Supplemental Material

We give a table to summarize the content of the supplemental material.

Section	Content
Appendix A	some useful lemmas as technical tools
Appendix B	proof of Theorem 1 for $K = 1$
Appendix C	proof of Theorem 2 for $K < \infty$
Appendix D	proof of Theorem 3 for $K = \infty$
Appendix E	a table for some important notations

Table 1: Outline of the supplemental material.

A USEFUL LEMMAS

In this section, we provide some useful lemmas. Specifically, Lemma 1 is used to support the claim of the convergence speed in Insight 4. Lemmas 2 to 4 are some results about the Gaussian random matrices that can be found in the literature. We want to highlight Lemma 5 as part of our technical novelty, which gives the exact values of terms related to the projection formed by each agent's training inputs. Lemma 6 is used to justify the definition of model error.

Lemma 1. *Recalling the definition of C in Eq.* (32)*, we have*

$$\lim_{t=p\ln p,\ p\to\infty} C^t = 0.$$

Proof. We have $C^t \geq 0$ and

$$\begin{aligned}
C^t &\leq \left(1-\frac{n}{p}\right)^t \quad (\text{since } C \leq \left(1-\frac{n}{p}\right) \text{ because } \left(1-\frac{n}{p}\right)^2 \leq \left(1-\frac{n}{p}\right))\\
&= \left(1+\frac{1}{\frac{p}{n}-1}\right)^{-t} \quad (\text{since } 1-\frac{n}{p} = \frac{1}{1+\frac{1}{\frac{p}{n}-1}})\\
&= \left(1+\frac{1}{\frac{p}{n}-1}\right)^{-p\ln p} \quad (\text{since } t = p\ln p)\\
&= \left(1+\frac{1}{\frac{p}{n}-1}\right)^{-\frac{p}{n}\cdot n\cdot\ln p}\\
&\leq \left(1+\frac{1}{\frac{p}{n}-1}\right)^{-\left(\frac{p}{n}-1\right)\cdot n\cdot\ln p}.
\end{aligned}$$

Notice that

$$\lim_{p\to\infty}\left(1+\frac{1}{\frac{p}{n}-1}\right)^{-\left(\frac{p}{n}-1\right)\cdot n\cdot\ln p} = \lim_{p\to\infty} e^{-n\ln p} = 0,$$

where we use the fact that $\lim_{x\to\infty}(1+x^{-1})^x = e$. The result of this lemma thus follows by the squeeze theorem. □

The result of the following lemma can be found in the literature (e.g., [19, 59]).

Lemma 2. *Consider a random matrix $\mathbf{K} \in \mathbb{R}^{p\times n}$ where p and n are two positive integers and $p > n+1$. Each element of $\mathbf{K}$ is* i.i.d. *according to standard Gaussian distribution. For any fixed vector $\boldsymbol{a} \in \mathbb{R}^p$, we must have*

$$\begin{aligned}
\mathbb{E}\left\|\left(\mathbf{I}_p - \mathbf{K}\left(\mathbf{K}^\top\mathbf{K}\right)^{-1}\mathbf{K}^\top\right)\boldsymbol{a}\right\|^2 &= \left(1-\frac{n}{p}\right)\|\boldsymbol{a}\|^2,\\
\mathbb{E}\left\|\mathbf{K}\left(\mathbf{K}^\top\mathbf{K}\right)^{-1}\mathbf{K}^\top\boldsymbol{a}\right\|^2 &= \frac{n}{p}\|\boldsymbol{a}\|^2.
\end{aligned}$$

The following lemma can be found in Lemma 8 of [60].

Lemma 3. *Consider a random matrix $\mathbf{K} \in \mathbb{R}^{a\times b}$ where $a > b+1$. Each element of $\mathbf{K}$ is* i.i.d. *following standard Gaussian distribution $\mathcal{N}(0,1)$. Consider three Gaussian random vectors $\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \mathbb{R}^a$ and $\boldsymbol{\beta} \in \mathbb{R}^b$ such that $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma_\alpha^2 \mathbf{I}_a)$, $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(d_1^2, d_2^2, \cdots, d_a^2))$, and $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_b)$. Here $\mathbf{K}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are independent of each other. We then must have*

$$\mathbb{E}\left[(\mathbf{K}^\top\mathbf{K})^{-1}\right] = \frac{\mathbf{I}_b}{a-b-1}, \tag{44}$$

$$\mathbb{E}\left\|\mathbf{K}(\mathbf{K}^\top\mathbf{K})^{-1}\boldsymbol{\beta}\right\|^2 = \frac{b\sigma_\beta^2}{a-b-1}, \tag{45}$$

$$\mathbb{E}\left\|(\mathbf{K}^\top\mathbf{K})^{-1}\mathbf{K}^\top\boldsymbol{\alpha}\right\|^2 = \frac{b\sigma_\alpha^2}{a-b-1}, \tag{46}$$

$$\mathbb{E}\left\|(\mathbf{K}^\top\mathbf{K})^{-1}\mathbf{K}^\top\boldsymbol{\gamma}\right\|^2 = \frac{b\sum_{i=1}^a d_i^2}{a(a-b-1)}. \tag{47}$$

The following lemma can be found in [61] and Lemma 13 of [59].

Lemma 4. *Consider a random matrix $\mathbf{K} \in \mathbb{R}^{a\times b}$ whose each element follows* i.i.d. *standard Gaussian distribution (i.e.,* i.i.d. *$\mathcal{N}(0,1)$). We mush have*

$$\mathbb{E}[\mathbf{K}^\top\mathbf{K}] = a\mathbf{I}_b,$$
$$\mathbb{E}[\mathbf{K}\mathbf{K}^\top] = b\mathbf{I}_a,$$
$$\mathbb{E}[\mathbf{K}\mathbf{K}^\top\mathbf{K}\mathbf{K}^\top] = b(b+a+1)\mathbf{I}_a.$$

Lemma 5. *For any $i \in [m]$ and t, we must have*

$$\mathop{\mathbb{E}}_{\mathbf{P}_{(i),t}}\left[\mathbf{P}_{(i),t}\Delta_{t-1}^{K=\infty}\right] = \frac{n_{(i),t}}{p}\Delta_{t-1}^{K=\infty}. \tag{48}$$

Consequently, when $i \neq j$, we have

$$\mathop{\mathbb{E}}_{\mathbf{P}_{(i),t},\mathbf{P}_{(j),t}}\left[\Delta_{t-1}^{K=\infty\top}\mathbf{P}_{(i),t}\mathbf{P}_{(j),t}\Delta_{t-1}^{K=\infty}\right] = \frac{n_{(j),t}n_{(i),t}}{p^2}\left\|\Delta_{t-1}^{K=\infty}\right\|^2.$$

