

A Technical assumptions and Some comments on Section 2

A.1 Technical Assumptions

For the validity of our theoretical results, we require some mild regularity conditions on the loss functions as well as the noise level of each client. We remark, that the following assumptions have appeared extensively in the theoretical analysis of iterative convex optimization algorithms, and here, are merely adapted to the federated learning setting.

Assumption A.1 (Strong convexity). *There exists $\mu > 0$ such that for each $k \in [K]$,*

$$\langle \nabla F_k(\theta) - \nabla F_k(\theta'), \theta - \theta' \rangle \geq \mu |\theta - \theta'|^2, \quad \theta, \theta' \in \mathbb{R}^d.$$

Assumption of strong convexity is common in the analysis of SGD iterates, appearing in [Ruppert \[1988\]](#), [Polyak and Juditsky \[1992\]](#), [Bottou et al. \[2018\]](#), [Chen et al. \[2020\]](#), and as such, Assumption A.1 adapts this condition to the decentralized setting.

Assumption A.2. (*Stochastic Lipschitzness of noisy gradients*) *There exists $L > 0$ such that for each $k \in [K]$,*

$$\mathbb{E}_{\mathcal{P}_k} [|\nabla f_k(\theta, \xi^k) - \nabla f_k(\theta', \xi^k)|^2] \leq L |\theta - \theta'|^2, \quad \theta, \theta' \in \mathbb{R}^d.$$

Assumption A.2 combines the L -smoothness condition on the risk functions F_k , with a stochastic Lipschitz condition on the gradient noise vectors $g_k(\theta, \xi^k) = \nabla F_k(\theta) - \nabla f_k(\theta, \xi^k)$.

Assumption A.3. (*Control on noisy gradients*) *The functions $f_k(\theta, \xi)$ is assumed to be continuously differentiable with respect to θ for any fixed ξ . Moreover, assume that $\max_{k \in [K]} \mathbb{E}[|g_k(\theta, \xi^k)|^p] < \infty$ for some $p \geq 2$.*

Assumption A.3 ensures that Newton-Leibnitz's integration rule holds and consequently, $\sum_{k=1}^K w_k g_k(\theta_t^k, \xi_t^k)$ constitutes a martingale difference sequence adapted to the filtration $\sigma(\Xi_s : s \leq t)$, where $\Xi_s = (\xi_s^1, \dots, \xi_s^K)$. Moreover, Assumptions A.2 and A.3 jointly imply that there exists a constant L_Q such that for all $\theta \in \mathbb{R}^d$

$$\max_{k \in K} |\nabla F_k(\theta) - \nabla_2 F(\theta_K^*)(\theta - \theta_K^*)| \leq L_Q |\theta - \theta_K^*|^2. \quad (\text{A.1})$$

See Lemma 5 of [Sheshukova et al. \[2025\]](#). The assumptions A.2 and A.3 are fairly ubiquitous in the stochastic optimization literature [Zhu et al. \[2023\]](#), [Wei et al. \[2023\]](#), [Li et al. \[2024\]](#). In particular, assumption A.3 is much weaker than the corresponding Assumption A2(p) in [Sheshukova et al. \[2025\]](#).

A.2 Is strong-convexity Assumption A.1 necessary?

It is important to note that Assumption A.1 fails to hold in certain M-estimation problems including logistic regression [Bach \[2010\]](#). [Gu and Chen \[2024\]](#) addressed this issue by invoking a weaker local strong-convexity assumption, also known as the “local concordance” condition.

Assumption A.4 (Local strong convexity). *There exists $\mu^* > 0$ such that $\nabla_2 F(\theta_K^*) \succeq \mu^*$. Moreover, there exists a constant $C > 0$, and compact set $\Phi \subseteq \mathbb{R}^d$, such that for all $\theta_1, \theta_2 \in \Phi$, it holds that*

$$|\varphi'''(u)| \leq C |\theta_1 - \theta_2| \varphi''(u), \quad \text{where } \varphi : u \mapsto F(\theta_1 + u(\theta_2 - \theta_1)), u \in \mathbb{R}.$$

In view of Assumption A.4, for theoretical validity of our results, one requires a projected local SGD updates $\Theta_t = M_\Phi((\Theta_{t-1} - \mathbf{G}_t)C_t)$ instead of (2.2), where M_Φ denotes the projection operator on the set compact Φ . The key difference in the treatment of Assumption A.4 compared to that of Assumption A.1 lies in the analysis of the term $|\theta - \eta \nabla F(\theta)|^2$ for some small enough $\eta > 0$. In particular, a recurring theme of our proofs is to show that

$$|\theta - \theta_K^* - \eta \nabla F(\theta)|^2 \leq (1 - \eta c) |\theta - \theta_K^*|^2 \text{ for some } c > 0, \theta \in \mathbb{R}^d. \quad (\text{A.2})$$

We highlight the different arguments leading up-to (A.2), leveraging Assumptions A.1 and A.4 respectively.

