta\mathbb{E}_t[\langle\nabla F(\theta^{(t-1)}),\tilde{g}^{(t)}\rangle] \\
&= -\eta\mathbb{E}_{t-1}\mathbb{E}_{\mathcal{B}^{(t)}}[\mathbb{E}_{\mathcal{P}^{(t)}}[\langle\nabla F(\theta^{(t-1)}),\tilde{g}^{(t)}\rangle]] \\
&= -\eta\mathbb{E}_{t-1}\mathbb{E}_{\mathcal{B}^{(t)}}[\mathbb{E}_{\mathcal{P}^{(t)}}[\langle\nabla F(\theta^{(t-1)}),\frac{1}{K}\sum_{k\in\mathcal{P}^{(t)}}\tilde{g}_k^{(t)}\rangle]] \\
&\overset{①}{=} -\eta\mathbb{E}_{t-1}\mathbb{E}_{\mathcal{P}^{(t)}}[\mathbb{E}_{\mathcal{B}^{(t)}}[\langle\nabla F(\theta^{(t-1)}),\frac{1}{K}\sum_{k\in\mathcal{P}^{(t)}}\tilde{g}_k^{(t)}\rangle]] \\
&= -\eta\mathbb{E}_{t-1}[\langle\nabla F(\theta^{(t-1)}),\mathbb{E}_{\mathcal{P}^{(t)}}[\frac{1}{K}\sum_{k\in\mathcal{P}^{(t)}}\mathbb{E}_{\mathcal{B}^{(t)}}[\tilde{g}_k^{(t)}]]\rangle] \\
&= -\eta\mathbb{E}_{t-1}[\langle\nabla F(\theta^{(t-1)}),\frac{1}{K}\mathbb{E}_{\mathcal{P}^{(t)}}[\sum_{k\in\mathcal{P}^{(t)}}g_k^{(t)}]\rangle] \\
&= -\eta\mathbb{E}_{t-1}[\langle\nabla F(\theta^{(t-1)}),\frac{1}{K}[K\sum_{k=1}^{M}\frac{1}{M}g_k^{(t)}]\rangle] \\
&\overset{②}{=} -\frac{\eta}{2}\mathbb{E}_{t-1}[\|\nabla F(\theta^{(t-1)})\|_2^2] - \frac{\eta}{2}\mathbb{E}_{t-1}[\|\frac{1}{M}\sum_{k=1}^{M}g_k^{(t)}\|_2^2] + \frac{\eta}{2}\mathbb{E}_{t-1}[\|\nabla F(\theta^{(t-1)}) - \frac{1}{M}\sum_{k=1}^{M}g_k^{(t)}\|_2^2] \\
&= -\frac{\eta}{2}\mathbb{E}_{t-1}\|\nabla F(\theta^{(t-1)})\|_2^2] - \frac{\eta}{2}\mathbb{E}_{t-1}[\|\frac{1}{M}\sum_{k=1}^{M}g_k^{(t)}\|_2^2] + \frac{\eta}{2}\mathbb{E}_{t-1}[\|\frac{1}{M}\sum_{k=1}^{M}(\nabla F_k(\theta^{(t-1)}) - g_k^{(t)})\|_2^2] \\
&\overset{③}{\leq} -\frac{\eta}{2}\mathbb{E}_{t-1}[\|\nabla F(\theta^{(t-1)})\|_2^2] - \frac{\eta}{2}\mathbb{E}_{t-1}[\|\frac{1}{M}\sum_{k=1}^{M}g_k^{(t)}\|_2^2] + \frac{\eta}{2}\frac{1}{M}\sum_{k=1}^{M}\mathbb{E}_{t-1}[\|\nabla F_k(\theta^{(t-1)}) - \nabla F_k(\theta_k^{(t-1)})\|_2^2] \\
&\overset{④}{\leq} -\frac{\eta}{2}\mathbb{E}_{t-1}[\|\nabla F(\theta^{(t-1)})\|_2^2] - \frac{\eta}{2}\mathbb{E}_{t-1}[\|\frac{1}{M}\sum_{k=1}^{M}g_k^{(t)}\|_2^2] + \frac{\eta}{2}\frac{L^2}{M}\sum_{k=1}^{M}\mathbb{E}_{t-1}[\|\theta^{(t-1)} - \theta_k^{(t-1)}\|_2^2],
\end{aligned}
$$

where ① is due to the fact that the randomness of $\mathcal{B}^{(t)}$ and $\mathcal{P}^{(t)}$ are independent, ② is due to the equation $2\langle a,b\rangle = \|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2$, ③ holds because of convexity of $\|\cdot\|_2$, and ④ follows from $L$-smoothness of loss function $f$ (Assumption 1). □