Proof. Let $C \coloneqq \left\|\Delta_{t-1}^{K=\infty}\right\|$. Since we are calculating expected projection of $\Delta_{t-1}^{K=\infty}$ onto the column space of $\mathbf{X}_{(i),t}$, by the symmetry of $\mathbf{X}_{(i),t}$, without loss of generality we let

$$\Delta_{t-1}^{K=\infty} = C \cdot \begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}. \tag{49}$$

Define

$$\tilde{\mathbf{X}}_{(i),t} \coloneqq \begin{bmatrix}-1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}\mathbf{X}_{(i),t}. \tag{50}$$

Since each element of $\mathbf{X}_{(i),t}$ follows *i.i.d.* standard Gaussian distribution, we know that $\tilde{\mathbf{X}}_{(i),t}$ and $\mathbf{X}_{(i),t}$ has identical distributioin. Thus, we have

$$\int \mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top\mathbf{X}_{(i),t})\mathbf{X}_{(i),t}^\top\Delta_{t-1}^{K=\infty}d\mu(\mathbf{X}_{(i),t}) = \int \tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top\tilde{\mathbf{X}}_{(i),t})\tilde{\mathbf{X}}_{(i),t}\Delta_{t-1}^{K=\infty}d\mu(\mathbf{X}_{(i),t}), \tag{51}$$

where $\mu(\mathbf{X}_{(i),t})$ denotes the joint probability distribution of $\mathbf{X}_{(i),t}$.

By Eq. (50), we have

$$\tilde{\mathbf{X}}_{(i),t}^\top\tilde{\mathbf{X}}_{(i),t} = \mathbf{X}_{(i),t}^\top\begin{bmatrix}-1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}\begin{bmatrix}-1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}\mathbf{X}_{(i),t} = \mathbf{X}_{(i),t}^\top\mathbf{X}_{(i),t},$$

$$\mathbf{X}_{(i),t}^\top\Delta_{t-1}^{K=\infty} = [\mathbf{X}_{(i),t}]_{1,:},\ \tilde{\mathbf{X}}_{(i),t}^\top\Delta_{t-1}^{K=\infty} = -[\mathbf{X}_{(i),t}]_{1,:} \text{ (here } [\cdot]_{1,:} \text{ denotes the first row of a matrix).}$$

Thus, we have

$$\tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})^{-1}\tilde{\mathbf{X}}_{(i),t}^\top \Delta_{t-1}^{K=\infty} = -\tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \Delta_{t-1}^{K=\infty}. \tag{52}$$

Therefore, we have

$$\begin{aligned}
&\mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \Delta_{t-1}^{K=\infty} + \tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})^{-1}\tilde{\mathbf{X}}_{(i),t}^\top \Delta_{t-1}^{K=\infty}\\
=&(\mathbf{X}_{(i),t} - \tilde{\mathbf{X}}_{(i),t})(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \Delta_{t-1}^{K=\infty} \text{ (by Eq. (52))}\\
=&\begin{bmatrix} 2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \Delta_{t-1}^{K=\infty} \text{ (by Eq. (50))}\\
=&\begin{bmatrix} 1\\ 0\\ \vdots\\ 0 \end{bmatrix} \begin{bmatrix} 2 & 0 & \cdots & 0 \end{bmatrix} \mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \Delta_{t-1}^{K=\infty}\\
=&2\frac{\Delta_{t-1}^{K=\infty}}{C^2}\Delta_{t-1}^{K=\infty\top}\mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \Delta_{t-1}^{K=\infty} \text{ (by Eq. (49))}\\
=&2\frac{\Delta_{t-1}^{K=\infty}}{C^2}\Delta_{t-1}^{K=\infty\top}\mathbf{P}_{(i),t}\Delta_{t-1}^{K=\infty} \text{ (by Eq. (83))}\\
=&2\frac{\Delta_{t-1}^{K=\infty}}{C^2}\left\|\mathbf{P}_{(i),t}\Delta_{t-1}^{K=\infty}\right\|^2 \text{ (since } \mathbf{P}_{(i),t}^\top\mathbf{P}_{(i),t} = \mathbf{P}_{(i),t} \text{ as } \mathbf{P}_{(i),t} \text{ is an orthogonal projection).}
\end{aligned} \tag{53}$$

Thus, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{X}_{(i),t}}[\mathbf{P}_{(i),t}\Delta_{t-1}^{K=\infty}]\\
=&\int \mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \Delta_{t-1}^{K=\infty} d\mu(\mathbf{X}_{(i),t})\\
=&\frac{1}{2}\int \left(\mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \Delta_{t-1}^{K=\infty} + \tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})\tilde{\mathbf{X}}_{(i),t}^\top \Delta_{t-1}^{K=\infty}\right) d\mu(\mathbf{X}_{(i),t}) \text{ (by Eq. (51))}\\
=&\int \frac{\Delta_{t-1}^{K=\infty}}{C^2}\left\|\mathbf{P}_{(i),t}\Delta_{t-1}^{K=\infty}\right\|^2 d\mu(\mathbf{X}_{(i),t})\\
=&\frac{\Delta_{t-1}^{K=\infty}}{C^2}\mathop{\mathbb{E}}_{\mathbf{X}_{(i),t}}\left\|\mathbf{P}_{(i),t}\Delta_{t-1}^{K=\infty}\right\|^2\\
=&\frac{n_{(i),t}}{p}\Delta_{t-1}^{K=\infty} \text{ (by Lemma 2).}
\end{aligned}$$

The result of this lemma thus follows. □

Lemma 6. *Let the noise in every test sample have zero mean and variance σ^2. For any learning result $\hat{\boldsymbol{w}}$, the mean square test error must equal to $\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|^2 + \sigma^2$. Therefore, the mean squared test error for noise-free test samples equals to the model error $L^{model}(\hat{\boldsymbol{w}}) = \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|^2$.*

Proof. Considering $(\boldsymbol{x}, y)$ as a randomly generated test sample by the ground truth $y = \boldsymbol{x}^\top\boldsymbol{w}^* + \epsilon$, the mean squared error is equal to

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x},y}\left\|\boldsymbol{x}^\top\hat{\boldsymbol{w}} - y\right\| &= \mathop{\mathbb{E}}_{\boldsymbol{x},\epsilon}\left\|\boldsymbol{x}^\top\hat{\boldsymbol{w}} - (\boldsymbol{x}^\top\boldsymbol{w}^* + \epsilon)\right\|^2\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x},\epsilon}\left\|\boldsymbol{x}^\top(\hat{\boldsymbol{w}} - \boldsymbol{w}^*) + \epsilon\right\|^2\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left\|\boldsymbol{x}^\top(\hat{\boldsymbol{w}} - \boldsymbol{w}^*)\right\|^2 + \mathop{\mathbb{E}}_{\epsilon}\|\epsilon\|^2\\
&\quad \text{(since the noise } \epsilon \text{ has zero mean and is independent of other random variables)}\\
&= \left\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\right\|^2 + \sigma^2\\
&\quad \text{(notice that } \boldsymbol{x} \text{ follows standard Gaussian distribution and is independent of } \hat{\boldsymbol{w}}\text{).}
\end{aligned}$$