1006 **A.2.1 Proof of (A.2) via Assumptions A.1 and A.2**

1007 Note that

$$\begin{aligned} |\theta - \theta_K^* - \eta \nabla F(\theta)|^2 &= |\theta|^2 - 2\eta(\theta - \theta_K^*)^\top \nabla F(\theta) + \eta^2 |\nabla F(\theta)|^2 \\ &\leq (1 - 2\eta\mu + \eta^2 L^2) |\theta - \theta_K^*|^2, \end{aligned} \quad (\text{A.3})$$

1008 and hence, (A.2) is inferred by choosing η to be small enough. In particular, since we work with
1009 decaying step size $\eta_t \propto t^{-\beta}$, it follows that $1 - 2\eta_t\mu + \eta_t^2 L^2 \leq 1 - \eta_t c$ for some $c > 0$ and all
1010 sufficiently large $t \in \mathbb{N}$.

1011 **A.2.2 Proof of (A.2) via Assumptions A.4, A.2 and $|x| \leq R$**

1012 Fix $\theta \in \mathbb{R}^d$, and choose $\phi(u) = F(\theta_K^* + u(\theta - \theta_K^*))$, $u \in [0, 1]$. Note that $\phi''(0) \geq \mu^* |\theta - \theta_K^*|^2$.
1013 From Assumption A.4, one directly has

$$\phi''(u) \geq \phi''(0) \exp(-C|\theta - \theta_K^*|u),$$

1014 and therefore, recalling $|x| \geq R$

$$\begin{aligned} (\theta - \theta_K^*)^\top \nabla F(\theta) &= \phi'(1) - \phi'(0) \\ &\geq \mu^* |\theta - \theta_K^*|^2 \int_0^1 \exp(-C|\theta - \theta_K^*|u) du \\ &= \mu^* |\theta - \theta_K^*|^2 \frac{1 - \exp(-C|\theta - \theta_K^*|)}{C|\theta - \theta_K^*|} \\ &\geq \mu^* C \exp(-R) |\theta - \theta_K^*|^2, \end{aligned} \quad (\text{A.4})$$

1015 which immediately can be applied to (A.3) to deduce (A.2).

1016 In view of the analysis in Sections A.2.1 and A.2.2 coming to the same conclusion, for the sake of
1017 simplicity, our subsequent theoretical findings are stated and proved using Assumption A.1 only. We
1018 remark that an accompanying result invoking Assumption A.4 and the projected local SGD updates
1019 can easily be obtained via minor modifications of our arguments following Section A.2.2. For a
1020 more detailed discussion on the implications of Assumption A.4, we refer the interested readers to
1021 Assumption 3.4 and the associated remark in Gu and Chen [2024].

1022 **A.3 A comment on step-size**

1023 Our choice of the step-size is motivated from the extensive literature of asymptotics of various
1024 stochastic approximation algorithm. In particular, it is well-known that SGD with a constant step-size
1025 is asymptotically biased Dieuleveut et al. [2020], Li et al. [2024], Glasgow et al. [2022], whereas
1026 central limit theory based on polynomially decaying schedule $\eta_t \propto t^{-\beta}$, $\beta \in (1/2, 1)$ has an extensive
1027 literature for different algorithms. In practice, often a combination of the two kinds of step-size is
1028 used, where a constant-step size algorithm provides a warm start, and after discarding initial few
1029 iterates pre-specified by the fixed *burn-in* period k_0 , local SGD can be run with the polynomially
1030 decaying step-size to ensure appropriate convergence. This is tantamount to the step-size choice
1031 $\eta_t = \eta_0(t - k_0)^{-\beta}$, $t > k_0$, which is also covered by our theory.