## C.3 Proof of Claim 5

Proof of Claim 5. Define $t_c \coloneqq \lfloor\frac{t-1}{E}\rfloor E$, which denotes the last iteration before entering the current round of iteration $t$. The local model at client $k$ in any iteration $t > t_c$ can be expressed as

$$
\theta_k^{(t)} = \theta_k^{(t-1)} - \eta\tilde{g}_k^{(t)} = \theta_k^{(t-2)} - [\eta\tilde{g}_k^{(t-2)} + \eta\tilde{g}_k^{t-1}] = \ldots = \theta_k^{t_c} - \sum_{\tau=t_c+1}^{t}\eta\tilde{g}_k^{(\tau)}. \qquad (12)
$$

From (12), we compute the virtual average model in $t$-th iteration as follows:

$$
\theta^{(t)} = \theta^{(t_c)} - \frac{\eta}{K}\sum_{k\in\mathcal{P}^{(t)}}\sum_{\tau=t_c+1}^{t}\tilde{g}_k^{(\tau)}.
$$

We express $t$ as $t = sE + r$, where $s \in \{0,1,\ldots,\lfloor\frac{T-1}{E}\rfloor\}$ denotes the indices of communication round and $r \in [E]$ denotes the indices of local updates. Observe that for $t_c < t \leq t_c + E$, $\mathbb{E}[\|\theta^{(t-1)} - \theta_k^{(t-1)}\|_2^2$ does not depend on time steps $t \leq t_c$ for all client $k \in [M]$. Thus, we obtain the following for all iterations $1 \leq t \leq T$:

$$
\frac{1}{TM}\sum_{t=1}^{T}\sum_{k=1}^{M}\mathbb{E}[\|\theta^{(t-1)} - \theta_k^{(t-1)}\|_2^2] = \frac{1}{TM}\sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{r=1}^{E}\sum_{k=1}^{M}\mathbb{E}[\|\theta^{(sE+r-1)} - \theta_k^{(sE+r-1)}\|_2^2].
$$

We first bound the term $\mathbb{E}[\|\theta^{(sE+r-1)} - \theta_k^{(sE+r-1)}\|_2^2]$ for any fixed tuple of $(s,r,k)$.