□

B PROOF OF THEOREM 1

Calculating the gradient of the training loss defined at Eq. (4), we have

$$\begin{aligned}\frac{\partial L(\hat{\boldsymbol{w}})}{\partial \hat{\boldsymbol{w}}} =&\frac{\partial(\boldsymbol{y}-\mathbf{X}^\top\hat{\boldsymbol{w}})}{\partial\hat{\boldsymbol{w}}}\cdot\frac{\partial\frac{1}{2n}\left\|\boldsymbol{y}-\mathbf{X}^\top\hat{\boldsymbol{w}}\right\|^2}{\partial(\boldsymbol{y}-\mathbf{X}^\top\hat{\boldsymbol{w}})}\ \text{ (by the chain rule)}\\ =&-\mathbf{X}\cdot\frac{1}{n}(\boldsymbol{y}-\mathbf{X}^\top\hat{\boldsymbol{w}})\\ =&\frac{1}{n}(\mathbf{X}\mathbf{X}^\top\hat{\boldsymbol{w}}-\mathbf{X}\boldsymbol{y}).\end{aligned}$$

When $K=1$, with step size $\alpha_{(i),t}>0$, we thus have

$$\begin{aligned}\hat{\boldsymbol{w}}_{(i),t}^{K=1}=&\left(\mathbf{I}_p-\frac{\alpha_{(i),t}}{n_{(i),t}}\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top\right)\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1}+\frac{\alpha_{(i),t}}{n_{(i),t}}\mathbf{X}_{(i),t}\boldsymbol{y}_{(i),t}\\ =&\left(\mathbf{I}_p-\frac{\alpha_{(i),t}}{n_{(i),t}}\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top\right)\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1}+\frac{\alpha_{(i),t}}{n_{(i),t}}\mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top\boldsymbol{w}_{(i),t}+\boldsymbol{\epsilon}_{(i),t}\right)\ \text{ (by Eq. (2)).}\end{aligned}$$

Thus, we have

$$\begin{aligned}\hat{\boldsymbol{w}}_{\text{avg},t}^{K=1}=&\frac{1}{\sum_{i\in[m]}n_{(i),t}}\sum_{i\in[m]}n_{(i),t}\hat{\boldsymbol{w}}_{(i),t}^{K=1}\\ =&\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1}+\frac{1}{\sum_{i\in[m]}n_{(i),t}}\sum_{i\in[m]}\alpha_{(i),t}\left(-\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1}+\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top\boldsymbol{w}_{(i),t}+\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}\right).\end{aligned}\tag{54}$$

By Eqs. (3) and (8), we have

$$\begin{aligned}&\Delta_t^{K=1}\\ =&\Delta_{t-1}^{K=1}+\frac{1}{\sum_{i\in[m]}n_{(i),t}}\sum_{i\in[m]}\alpha_{(i),t}\left(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top(\boldsymbol{\gamma}_{(i),t}-\Delta_{t-1}^{K=1})-\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}\right)\\ =&\frac{1}{\sum_{i\in[m]}n_{(i),t}}\sum_{i\in[m]}\Bigg(\underbrace{\left(n_{(i),t}\mathbf{I}_p-\alpha_{(i),t}\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top\right)\Delta_{t-1}^{K=1}}_{\boldsymbol{q}_{1i}}+\underbrace{\alpha_{(i),t}\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top\boldsymbol{\gamma}_{(i),t}}_{\boldsymbol{q}_{2i}}-\underbrace{\alpha_{(i),t}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}}_{\boldsymbol{q}_{3i}}\Bigg)\\ &\text{(since }\Delta_{t-1}^{K=1}=\frac{1}{\sum_{i\in[m]}n_{(i),t}}\sum_{i\in[m]}n_{(i),t}\Delta_{t-1}^{K=1}\text{).}\end{aligned}\tag{55}$$

Considering the three types of terms $\boldsymbol{q}_{1i},\boldsymbol{q}_{2i},\boldsymbol{q}_{3i}$ defined in Eq. (55), by Assumption 1, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_t\boldsymbol{q}_{1i}=n_{(i),t}\left(1-\alpha_{(i),t}\right)\Delta_{t-1}^{K=1},\\ &\mathop{\mathbb{E}}_t\boldsymbol{q}_{2i}=\alpha_{(i),t}n_{(i),t}\boldsymbol{\gamma}_{(i),t},\\ &\mathop{\mathbb{E}}_t\boldsymbol{q}_{3i}=\mathbf{0}.\end{aligned}\tag{56}$$

Notice that we use $\mathbb{E}$ to denote the expectation on all randomness and use $\mathbb{E}_t$ to denote the expectation on the randomness at the t-th round, i.e., on the randomness of $\mathbf{X}_{(i),t}$ and $\boldsymbol{\epsilon}_{(i),t}$ for all $i\in[m]$. By Eqs. (55) and (56), we thus have

$$\mathop{\mathbb{E}}_t\Delta_t^{K=1}=\frac{1}{\sum_{i\in[m]}n_{(i),t}}\sum_{i\in[m]}\left(n_{(i),t}\left(1-\alpha_{(i),t}\right)\Delta_{t-1}^{K=1}+\alpha_{(i),t}n_{(i),t}\boldsymbol{\gamma}_{(i),t}\right).\tag{57}$$

Applying Eq. (57) recursively and recalling Eq. (15), we thus have

$$\mathbb{E}[\Delta_t^{K=1}]=\boldsymbol{g}_t^{K=1}.\tag{58}$$

By Assumption 1, we know that $\boldsymbol{\epsilon}_{(i),t}$ is independent of $\mathbf{X}_{(j),t}$ for all $i,j\in[m]$ and $\mathbb{E}\,\boldsymbol{\epsilon}_{(i),t}=\mathbf{0}$. Thus, we have

$$\mathop{\mathbb{E}}_t[\boldsymbol{q}_{1i}^\top\boldsymbol{q}_{3j}]=\mathop{\mathbb{E}}_t[\boldsymbol{q}_{2i}^\top\boldsymbol{q}_{3j}]=0.$$