1032 **B Proof of Theorems 2.1 and 2.2**

1033 In this section we rigorously derive the Berry-Esseen bounds on \bar{Y}_n and Y_n , as stated in Theorems
1034 2.1 and 2.2 respectively. Similar to the simpler analysis for stochastic gradient descent in Samsonov
1035 et al. [2024], we aim to leverage Theorem 2.1 of Shao and Zhang [2022]. However, the regular
1036 synchronization step, as well as the general connection matrix \mathbf{C} , induces some significant non-
1037 triviality in the problem, requiring, in particular, careful analysis of the difference between client-wise
1038 estimates and the aggregated estimate. Before we delve deeper into the mathematical details, we
1039 summarize the road-map to prove Theorem 2.1 below.

- 1040 • In Section B.1.1, we decompose the local SGD updates into a linear component and the
1041 remainder terms.

- In Section [B.1.2](#), we echo the Lindeberg method, and define a coupling for the remainder terms. In particular, our choice of the coupling is novel, and rooted into the uniqueness of the decentralized setting.
- Finally, in Sections [B.1.3-B.1.7](#), we control the different terms arising out of the application of the abstract Theorem 2.1 of [Shao and Zhang \[2022\]](#) to the steps above. We remark that this is where our treatment diverges from the preceding works proving Berry-Esseen in a stochastic approximation framework. To accommodate an increasing number of clients K as well as to control the error of the each local client-level iterates, we derive and apply the Auxiliary results [4-8](#).

The proof for Theorem [2.2](#) will follow a similar structure.

B.1 Proof of Theorem [2.1](#)

B.1.1 Linearization

Noting that $Y_t = K^{-1}R_t\mathbf{1}$ no matter if $t \in E$ or $t \notin E$, it is easy to observe

$$Y_t = Y_{t-1} - \eta_t \sum_{k=1}^K w_k \nabla f_k(\theta_{t-1}^k, \xi_t^k), \quad t \in [n], \quad Y_0 = K^{-1} \sum_{k=1}^K \theta_0^k. \quad (\text{B.1})$$

Write [\(B.1\)](#) as follows:

$$\begin{aligned} Y_t &= Y_{t-1} - \eta_t \nabla F(Y_{t-1}) + \eta_t \sum_{k=1}^K w_k (\nabla F_k(Y_{t-1}) - \nabla F_k(\theta_{t-1}^k)) + \eta_t \sum_{k=1}^K w_k g_k(\theta_{t-1}^k, \xi_t^k) \\ &= (I - \eta_t A)(Y_{t-1} - \theta_K^*) + \eta_t (A(Y_{t-1} - \theta_K^*) - \nabla F(Y_{t-1})) \\ &\quad + \eta_t \sum_{k=1}^K w_k (\nabla F_k(Y_{t-1}) - \nabla F_k(\theta_{t-1}^k)) + \eta_t \sum_{k=1}^K w_k g_k(\theta_{t-1}^k, \xi_t^k), \end{aligned} \quad (\text{B.2})$$

where $g_k(\theta, \xi) := \nabla F_k(\theta) - \nabla f_k(\theta, \xi)$ denote the gradient noise. Denote $\mathcal{A}_s^t = \prod_{j=s+1}^t (I - \eta_j A)$, $\mathcal{A}_t^t = I$ with $A := \nabla_2 F(\theta_K^*)$, and define $Q_s = \eta_s \sum_{j=s}^n \mathcal{A}_s^j$. Recursively, [\(B.2\)](#) can be simplified to

$$\begin{aligned} Y_t - \theta_K^* &= \mathcal{A}_0^t (Y_0 - \theta_K^*) + \sum_{s=1}^t \eta_s \mathcal{A}_s^t \left((A(Y_{s-1} - \theta_K^*) - \nabla F(Y_{s-1})) + \right. \\ &\quad \left. + \sum_{k=1}^K w_k (\nabla F_k(Y_{s-1}) - \nabla F_k(\theta_{s-1}^k)) + \sum_{k=1}^K w_k g_k(\theta_{s-1}^k, \xi_s^k) \right), \end{aligned} \quad (\text{B.3})$$

which immediately yields,

$$\begin{aligned} \bar{Y}_n - \theta_K^* &= n^{-1} \eta_0^{-1} Q_0 (Y_0 - \theta_K^*) + n^{-1} \sum_{s=1}^n Q_s \mathcal{N}_s + n^{-1} \sum_{s=1}^n Q_s \left((A(Y_{s-1} - \theta_K^*) - \nabla F(Y_{s-1})) \right. \\ &\quad \left. + \sum_{k=1}^K w_k (\nabla F_k(Y_{s-1}) - \nabla F_k(\theta_{s-1}^k)) + \sum_{k=1}^K w_k (g_k(\theta_{s-1}^k, \xi_s^k) - g_k(\theta_K^*, \xi_s^k)) \right), \end{aligned} \quad (\text{B.4})$$