$$
\mathbb{E}[\|\theta^{(sE+r-1)} - \theta_k^{(sE+r-1)}\|_2^2]
$$

$$
\begin{aligned}
&= \mathbb{E}[\|\theta^{(sE)} - \frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta\tilde{g}_l^{(sE+i)} - \theta^{(sE)} + \sum_{i=1}^{r-1}\eta\tilde{g}_k^{(sE+i)}\|_2^2] \\
&= \mathbb{E}[\|\sum_{i=1}^{r-1}\eta\tilde{g}_k^{(sE+i)} - \frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta\tilde{g}_l^{(sE+i)}\|_2^2] \\
&\overset{\text{①}}{\leq} 2\mathbb{E}[\|\sum_{i=1}^{r-1}\eta\tilde{g}_k^{(sE+i)}\|_2^2 + \|\frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta\tilde{g}_l^{(sE+i)}\|_2^2] \\
&\overset{\text{②}}{=} 2\mathbb{E}[\|\sum_{i=1}^{r-1}\eta\tilde{g}_k^{(sE+i)} - \mathbb{E}[\sum_{i=1}^{r-1}\eta\tilde{g}_k^{(sE+i)}]\|_2^2 + \|\mathbb{E}[\sum_{i=1}^{r-1}\eta\tilde{g}_k^{(sE+i)}]\|_2^2 + \|\frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta\tilde{g}_l^{(sE+i)} - \mathbb{E}[\frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta\tilde{g}_l^{(sE+i)}]\|_2^2 + \|\mathbb{E}[\frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta\tilde{g}_l^{(sE+i)}]\|_2^2] \\
&= 2\mathbb{E}\left[\|\sum_{i=1}^{r-1}\eta(\tilde{g}_k^{(sE+i)} - g_k^{(sE+i)})\|_2^2 + \|\sum_{i=1}^{r-1}\eta g_k^{(sE+i)}\|_2^2 + \|\frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta(\tilde{g}_l^{(sE+i)} - g_l^{(sE+i)})\|_2^2 + \|\frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta g_l^{(sE+i)}\|_2^2\right] \\
&= 2\mathbb{E}\Bigg[\sum_{i=1}^{r-1}\eta^2\|\tilde{g}_k^{(sE+i)} - g_k^{(sE+i)}\|_2^2 + \sum_{u\neq v}\langle\eta\tilde{g}_k^{(u)} - \eta g_k^{(u)}, \eta\tilde{g}_k^{(v)} - \eta g_k^{(v)}\rangle + \|\sum_{i=1}^{r-1}\eta g_k^{(sE+i)}\|_2^2 \\
&\quad + \frac{1}{K^2}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta^2\|\tilde{g}_l^{(sE+i)} - g_l^{(sE+i)}\|_2^2 + \frac{1}{K^2}\sum_{u\neq v, y\neq z}\langle\eta\tilde{g}_y^{(u)} - \eta g_y^{(u)}, \eta\tilde{g}_z^{(v)} - \eta g_z^{(v)}\rangle + \|\frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta g_l^{(sE+i)}\|_2^2\Bigg] \\
&\overset{\text{③}}{=} 2\mathbb{E}\left[\sum_{i=1}^{r-1}\eta^2\|\tilde{g}_k^{(sE+i)} - g_k^{(sE+i)}\|_2^2 + \|\sum_{i=1}^{r-1}\eta g_k^{(sE+i)}\|_2^2 + \frac{1}{K^2}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta^2\|\tilde{g}_l^{(sE+i)} - g_l^{(sE+i)}\|_2^2 + \|\frac{1}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta g_l^{(sE+i)}\|_2^2\right] \\
&\overset{\text{④}}{\leq} 2\mathbb{E}[\sum_{i=1}^{r-1}\eta^2\|\tilde{g}_k^{(sE+i)} - g_k^{(sE+i)}\|_2^2 + (r-1)\eta^2\sum_{i=1}^{r-1}\|g_k^{(sE+i)}\|_2^2 + \frac{1}{K^2}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\eta^2\|\tilde{g}_l^{(sE+i)} - g_l^{(sE+i)}\|_2^2 + \frac{(r-1)\eta^2}{K}\sum_{l\in\mathcal{P}^{(t)}}\sum_{i=1}^{r-1}\|g_l^{(sE+i)}\|_2^2] \\
&= 2\sum_{i=1}^{r-1}\eta^2\mathbb{E}[\|\tilde{g}_k^{(sE+i)} - g_k^{(sE+i)}\|_2^2] + 2(r-1)\eta^2\sum_{i=1}^{r-1}\mathbb{E}[\|g_k^{(sE+i)}\|_2^2] + \frac{2\eta^2}{KM}\sum_{i=1}^{r-1}\sum_{l=1}^{M}\mathbb{E}[\|\tilde{g}_l^{(sE+i)} - g_l^{(sE+i)}\|_2^2] + \frac{2(r-1)\eta^2}{M}\sum_{i=1}^{r-1}\sum_{l=1}^{M}\mathbb{E}[\|g_l^{(sE+i)}\|_2^2] \\
&\overset{\text{⑤}}{\leq} 2(r-1)\eta^2\frac{G^2}{b} + 2(r-1)\eta^2\sum_{i=1}^{r-1}\mathbb{E}[\|g_k^{(sE+i)}\|_2^2] + \frac{2(r-1)\eta^2}{K}\frac{G^2}{b} + \frac{2(r-1)\eta^2}{M}\sum_{i=1}^{r-1}\sum_{l=1}^{M}\mathbb{E}[\|g_l^{(sE+i)}\|_2^2] \\
&= \frac{2(K+1)(r-1)\eta^2G^2}{Kb} + 2(r-1)\eta^2\sum_{i=1}^{r-1}\mathbb{E}[\|g_k^{(sE+i)}\|_2^2] + \frac{2(r-1)\eta^2}{M}\sum_{i=1}^{r-1}\sum_{l=1}^{M}\mathbb{E}[\|g_l^{(sE+i)}\|_2^2],
\end{aligned}
$$