Thus, we have

$$
\begin{aligned}
\mathbb{E}_t \left\|\Delta_t^{K=1}\right\|^2 =& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\left(\sum_{i\in[m]}\left(\mathbb{E}_t \|\boldsymbol{q}_{1i}\|^2 + \mathbb{E}_t \|\boldsymbol{q}_{2i}\|^2 + \mathbb{E}_t \|\boldsymbol{q}_{3i}\|^2 + 2\,\mathbb{E}_t[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2i}]\right)\right.\\
&\left.+\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}}\left(\mathbb{E}_t[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{1j}] + \mathbb{E}_t[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j}] + \mathbb{E}_t[\boldsymbol{q}_{1j}^\top \boldsymbol{q}_{2i}] + \mathbb{E}_t[\boldsymbol{q}_{2i}^\top \boldsymbol{q}_{2j}]\right)\right)\\
=& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\left(\sum_{i\in[m]}\left(\mathbb{E}_t \|\boldsymbol{q}_{1i}\|^2 + \mathbb{E}_t \|\boldsymbol{q}_{2i}\|^2 + \mathbb{E}_t \|\boldsymbol{q}_{3i}\|^2 + 2\,\mathbb{E}_t[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2i}]\right)\right.\\
&\left.+\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}}\left(\mathbb{E}_t[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{1j}] + 2\,\mathbb{E}_t[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j}] + \mathbb{E}_t[\boldsymbol{q}_{2i}^\top \boldsymbol{q}_{2j}]\right)\right)\\
&\left(\text{since} \sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} \boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j} + \boldsymbol{q}_{1j}^\top \boldsymbol{q}_{2i} = 2\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} \boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j}\right).
\end{aligned}
\tag{59}
$$

By Lemma 4, for any $i \in [m]$, we have

$$
\begin{aligned}
\mathbb{E}_t \|\boldsymbol{q}_{1i}\|^2 =& \left(n_{(i),t}^2 - 2\alpha_{(i),t} n_{(i),t}^2 + \alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1)\right)\left\|\Delta_{t-1}^{K=1}\right\|^2\\
=& \left(\left(1-\alpha_{(i),t}\right)^2 n_{(i),t}^2 + \alpha_{(i),t}^2 n_{(i),t}(p+1)\right)\left\|\Delta_{t-1}^{K=1}\right\|^2,\\
\mathbb{E}_t \|\boldsymbol{q}_{2i}\|^2 =& \alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2,\\
\mathbb{E}_t \|\boldsymbol{q}_{3i}\|^2 =& \alpha_{(i),t}^2 p n_{(i),t}\sigma_{(i),t}^2,\\
\mathbb{E}_t[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2i}] =& \left(\alpha_{(i),t} n_{(i),t}^2 - \alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1)\right)\Delta_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(i),t}.
\end{aligned}
\tag{60}
$$

Similarly, by Lemma 4, for any $i, j \in [m]$ where $i \neq j$, we have

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{1j}] =& n_{(i),t} n_{(j),t}\left(1-\alpha_{(i),t}\right)\left(1-\alpha_{(j),t}\right)\left\|\Delta_{t-1}^{K=1}\right\|^2,\\
\mathbb{E}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j}] =& \left(\alpha_{(j),t} n_{(i),t} n_{(j),t} - \alpha_{(i),t}\alpha_{(j),t} n_{(i),t} n_{(j),t}\right)\Delta_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(j),t}\\
=& n_{(i),t} n_{(j),t}\alpha_{(j),t}\left(1-\alpha_{(i),t}\right)\Delta_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(j),t},\\
\mathbb{E}[\boldsymbol{q}_{2i}^\top \boldsymbol{q}_{2j}] =& \alpha_{(i),t}\alpha_{(j),t} n_{(i),t} n_{(j),t}\boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{\gamma}_{(j),t}.
\end{aligned}
\tag{61}
$$

Plugging Eqs. (60) and (61) into Eq. (59), we thus have

$$
\begin{aligned}
&\mathbb{E}_t[\left\|\Delta_t^{K=1}\right\|^2]\\
=&\frac{\|\Delta_{t-1}^{K=1}\|^2}{(\sum_{i\in[m]} n_{(i),t})^2}\left(\sum_{i\in[m]}\left((1-\alpha_{(i),t})^2 n_{(i),t}^2 + \alpha_{(i),t}^2 n_{(i),t}(p+1)\right) + \sum_{i\in[m]}\sum_{j\in[m]\setminus\{j\}} n_{(i),t} n_{(j),t}(1-\alpha_{(i),t})(1-\alpha_{(j),t})\right)\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\alpha_{(i),t}^2\left(p n_{(i),t}\sigma_{(i),t}^2 + n_{(i),t}(n_{(i),t}+p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2\right)\\
&+2\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\left(\alpha_{(i),t} n_{(i),t}^2 - \alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1)\right)\Delta_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(i),t}\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}}\left(2n_{(i),t} n_{(j),t}\alpha_{(j),t}\left(1-\alpha_{(i),t}\right)\Delta_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(j),t} + \alpha_{(i),t}\alpha_{(j),t} n_{(i),t} n_{(j),t}\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\right).
\end{aligned}
\tag{62}
$$

Notice that

$$\begin{aligned}
&\left(\sum_{i\in[m]}\left((1-\alpha_{(i),t})^2 n_{(i),t}^2+\alpha_{(i),t}^2 n_{(i),t}(p+1)\right)+\sum_{i\in[m]}\sum_{j\in[m]\backslash\{j\}} n_{(i),t}n_{(j),t}(1-\alpha_{(i),t})(1-\alpha_{(j),t})\right)\\
=&\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\left(\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})^2\right)^2+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\alpha_{(i),t}^2 n_{(i),t}(p+1)\\
=&H_t \text{ (recalling Eq. (16))},
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\backslash\{i\}}\alpha_{(i),t}\alpha_{(j),t}n_{(i),t}n_{(j),t}\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\\
=&\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\left\|\sum_{i\in[m]}\alpha_{(i),t}n_{(i),t}\boldsymbol{\gamma}_{(i),t}\right\|^2+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\alpha_{(i),t}^2 n_{(i),t}(p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2,
\end{aligned}$$

and

$$\begin{aligned}
&2\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\left(\alpha_{(i),t}n_{(i),t}^2-\alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1)\right)\boldsymbol{\Delta}_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(i),t}\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\backslash\{i\}}\left(2n_{(i),t}n_{(j),t}\alpha_{(j),t}\left(1-\alpha_{(i),t}\right)\boldsymbol{\Delta}_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(j),t}\right)\\
=&\frac{2}{(\sum_{i\in[m]} n_{(i),t})^2}\left(\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})\right)\cdot\left(\sum_{i\in[m]} n_{(i),t}\alpha_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(i),t}\right)\\
&-\frac{2\sum_{i\in[m]}\alpha_{(i),t}^2 n_{(i),t}(p+1)\boldsymbol{\Delta}_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(i),t}}{(\sum_{i\in[m]} n_{(i),t})^2}.
\end{aligned}$$

Further, by Eq. (58) and recalling Eq. (17), we thus can rewrite Eq. (62) as

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=1}\right\|^2=H_t\,\mathbb{E}\left\|\boldsymbol{\Delta}_{t-1}^{K=1}\right\|^2+G_t. \tag{63}$$

Applying Eq. (63) recursively, we thus have Eq. (18).