where we define that $\mathcal{N}_t = \sum_{k=1}^K w_k W_t^k$, with $W_t^k = g_k(\theta_K^*, \xi_t^k)$. Let $H = n^{-1/2} \sum_{s=1}^n Q_s \mathcal{N}_s$, and let $\Sigma_n = \mathbb{E}[HH^\top]$. Then, [\(B.4\)](#) can be re-written as

$$\sqrt{n} \Sigma_n^{-1/2} (\bar{Y}_n - \theta_K^*) = W + D_1 + D_2 + D_3 + D_4, \quad (\text{B.5})$$

1061 where

$$W = \Sigma_n^{-1/2} H = \sum_{s=1}^n u_s, \text{ where } u_s = \frac{1}{\sqrt{n}} \Sigma_n^{-1/2} Q_s \mathcal{N}_s, \quad (\text{B.6})$$

$$D_1 = \frac{1}{\sqrt{n}\eta_0} \Sigma_n^{-1/2} Q_0 (Y_0 - \theta_K^*), \quad (\text{B.7})$$

$$D_2 = \frac{1}{\sqrt{n}} \Sigma_n^{-1/2} \sum_{s=1}^n Q_s (A(Y_{s-1} - \theta_K^*) - \nabla F(Y_{s-1})), \quad (\text{B.8})$$

$$D_3 = \frac{1}{\sqrt{n}} \Sigma_n^{-1/2} \sum_{s=1}^n Q_s \left(\sum_{k=1}^K w_k (\nabla F_k(Y_{s-1}) - \nabla F_k(\theta_{s-1}^k)) \right), \quad (\text{B.9})$$

$$D_4 = \frac{1}{\sqrt{n}} \Sigma_n^{-1/2} \sum_{s=1}^n Q_s \left(\sum_{k=1}^K w_k (g_k(\theta_{s-1}^k, \xi_s^k) - g_k(\theta_K^*, \xi_s^k)) \right). \quad (\text{B.10})$$

1062 B.1.2 Definition of the Lindeberg Coupling

1063 Note that

$$|D_3|_2 \leq C \frac{b_2 \sqrt{L}}{K \sqrt{n}} |\Sigma_n^{-1/2}|_F \sum_{s=1}^n \sum_{k=1}^K |Y_{s-1} - \theta_{s-1}^k|_2 \quad (\text{B.11})$$

$$\leq C \frac{b_2 \sqrt{L}}{\sqrt{nK}} |\Sigma_n^{-1/2}|_F \sum_{s=1}^n \sqrt{\sum_{k=1}^K |Y_{s-1} - \theta_{s-1}^k|^2} \quad (\text{B.12})$$

$$= C \frac{b_2 \sqrt{L}}{\sqrt{nK}} |\Sigma_n^{-1/2}|_F \sum_{s=1}^n |\Theta_s(I - J)|_F := \Delta_3. \quad (\text{B.13})$$

In the above series of inequalities, (B.11) follows from $w_k \leq b_2 K^{-1}$, $\max_s |Q_s|_F \leq C$, and Assumption A.2; (B.12) is a trivial consequence of Cauchy-Schwarz inequality. Additionally, define $\Delta_l = |D_l|_2$ for $l = 1, 2, 4$. Let $\Xi_t = (\xi_t^1, \dots, \xi_t^K)$, and for each $i \in [n]$, let us denote

$$\Xi_{t,\{i\}} = \begin{cases} \Xi_t, & t \neq i \\ \Xi'_i := (\xi_t^{1'}, \dots, \xi_t^{K'}), & t = i, \end{cases}$$

1064 where $\xi_t^{k'}, \xi_t^k \stackrel{i.i.d.}{\sim} \mathcal{P}_k$, $k \in [K]$, $t \in [n]$. For each $i \in [n]$, define the coupled DFL iterates as

$$\Theta_{t,\{i\}} = (\Theta_{t-1,\{i\}} - \eta_t G_{t,\{i\}}) C_t, \quad \Theta_{0,\{i\}} = \Theta_0, \quad (\text{B.14})$$