where ① is due to the triangle inequality $\|a+b\|_2 \leq \|a\|_2 + \|b\|_2$ and the Cauchy—Schwarz inequality, ② follows from $\mathbb{E}[w^2] = \mathbb{E}[(w - \mathbb{E}[w])^2] + (\mathbb{E}[w])^2$, ③ is due to the independent mini-batch sampling and unbiased gradient estimation, ④ uses Cauchy—Schwarz inequality again, and ⑤ is due to bounded local variance assumption of local stochastic gradients (Assmption 2).

Next, we let $(s, r, k)$ vary and bound $\sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{r=1}^{E}\sum_{k=1}^{M}\mathbb{E}[\|\theta^{(sE+r-1)} - \theta_k^{(sE+r-1)}\|_2^2]$.

$$
\begin{aligned}
&\sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{r=1}^{E}\sum_{k=1}^{M}\mathbb{E}[\|\theta^{(sE+r-1)} - \theta_k^{(sE+r-1)}\|_2^2] \\
&= \sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{r=1}^{E}\sum_{k=1}^{M}\left(\frac{2(K+1)(r-1)\eta^2G^2}{Kb} + 2(r-1)\eta^2\sum_{i=1}^{r-1}\mathbb{E}[\|g_k^{(sE+i)}\|_2^2] + \frac{2(r-1)\eta^2}{M}\sum_{i=1}^{r-1}\sum_{l=1}^{M}\mathbb{E}[\|g_l^{(sE+i)}\|_2^2]\right) \\
&\leq \frac{TME(K+1)\eta^2G^2}{Kb} + 2\eta^2\sum_{r=1}^{E}(r-1)\sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{k=1}^{M}\sum_{i=1}^{r-1}\mathbb{E}[\|g_k^{(sE+i)}\|_2^2] + \frac{2\eta^2}{M}\sum_{r=1}^{E}(r-1)\sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{k=1}^{M}\sum_{i=1}^{r-1}\sum_{l=1}^{M}\mathbb{E}[\|g_l^{(sE+i)}\|_2^2] \\
&\leq \frac{TME(K+1)\eta^2G^2}{Kb} + \eta^2E(E-1)\sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{i=1}^{E}\sum_{k=1}^{M}\mathbb{E}[\|g_k^{(sE+i)}\|_2^2] + \eta^2E(E-1)\sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{i=1}^{E}\sum_{l=1}^{M}\mathbb{E}[\|g_l^{(sE+i)}\|_2^2]
\end{aligned}
$$

$$= \frac{TME(K+1)\eta^2G^2}{Kb} + 2\eta^2E(E-1) \sum_{s=0}^{\lfloor \frac{T-1}{E} \rfloor} \sum_{i=1}^{E} \sum_{k=1}^{M} \mathbb{E}[\|g_k^{(sE+i)}\|_2^2]$$
$$= \frac{TME(K+1)\eta^2G^2}{Kb} + 2\eta^2E(E-1) \sum_{t=1}^{T} \sum_{k=1}^{M} \mathbb{E}[\|g_k^{(t)}\|_2^2]$$

Finally, we get

$$\begin{aligned}\frac{1}{TM} \sum_{t=1}^{T} \sum_{k=1}^{M} \mathbb{E}[\|\theta^{(t-1)} - \theta_k^{(t-1)}\|_2^2] &\leq \frac{E(K+1)\eta^2G^2}{Kb} + \frac{2\eta^2E(E-1)}{TM} \sum_{t=1}^{T} \sum_{k=1}^{M} \mathbb{E}[\|g_k^{(t)}\|_2^2] \\ &\leq \frac{E(K+1)\eta^2G^2}{Kb} + \frac{2\eta^2\lambda E(E-1)}{T} \sum_{t=1}^{T} \mathbb{E}[\|\frac{1}{M} \sum_{k=1}^{M} g_k^{(t)}\|_2^2],\end{aligned}$$

where the last inequality is due to our bounded heterogeneity assumption (Assumption 3). □