C PROOF OF THEOREM 2

Define

$$\boldsymbol{g}_l^{K<\infty}:=\mathcal{F}\left(l,\boldsymbol{\Delta}_0,\mathrm{seq}_t\left(\frac{\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})^K}{\sum_{i\in[m]} n_{(i),t}}\right),\mathrm{seq}_t\left(\frac{\sum_{i\in[m]} n_{(i),t}\left(1-(1-\alpha_{(i),t})^K\right)\boldsymbol{\gamma}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}}\right)\right) \tag{64}$$

$$\mathcal{A}_{(i),t}:=(1-\alpha_{(i),t})^2+\frac{\alpha_{(i),t}^2(p+1)}{\tilde{n}_{(i),t}}, \tag{65}$$

$$\begin{aligned}
\mathcal{B}_{(i),t,k}:=&\frac{\alpha_{(i),t}^2 p\sigma_{(i),t}^2}{\tilde{n}_{(i),t}}\\
&+\left(\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)+2\alpha_{(i),t}\left(1-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\left(1-(1-\alpha_{(i),t})^{k-1}\right)\right)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2\\
&+2\left(\alpha_{(i),t}-\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)(1-\alpha_{(i),t})^{k-1}\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{g}_{t-1}^{K<\infty},
\end{aligned} \tag{66}$$

$$\mathcal{J}_t \coloneqq \frac{\sum_{i\in[m]} n_{(i),t}^2 \mathcal{A}_{(i),t}^K}{(\sum_{i\in[m]} n_{(i),t})^2} + \frac{\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}(1-\alpha_{(i),t})^K(1-\alpha_{(j),t})^K}{(\sum_{i\in[m]} n_{(i),t})^2}, \tag{67}$$

$$\begin{aligned}\mathcal{Q}_t \coloneqq& \frac{\sum_{i\in[m]} n_{(i),t}^2 \sum_{k=1}^K \mathcal{B}_{(i),t,k}\mathcal{A}_{(i),t}^{K-k}}{(\sum_{i\in[m]} n_{(i),t})^2}\\ &+ \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\Big(2(1-\alpha_{(i),t})^K(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(j),t}^\top \boldsymbol{g}_{t-1}^{K<\infty}\\ &+(1-(1-\alpha_{(i),t})^K)(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\Big).\end{aligned} \tag{68}$$

In the following, we use $\mathbb{E}_k$ to denote the expectation with respect to the randomness in the k-th batch.
We have

$$\begin{aligned}\Delta_t^{K<\infty} =& \boldsymbol{w}^* - \hat{\boldsymbol{w}}_{\text{avg},t}^{K<\infty}\\ =& \boldsymbol{w}^* - \frac{1}{\sum_{i\in[m]} n_{(i),t}}\sum_{i\in[m]} n_{(i),t}\hat{\boldsymbol{w}}_{(i),t}\\ =& \frac{1}{\sum_{i\in[m]} n_{(i),t}}\sum_{i\in[m]} n_{(i),t}(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}) \ \ (\text{since } \boldsymbol{w}^* = \frac{1}{\sum_{i\in[m]} n_{(i),t}}\sum_{i\in[m]} n_{(i),t}\boldsymbol{w}^*).\end{aligned}$$

Thus, we have

$$\begin{aligned}\left\|\Delta_t^{K<\infty}\right\|^2 =& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]} n_{(i),t}^2\left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}\right\|^2\\ &+ \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t})^\top(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(j),t}).\end{aligned} \tag{69}$$

By Assumption 1, we know that at round t, different agents' data are independent with each other. Thus, we have

$$\mathop{\mathbb{E}}_t(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t})^\top(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(j),t}) = \mathop{\mathbb{E}}_t(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t})^\top \mathop{\mathbb{E}}_t(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(j),t}).$$

Thus, by Eq. (69), to calculate $\mathbb{E}_t\left\|\Delta_t^{K<\infty}\right\|^2$, it remains to calculate $\mathbb{E}_t\left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}\right\|^2$ and $\mathbb{E}_t(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t})$ for all $i \in [m]$. To that end, we have

$$\hat{\boldsymbol{w}}_{(i),t,k} = \left(\mathbf{I}_p - \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\right)\hat{\boldsymbol{w}}_{(i),t,k-1} + \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}(\mathbf{X}_{(i),t,k}^\top\boldsymbol{w}_{(i),t} + \boldsymbol{\epsilon}_{(i),t,k}).$$

We thus have

$$\begin{aligned}\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k} =& \left(\mathbf{I}_p - \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\right)(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k-1}) + \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top(\boldsymbol{w}^* - \boldsymbol{w}_{(i),t})\\ &+ \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}\boldsymbol{\epsilon}_{(i),t,k}.\end{aligned} \tag{70}$$

By Lemma 4 and recalling Eq. (3), we thus have

$$\mathop{\mathbb{E}}_k(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k}) = (1-\alpha_{(i),t})(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k-1}) + \alpha_{(i),t}\boldsymbol{\gamma}_{(i),t}. \tag{71}$$

Applying Eq. (71) recursively and recalling that $\hat{\boldsymbol{w}}_{(i),t,0} = \Delta_{t-1}^{K<\infty}$, we thus have

$$\mathop{\mathbb{E}}_{1,2,\cdots,k}(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k}) = (1-\alpha_{(i),t})^k\Delta_{t-1}^{K<\infty} + \left(1-(1-\alpha_{(i),t})^k\right)\boldsymbol{\gamma}_{(i),t}. \tag{72}$$

By letting $k = K$ in Eq. (72) and $\hat{\boldsymbol{w}}_{(i),t,K} = \hat{\boldsymbol{w}}_{(i),t}$, we thus have

$$\mathop{\mathbb{E}}_t(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}) = (1-\alpha_{(i),t})^K\Delta_{t-1}^{K<\infty} + \left(1-(1-\alpha_{(i),t})^K\right)\boldsymbol{\gamma}_{(i),t}. \tag{73}$$