1065 where $G_{t,\{i\}} = K(w_1 \nabla f_1(\theta_{t-1,\{i\}}^1, \xi_{t,\{i\}}^1), \dots, w_K \nabla f_K(\theta_{t-1,\{i\}}^K, \xi_{t,\{i\}}^K))$. Let $Y_{t,\{i\}} =$

1066 $K^{-1} \Theta_{t,\{i\}} \mathbf{1}$. Based on (B.14), we can define coupled versions of Δ_l , $l = 2, 3, 4$ as follows:

$$\Delta_{2,\{i\}} = \frac{1}{\sqrt{n}} \left| \Sigma_n^{-1/2} \sum_{s=1}^n Q_s (A(Y_{s-1,\{i\}} - \theta_K^*) - \nabla F(Y_{s-1,\{i\}})) \right|, \quad (\text{B.15})$$

$$\Delta_{3,\{i\}} = C \frac{b_2 \sqrt{L}}{\sqrt{nK}} |\Sigma_n^{-1/2}|_F \sum_{s=1}^n |\Theta_{s,\{i\}}(I - J)|_F, \quad (\text{B.16})$$

$$\Delta_{4,\{i\}} = \frac{1}{\sqrt{n}} \left| \Sigma_n^{-1/2} \sum_{s=1}^n Q_s \left(\sum_{k=1}^K w_k (g_k(\theta_{s-1,\{i\}}^k, \xi_{s,\{i\}}^k) - g_k(\theta_K^*, \xi_{s,\{i\}}^k)) \right) \right|. \quad (\text{B.17})$$

1067 Note that $D_{1,\{i\}} = D_1$ for all $i \in [n]$. With these definitions, along with the fact that $\mathbb{E}[WW^\top] = I$
1068 allows us to apply Shao and Zhang [2022], Theorem 2.1 on (B.5) to obtain

$$d_C(\sqrt{n} \Sigma_n^{-1/2} (\bar{Y}_n - \theta_K^*), Z) \leq c_1 \sqrt{d} \Upsilon_n + \mathbb{E}[|W| |\Delta_n|] + \sum_{i=1}^n \mathbb{E}[|u_i| |\Delta_n - \Delta_{n,\{i\}}|], \quad (\text{B.18})$$

1069 where $Z \sim N(0, I)$, $\Upsilon_n = \sum_{s=1}^n \mathbb{E}[|u_s|^3]$, $\Delta_n = \sum_{l=1}^4 |\Delta_l|$, and $\Delta_{n,\{i\}} = \sum_{l=1}^4 \Delta_{l,\{i\}}$.

1070 **B.1.3 Bound on $\sum_{i=1}^n \mathbb{E}[|u_i| |\Delta_n - \Delta_{n,\{i\}}|]$**

1071 Recall that $\max_k |\text{Var}[W_s^k]| = O(1)$. Clearly \mathcal{N}_s are i.i.d. and $\mathbb{E}[\mathcal{N}_s \mathcal{N}_s^\top] = \sum_{k=1}^K w_k^2 \text{Var}[W_s^k]$,
 1072 which directly implies $|\Sigma_n| = O(K^{-1})$ in view of the fact $w_k \asymp K^{-1}$ for $k \in [K]$. Therefore, from
 1073 (B.6) it follows $\mathbb{E}[|u_s|^2] = O(1/n)$, and consequently,

$$\sum_{i=1}^n \mathbb{E}[|u_i| |\Delta - \Delta_{n,\{i\}}|] \lesssim \frac{1}{\sqrt{n}} \sum_{l=2}^4 \sum_{i=1}^n \sqrt{\mathbb{E}[|\Delta_l - \Delta_{l,\{i\}}|^2]}. \quad (\text{B.19})$$

1074 We will deal with the three terms in the right side of (B.19) one-by-one.

1075 **Bound on $\Delta_2 - \Delta_{2,\{i\}}$** We start with controlling $\mathbb{E}[|\Delta_2 - \Delta_{2,\{i\}}|^2]$. It is easy to see from (B.8) and
 1076 (B.15) that

$$\begin{aligned} \mathbb{E}[|\Delta_2 - \Delta_{2,\{i\}}|^2] &\lesssim \frac{K}{n} \mathbb{E} \left[\left| \sum_{s=i}^n (A(Y_s - Y_{s,\{i\}}) - \nabla F(Y_s) + \nabla F(Y_{s,\{i\}})) \right|^2 \right] \\ &\lesssim K \sum_{s=i}^n \mathbb{E}[|Y_s - Y_{s,\{i\}}|^4] = O(Kn^{1-4\beta} - Ki^{1-4\beta}), \end{aligned} \quad (\text{B.20})$$

1077 where (B.20) follows from Cauchy-Schwarz inequality and Proposition 8.