## C.4 Proof of Claim 6

Proof of Claim 6.

$$\begin{aligned}&\mathbb{E}[\|\tilde{g}^{(t)}\|_2^2] \\ &= \mathbb{E}_{t-1}\mathbb{E}_{\mathcal{P}^{(t)}}\mathbb{E}_{\mathcal{B}^{(t)}}[\|\tilde{g}^{(t)}\|_2^2] \\ &= \mathbb{E}_{t-1}\mathbb{E}_{\mathcal{P}^{(t)}}\mathbb{E}_{\mathcal{B}^{(t)}}[\|\tilde{g}^{(t)} - \mathbb{E}_{\mathcal{B}^{(t)}}[\tilde{g}^{(t)}]\|_2^2 + \|\mathbb{E}_{\mathcal{B}^{(t)}}[\tilde{g}^{(t)}]\|_2^2] \\ &= \mathbb{E}_{t-1}\mathbb{E}_{\mathcal{P}^{(t)}}\mathbb{E}_{\mathcal{B}^{(t)}}[\|(\frac{1}{K}\sum_{k\in\mathcal{P}^{(t)}} \tilde{g}_k^{(t)}) - (\frac{1}{K}\sum_{k\in\mathcal{P}^{(t)}} g_k^{(t)})\|_2^2] + \mathbb{E}_{t-1}\mathbb{E}_{\mathcal{P}^{(t)}}[\|\frac{1}{K}\sum_{k\in\mathcal{P}^{(t)}} g_k^{(t)}\|_2^2] \\ &\leq \frac{1}{K^2}\mathbb{E}_{t-1}\mathbb{E}_{\mathcal{P}^{(t)}}\mathbb{E}_{\mathcal{B}^{(t)}}[\sum_{k\in\mathcal{P}^{(t)}} \|\tilde{g}_k^{(t)} - g_k^{(t)}\|_2^2] + \frac{1}{K}\mathbb{E}_{t-1}\mathbb{E}_{\mathcal{P}^{(t)}}[\sum_{k\in\mathcal{P}^{(t)}} \|g_k^{(t)}\|_2^2] \\ &\leq \frac{1}{K^2}\mathbb{E}_{t-1}[K\sum_{k=1}^{M}\frac{1}{M}\frac{G^2}{b}] + \frac{1}{K}\mathbb{E}_{t-1}[K\sum_{k=1}^{M}\frac{1}{M}\|g_k^{(t)}\|_2^2] \\ &\leq \frac{G^2}{Kb} + \lambda\mathbb{E}_{t-1}[\|\frac{1}{M}\sum_{k=1}^{M} g_k^{(t)}\|_2^2],\end{aligned}$$

where in the last inequality we use the bounded heterogeneity assumption (Assumption 3). □

## C.5 Proof of Theorem 2

Proof of Theorem 2. The choice of $\eta = \frac{1}{L\sqrt{\Gamma T}}$ comes by balancing the first and the last term in (9). With this choice of $\eta$, condition (7) becomes $\frac{E(E-1)}{T^2} < \frac{\Gamma}{2\lambda} - \frac{\sqrt{\Gamma}}{2T}$. By letting $T > \frac{2\lambda}{\sqrt{\Gamma}}$, it suffices to have $\frac{E}{T} < \frac{1}{2}\sqrt{\frac{\Gamma}{\lambda}}$ hold. Next we bound the mean-squared loss gradient norm. By the choice of $\eta$, we can calculate the sum of the first and the last term in (9) as

$$\frac{2(F(\theta^{(0)}) - F^*)}{\eta T} + \frac{\eta L G^2 T}{\rho_S MN} = \frac{3G\sqrt{L(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}}.$$

By taking the choice of $\eta$, we can calculate the the second term in (9) as

$$\frac{\eta^2L^2G^2E(\rho_C EM + T)}{\rho_S MN} = L(F(\theta^{(0)}) - F^*)\frac{E}{T}(\frac{\rho_C ME}{T} + 1).$$