Plugging Eq. (73) into Eq. (69), we thus have

$$\begin{aligned}
\mathop{\mathbb{E}}_{t}\left\|\Delta_t^{K<\infty}\right\|^2 =&\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]} n_{(i),t}^2 \mathop{\mathbb{E}}_{t}\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t}\right\|^2\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\mathop{\mathbb{E}}_{t}(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t})^\top \mathop{\mathbb{E}}_{t}(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(j),t}) \qquad (74)\\
=&\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]} n_{(i),t}^2 \mathop{\mathbb{E}}_{t}\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t}\right\|^2\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\left((1-\alpha_{(i),t})^K(1-\alpha_{(j),t})^K\left\|\Delta_{t-1}^{K<\infty}\right\|^2\right.\\
&+(1-\alpha_{(i),t})^K(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(j),t}^\top\Delta_{t-1}^{K<\infty}+(1-\alpha_{(j),t})^K(1-(1-\alpha_{(i),t})^K)\boldsymbol{\gamma}_{(i),t}^\top\Delta_{t-1}^{K<\infty}\\
&\left.+(1-(1-\alpha_{(i),t})^K)(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\right)\\
=&\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]} n_{(i),t}^2 \mathop{\mathbb{E}}_{t}\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t}\right\|^2\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\left((1-\alpha_{(i),t})^K(1-\alpha_{(j),t})^K\left\|\Delta_{t-1}^{K<\infty}\right\|^2\right.\\
&+2(1-\alpha_{(i),t})^K(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(j),t}^\top\Delta_{t-1}^{K<\infty}\\
&\left.+(1-(1-\alpha_{(i),t})^K)(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\right). \qquad (75)
\end{aligned}$$

Notice that in Eq. (74) we use $\mathbb{E}_t(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t})^\top(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(j),t}) = \mathbb{E}_t(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t})^\top\,\mathbb{E}_t(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(j),t})$ for $i\neq j$, since $\hat{\boldsymbol{w}}_{(i),t}$ and $\hat{\boldsymbol{w}}_{(j),t}$ are independent with respect to the randomness during the local updates at round t.

By Eqs. (5) and (73), we thus have

$$\mathbb{E}\,\Delta_t^{K<\infty}=\frac{\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})^K}{\sum_{i\in[m]} n_{(i),t}}\mathbb{E}\,\Delta_{t-1}^{K<\infty}+\frac{\sum_{i\in[m]} n_{(i),t}\left(1-(1-\alpha_{(i),t})^K\right)\boldsymbol{\gamma}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}}. \qquad (76)$$

Applying Eq. (76) recursively and recalling Eq. (9), we thus have

$$\mathbb{E}[\Delta_l^{K<\infty}]=\boldsymbol{g}_l^{K<\infty}, \qquad (77)$$

where $\boldsymbol{g}_l^{K<\infty}$ is defined in Eq. (64).

By Eq. (70), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{k}\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k}\right\|^2\\
=&(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1})^\top\left(\mathbf{I}_p-2\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top+\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}^2}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\right)(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1})\\
&+\boldsymbol{\gamma}_{(i),t}^\top\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}^2}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}\boldsymbol{\gamma}_{(i),t}+\boldsymbol{\epsilon}_{(i),t,k}^\top\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}^2}\mathbf{X}_{(i),t,k}^\top\mathbf{X}_{(i),t,k}\boldsymbol{\epsilon}_{(i),t,k}\\
&+2\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\boldsymbol{\gamma}_{(i),t}^\top\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\left(\mathbf{I}_p-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\right)(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1})\\
=&\left(1-2\alpha_{(i),t}+\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1}\right\|^2+\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2\\
&+\alpha_{(i),t}^2\frac{p}{\tilde{n}_{(i),t}}\sigma_{(i),t}^2+2\alpha_{(i),t}\left(1-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\boldsymbol{\gamma}_{(i),t}^\top(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1})\ \text{(by Lemma 4)}. \qquad (78)
\end{aligned}$$

Plugging Eq. (72) into Eq. (78), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{1,2,\cdots,k}\left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k}\right\|^2\\
=&\left((1-\alpha_{(i),t})^2 + \frac{\alpha_{(i),t}^2(p+1)}{\tilde{n}_{(i),t}}\right)\mathop{\mathbb{E}}_{1,2,\cdots,k-1}\left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k-1}\right\|^2 + \frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2\\
&+\alpha_{(i),t}^2\frac{p}{\tilde{n}_{(i),t}}\sigma_{(i),t}^2 + 2\alpha_{(i),t}\left(1-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)(1-\alpha_{(i),t})^{k-1}\boldsymbol{\gamma}_{(i),t}^\top\Delta_{t-1}^{K<\infty}\\
&+2\alpha_{(i),t}\left(1-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\left(1-(1-\alpha_{(i),t})^{k-1}\right)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2\\
=&\mathcal{A}_{(i),t}\,\mathbb{E}\left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k-1}\right\|^2 + \mathcal{B}'_{(i),t,k},
\end{aligned}\tag{79}$$

where $\mathcal{A}_{(i),t}$ is defined in Eq. (65) and

$$\begin{aligned}
&\mathcal{B}'_{(i),t,k}\\
:=&\frac{\alpha_{(i),t}^2 p\sigma_{(i),t}^2}{\tilde{n}_{(i),t}}\\
&+\left(\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1) + 2\alpha_{(i),t}\left(1-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\left(1-(1-\alpha_{(i),t})^{k-1}\right)\right)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2\\
&+2\left(\alpha_{(i),t} - \frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)(1-\alpha_{(i),t})^{k-1}\boldsymbol{\gamma}_{(i),t}^\top\Delta_{t-1}^{K<\infty}.
\end{aligned}$$

We also define $\mathcal{B}_{(i),t,k}$ by replacing $\Delta_{t-1}^{K<\infty}$ in $\mathcal{B}'_{(i),t,k}$ with $\mathcal{F}_{t-1}$, i.e., Eq. (66).

Applying Eq. (79) recursively over $k = 1, 2, \cdots, K$, we thus have

$$\mathop{\mathbb{E}}_{t}\left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}\right\|^2 = \mathcal{A}_{(i),t}^K\left\|\Delta_{t-1}^{K<\infty}\right\|^2 + \sum_{k=1}^K \mathcal{B}_{(i),t,k}\mathcal{A}_{(i),t}^{K-k}.\tag{80}$$

Plugging Eqs. (77) and (80) into Eq. (75), we thus have

$$\mathbb{E}\left\|\Delta_t^{K<\infty}\right\|^2 = \mathcal{J}_t\,\mathbb{E}\left\|\Delta_{t-1}^{K<\infty}\right\|^2 + \mathcal{Q}_t,\tag{81}$$

where $\mathcal{J}_t$ is defined in Eq. (67) and $\mathcal{Q}_t$ is defined in Eq. (68).

Applying Eq. (81) recursively, we thus have Eq. (20).

D PROOF OF THEOREM 3

Proof. In the overparameterized situation, after each agent trains to converge, we have

$$\hat{\boldsymbol{w}}_{(i),t}^{K=\infty} = \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top\mathbf{X}_{(i),t}\right)^{-1}\left(\boldsymbol{y}_{(i),t} - \mathbf{X}_{(i),t}^\top\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty}\right) + \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty}.\tag{82}$$

For any $i \in [m]$, we define $\mathbf{P}_{(i),t} \in \mathbb{R}^{p\times p}$ as

$$\mathbf{P}_{(i),t} := \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top\mathbf{X}_{(i),t}\right)^{-1}\mathbf{X}_{(i),t}^\top.\tag{83}$$

(We know $\mathbf{P}_{(i),t}$ is an orthogonal projection since $\mathbf{P}_{(i),t}\mathbf{P}_{(i),t} = \mathbf{P}_{(i),t}$ and $\mathbf{P}_{(i),t}^\top = \mathbf{P}_{(i),t}$.) By Eqs. (2), (82) and (83), we thus have

$$\hat{\boldsymbol{w}}_{(i),t}^{K=\infty} = \mathbf{P}_{(i),t}\boldsymbol{w}_{(i),t} + (\mathbf{I}_p - \mathbf{P}_{(i),t})\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty} + \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top\mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}.\tag{84}$$

We thus have

$$\begin{aligned}
&\Delta_t^{K=\infty}\\
=&\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{\text{avg},t}^{K=\infty} \quad \text{(by Eq. (8))}\\
=&\boldsymbol{w}^* - \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \left(\mathbf{P}_{(i),t}\boldsymbol{w}_{(i),t} + (\mathbf{I}_p - \mathbf{P}_{(i),t})\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty} + \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}\right)\\
&\text{(by Eqs. (5) and (84))}\\
=&\frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \left(\mathbf{P}_{(i),t}(\boldsymbol{w}^* - \boldsymbol{w}_{(i),t}) + (\mathbf{I}_p - \mathbf{P}_{(i),t})(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty}) - \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}\right)\\
&\left(\text{since } \boldsymbol{w}^* = \frac{\sum_{i\in[m]} n_{(i),t}(\mathbf{P}_{(i),t} + \mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{w}^*}{\sum_{i\in[m]} n_{(i),t}}\right)\\
=&\frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \left(\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t} + (\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty} - \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}\right)\\
&\text{(by Eqs. (3) and (8)).}
\end{aligned} \tag{85}$$

For any $i, j \in [m]$, because $\boldsymbol{\epsilon}_{(j),t}$ is independent of $\Delta_{t-1}^{K=\infty}$ and $\mathbf{X}_{(i),t}$, and also because $\boldsymbol{\epsilon}_{(j),t}$ has zero mean (by Assumption 1), we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right)^\top \mathbf{X}_{(j),t}\left(\mathbf{X}_{(j),t}^\top \mathbf{X}_{(j),t}\right)^{-1}\boldsymbol{\epsilon}_{(j),t}\right]\\
=&\mathbb{E}\left[((\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty})^\top \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}\right]\\
=&0,
\end{aligned} \tag{86}$$

and

$$\mathbb{E}\left[\mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}\right] = \mathbf{0}. \tag{87}$$

Since $\mathbf{P}_{(i),t}(\mathbf{I}_p - \mathbf{P}_{(i),t}) = \mathbf{0}$, we have

$$\left(\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right)^\top (\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty} = 0. \tag{88}$$

Thus, by Eqs. (85), (86) and (88), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_t \left\|\Delta_t^{K=\infty}\right\|^2\\
=&\frac{\sum_{i\in[m]} n_{(i),t}^2 \left(\mathbb{E}_t \left\|(\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty}\right\|^2 + \mathbb{E}_t \left\|\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right\|^2 + \mathbb{E}_t \left\|\mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}\right\|^2\right)}{(\sum_{i\in[m]} n_{(i),t})^2}\\
&+ \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2} \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} n_{(i),t} n_{(j),t} \left(\boldsymbol{\gamma}_{(j),t}^\top \mathbf{P}_{(j),t}\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right.\\
&\left.+\Delta_{t-1}^{K=\infty\top}(\mathbf{I}_p - \mathbf{P}_{(j),t})(\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty} + 2\boldsymbol{\gamma}_{(j),t}^\top \mathbf{P}_{(j),t}(\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty}\right).
\end{aligned} \tag{89}$$

For any $i \in [m]$, we have

$$\mathop{\mathbb{E}}_t \left\|\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right\|^2 = \frac{n_{(i),t}}{p}\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 \quad \text{(by Lemma 2)}, \tag{90}$$

$$\mathop{\mathbb{E}}_t \left\|(\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty}\right\|^2 = \left(1 - \frac{n_{(i),t}}{p}\right)\left\|\Delta_{t-1}^{K=\infty}\right\|^2 \quad \text{(by Lemma 2)}, \tag{91}$$

$$\mathop{\mathbb{E}}_t \left\|\mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}\right\|^2 = \frac{n_{(i),t}\sigma_i^2}{p - n_{(i),t} - 1} \quad \text{(by Lemma 3)}. \tag{92}$$

For any $i, j \in [m]$ where $i \neq j$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t}\left[\Delta_{t-1}^{K=\infty\top}(\mathbf{I}_p - \mathbf{P}_{(j),t})(\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty}\right]\\
=&\mathop{\mathbb{E}}_{t}\left[(\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty}\right]^\top \mathop{\mathbb{E}}_{t}\left[(\mathbf{I}_p - \mathbf{P}_{(j),t})\Delta_{t-1}^{K=\infty}\right]\\
&\text{(since } \mathbf{P}_{(i),t} \text{ and } \mathbf{P}_{(j),t} \text{ are independent when } i \neq j)\\
=&\left(1 - \frac{n_{(i),t}}{p}\right)\left(1 - \frac{n_{(j),t}}{p}\right)\left\|\Delta_{t-1}^{K=\infty}\right\|^2 \text{ (by Lemma 5).}
\end{aligned} \tag{93}$$

Similarly, for $i \neq j$, we have

$$\mathop{\mathbb{E}}_{t}\left[\boldsymbol{\gamma}_{(j),t}^\top \mathbf{P}_{(j),t}\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right] = \frac{n_{(i),t}n_{(j),t}}{p^2}\boldsymbol{\gamma}_{(j),t}^\top\boldsymbol{\gamma}_{(i),t} \text{ (by Lemma 5),} \tag{94}$$

and

$$\mathop{\mathbb{E}}_{t}\left[\boldsymbol{\gamma}_{(j),t}^\top \mathbf{P}_{(j),t}(\mathbf{I}_p - \mathbf{P}_{(i),t})\Delta_{t-1}^{K=\infty}\right] = \frac{n_{(j),t}}{p}\left(1 - \frac{n_{(i),t}}{p}\right)\boldsymbol{\gamma}_{(j),t}^\top\Delta_{t-1}^{K=\infty} \text{ (by Lemma 5).} \tag{95}$$