Combining these results, we obtain the desired bound. □

## C.6 Proof of Corollary 1

Proof of Corollary 1. We start with the case of $E = T^{1-\alpha}$. Firstly, in order to have condition (7) hold, we require $T \geq \frac{2\lambda}{\sqrt{\Gamma}}$ and $\frac{E}{T} < \frac{1}{2}\sqrt{\frac{\Gamma}{\lambda}}$, which gives $T \geq \max\left\{\frac{2\lambda}{\sqrt{\Gamma}}, \left(2\sqrt{\frac{\lambda}{\Gamma}}\right)^{\frac{1}{\alpha}}\right\}$. Taking $E = T^{1-\alpha}$ into (10), we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] \leq \frac{3\sqrt{LG^2(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}} + \frac{L(F(\theta^{(0)}) - F^*)}{T^\alpha}(\frac{\rho_C M}{T^\alpha} + 1). \tag{13}$$

Furthermore, when $T$ is large enough such that $T > (\rho_C M)^{\frac{1}{\alpha}}$, we have $\frac{\rho_C M}{T^\alpha} < 1$, which leads to the following

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] \leq \frac{3G\sqrt{L(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}} + \frac{2L(F(\theta^{(0)}) - F^*)}{T^\alpha}. \tag{14}$$

Note that we have required $T \geq \left(2\sqrt{\frac{\lambda}{\Gamma}}\right)^{\frac{1}{\alpha}}$, we have

$$\begin{aligned}
\frac{2L(F(\theta^{(0)}) - F^*)}{T^\alpha} &\leq \frac{L(F(\theta^{(0)}) - F^*)}{\sqrt{\lambda}}\sqrt{\frac{G^2}{L(F(\theta^{(0)}) - F^*)\rho_S MN}} && (15)\\
&= \frac{G\sqrt{L(F(\theta^{(0)}) - F^*)}}{\sqrt{\lambda\rho_S MN}} && (16)\\
&\leq \frac{G\sqrt{L(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}}, && (17)
\end{aligned}$$

where the last inequality is due to the fact that $\lambda \geq 1$. Therefore, when $T \geq \max\left\{\frac{2\lambda}{\sqrt{\Gamma}}, \left(2\sqrt{\frac{\lambda}{\Gamma}}\right)^{\frac{1}{\alpha}}, (\rho_C M)^{\frac{1}{\alpha}}\right\}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] \leq \frac{4G\sqrt{L(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}} = O\left(\frac{1}{\sqrt{\rho_S MN}}\right). \tag{18}$$

Next, we move onto the case of $E = c \cdot T$ for some constant $c < \frac{1}{2}\sqrt{\frac{\Gamma}{\lambda}}$ and $T \geq \frac{2\lambda}{\sqrt{\Gamma}}$. By direct calculation, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] &\leq \frac{3\sqrt{LG^2(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}} + L(F(\theta^{(0)}) - F^*)\frac{E}{T}(\frac{\rho_C ME}{T} + 1) && (19)\\
&\leq \frac{3\sqrt{LG^2(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}} + \frac{L(F(\theta^{(0)}) - F^*)\rho_C M\Gamma}{4\lambda} + \frac{L(F(\theta^{(0)}) - F^*)}{2}\sqrt{\frac{\Gamma}{\lambda}} && (20)\\
&= \frac{3\sqrt{LG^2(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}} + \frac{\rho_C G^2}{4\lambda\rho_S N} + \frac{\sqrt{LG^2(F(\theta^{(0)}) - F^*)}}{2\sqrt{\lambda\rho_S MN}} && (21)\\
&\leq \frac{7\sqrt{LG^2(F(\theta^{(0)}) - F^*)}}{2\sqrt{\rho_S MN}} + \frac{\rho_C G^2}{4\lambda\rho_S N}, && (22)
\end{aligned}$$

where the last inequality follows from the fact of $\lambda \geq 1$. When $M < \frac{4\lambda^2 L(F(\theta^{(0)})-F^*)\rho_S}{G^2\rho_C^2}N$, we obtain that $\frac{\rho_C G^2}{4\lambda\rho_S N} < \frac{\sqrt{LG^2(F(\theta^{(0)})-F^*)}}{2\sqrt{\rho_S MN}}$, which implies that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\theta^{(t-1)})\|_2^2] \leq \frac{4\sqrt{LG^2(F(\theta^{(0)}) - F^*)}}{\sqrt{\rho_S MN}} = O\left(\frac{1}{\sqrt{\rho_S MN}}\right). \tag{23}$$

□