Plugging Eqs. (93) to (95) and (90) to (92) into Eq. (89), we thus have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t}\left\|\Delta_{t}^{K=\infty}\right\|^2\\
=&\frac{\sum_{i\in[m]} n_{(i),t}^2\left(\left(1 - \frac{n_{(i),t}}{p}\right)\left\|\Delta_{t-1}^{K=\infty}\right\|^2 + \frac{n_{(i),t}}{p}\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 + \frac{n_{(i),t}\sigma_{(i),t}^2}{p-n_{(i),t}-1}\right)}{(\sum_{i\in[m]} n_{(i),t})^2}\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\left(\frac{n_{(i),t}n_{(j),t}}{p^2}\boldsymbol{\gamma}_{(j),t}^\top\boldsymbol{\gamma}_{(i),t}\right.\\
&\left.+\left(1 - \frac{n_{(i),t}}{p}\right)\left(1 - \frac{n_{(j),t}}{p}\right)\left\|\Delta_{t-1}^{K=\infty}\right\|^2 + 2\frac{n_{(j),t}}{p}\left(1 - \frac{n_{(i),t}}{p}\right)\boldsymbol{\gamma}_{(j),t}^\top\Delta_{t-1}^{K=\infty}\right).
\end{aligned} \tag{96}$$

By Eq. (85), we also have

$$\mathop{\mathbb{E}}_{t}[\Delta_{t}^{K=\infty}] = \frac{1}{\sum_{i\in[m]} n_{(i),t}}\sum_{i\in[m]} n_{(i),t}\left(\frac{n_{(i),t}}{p}\boldsymbol{\gamma}_{(i),t} + \left(1 - \frac{n_{(i),t}}{p}\right)\Delta_{t-1}^{K=\infty}\right). \tag{97}$$

Applying Eq. (97) recursively, we thus have

$$\mathbb{E}[\Delta_{l}^{K=\infty}] = \boldsymbol{g}_{l}^{K=\infty}, \tag{98}$$

where $\boldsymbol{g}_{l}^{K=\infty}$ is defined in Eq. (23).

By Eqs. (96) and (98), we thus have

$$\mathbb{E}\left\|\Delta_{t}^{K=\infty}\right\|^2 = C_t \cdot \mathbb{E}\left\|\Delta_{t-1}^{K=\infty}\right\|^2 + D_t, \tag{99}$$

where C_t denotes the coefficient of $\left\|\Delta_{t-1}^{K=\infty}\right\|^2$ and D_t denotes the remaining parts. The specific expressions of C_t and D_t are in Eqs. (26) and (27). Applying Eq. (99) recursively, we thus have Eq. (28).

Underparameterized situation

In the underparameterized situation, the convergence point of local steps in each round corresponds to the solution that minimizes the training loss, i.e.,

$$\begin{aligned}
\hat{\boldsymbol{w}}_{(i),t}^{K=\infty} =&(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{y}_{(i),t}\\
=&(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top\boldsymbol{w}_{(i),t} + \boldsymbol{\epsilon}_{(i),t}) \text{ (by Eq. (2))}\\
=&\boldsymbol{w}_{(i),t} + (\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}.
\end{aligned}$$

Also recalling Eqs. (3) and (8), we thus have

$$\Delta_{t}^{K=\infty} = \frac{1}{\sum_{i\in[m]} n_{(i),t}}\sum_{i\in[m]} n_{(i),t}(\boldsymbol{\gamma}_{(i),t} - (\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}). \tag{100}$$

For any $i, j \in [m]$, because $\boldsymbol{\epsilon}_{(j),t}$ is independent of $\mathbf{X}_{(i),t}$ and $\boldsymbol{\epsilon}_{(i),t}$, and also because $\boldsymbol{\epsilon}_{(j),t}$ has zero mean (by Assumption 1), we have

$$\mathbb{E}\left[\boldsymbol{\gamma}_{(j),t}^\top (\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}\right] = 0 \;\text{ for all } i, j \in [m],$$

$$\mathbb{E}\left[\left(\mathbf{X}_{(j),t}\mathbf{X}_{(j),t}^\top)^{-1}\mathbf{X}_{(j),t}\boldsymbol{\epsilon}_{(j),t}\right)^\top (\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}\right] = 0 \;\text{ for all } i \neq j.$$

Thus, by Eq. (100), we have

$$\begin{aligned}\mathbb{E}\left\|\Delta_t^{K=\infty}\right\|^2 =& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]} n_{(i),t}^2\left(\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 + \mathbb{E}\left\|(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}\right\|^2\right)\\ &+ \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\\ =& \left\|\frac{\sum_{i\in[m]} n_{(i),t}\boldsymbol{\gamma}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}}\right\|^2 + \frac{\sum_{i\in[m]}\frac{n_{(i),t}^2 p\sigma_{(i),t}^2}{n_{(i),t}-p-1}}{(\sum_{i\in[m]} n_{(i),t})^2} \text{ (by Eq. (46) in Lemma 3).}\end{aligned}$$

We thus have proven Eq. (29).

The result of this theorem thus follows. □

E A TABLE FOR NOTATIONS

We provide a table of some important notations used in this paper.

symbol	**meaning**
$n_{(i),t}$	number of training samples
$\tilde{n}_{(i),t}$	batch size
p	number of parameters
$\sigma_{(i),t}$	noise level
$\mathbf{X}_{(i),t}$	matrix for input of training samples
$\boldsymbol{y}_{(i),t}$	vector for output of training samples
$\boldsymbol{\epsilon}_{(i),t}$	vector for noise of training samples
$\hat{\boldsymbol{w}}_0$	the pre-trained parameters (initialization)
$\boldsymbol{w}^*$	the learning target
$\boldsymbol{w}_{(i),t}$	the ground-truth of agent i at round t
$\hat{\boldsymbol{w}}_{(i),t}^{K=1}, \hat{\boldsymbol{w}}_{(i),t}, \hat{\boldsymbol{w}}_{(i),t}^{K=\infty}$	the local learning result of agent i at round t
$\hat{\boldsymbol{w}}_{(i),t,k}$	learning result after k-th batch (for $K < \infty$ case)
$\hat{\boldsymbol{w}}_{\text{avg},t}^{K=1}, \hat{\boldsymbol{w}}_{\text{avg},t}^{K<\infty}, \hat{\boldsymbol{w}}_{\text{avg},t}^{K=\infty}$	the FedAvg result at round t
$\left\|\Delta_t^{K=1}\right\|^2, \left\|\Delta_t^{K<\infty}\right\|^2, \left\|\Delta_t^{K=\infty}\right\|^2$	model error
$\|\Delta_0\|^2$	initial (pre-trained) model error
$\alpha_{(i),t}$	learning rate (step size)
$\boldsymbol{\gamma}_{(i),t}$	measurement of heterogeneity

Table 2: Table for some notations.