1078 **B.1.4 Bound on $\Delta_3 - \Delta_{3,\{i\}}$**

1079 Note that, since $\Theta_{t,\{i\}} = \Theta_t$ for all $t < i$, hence we must have

$$\begin{aligned} \mathbb{E}[|\Delta_3 - \Delta_{3,\{i\}}|^2] &\lesssim \frac{1}{n} \mathbb{E} \left[\left(\sum_{s=i}^n |(\Theta_s - \Theta_{s,\{i\}})(I - J)|_F \right)^2 \right] \\ &\leq \sum_{s=i}^n \mathbb{E}[|(\Theta_s - \Theta_{s,\{i\}})(I - J)|_F^2] \\ &= \mathbb{E}[|(\Theta_i - \Theta_{i,\{i\}})(I - J)|_F^2] + \sum_{s=i+1}^n \mathbb{E}[|(\Theta_s - \Theta_{s,\{i\}})(I - J)|_F^2]. \end{aligned} \quad (\text{B.21})$$

1080 Note that,

$$\begin{aligned} \mathbb{E}[|(\Theta_i - \Theta_{i,\{i\}})(I - J)|_F^2] &= \eta_i^2 \mathbb{E}[|(G_i - G_{i,\{i\}})(C_i - J)|_F^2] \\ &\leq \eta_i^2 \mathbb{E} \left[\left| \sum_{k=1}^K w_k (g_k(\theta_{i-1}^k, \xi_i^k) - g_k(\theta_{i-1}^k, \xi_i^{k'})) \right|^2 \right] \\ &\leq 2\eta_i^2 \mathbb{E} \left[\sum_{k=1}^K w_k g_k(\theta_{i-1}^k, \xi_i^k)^2 \right] = O\left(\frac{\eta_i^2}{K}\right). \end{aligned} \quad (\text{B.22})$$

1081 Hence, Proposition 5 and (B.22) simultaneously imply via (B.21) that

$$\mathbb{E}[|\Delta_3 - \Delta_{3,\{i\}}|^2] \lesssim \frac{\eta_i^2}{K} + K \sum_{s=i+1}^n \eta_s^4 = O(i^{-2\beta} K^{-1} + Kn^{1-4\beta} - Ki^{1-4\beta}). \quad (\text{B.23})$$

1082 **B.1.5 Bound on $\Delta_4 - \Delta_{4,\{i\}}$**

1083 This term is the simplest to deal with. In view of the facts (i) $\sum_{k=1}^K w_k (g_k(\theta_{t-1}^k, \xi_t^k) -$
 1084 $g_k(\theta_{t-1,\{i\}}^k, \xi_t^k))$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_t = \sigma(\Xi_s : s \leq$
 1085 $t) \vee \sigma(\Xi'_i)$, and (ii) for a fixed t , $g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1,\{i\}}^k, \xi_t^k)$ are independent conditional on

1086 \mathcal{F}_{t-1} , one readily obtains

$$\begin{aligned}
\mathbb{E}[|\Delta_4 - \Delta_{4,\{i\}}|^2] &\lesssim \frac{K}{n} \left(\sum_{s=i}^{n-1} \mathbb{E} \left[\sum_{k=1}^K w_k(g_k(\theta_s^k, \xi_{s+1}^k) - g_k(\theta_{s,\{i\}}^k, \xi_{s+1}^k))|^2 \right] \right. \\
&\quad \left. + \mathbb{E} \left[\sum_{k=1}^K w_k(g_k(\theta_{i-1}^k, \xi_i^k) - g_k(\theta_K^*, \xi_i^k) - g_k(\theta_{i-1}^k, \xi_i^{k'}) + g_k(\theta_K^*, \xi_i^{k'}))^2 \right] \right) \\
&\lesssim \frac{K}{n} (K^{-2} \sum_{s=i}^{n-1} \sum_{k=1}^K \mathbb{E}[|\theta_s^k - \theta_{s,\{i\}}^k|^2] + 2K^{-2} \mathbb{E}[\sum_{k=1}^K |\theta_{i-1}^k - \theta_K^*|^2]) \\
&\lesssim \frac{1}{nK} \left(\sum_{s=i}^{n-1} \mathbb{E}[|\Theta_s(I - J)|_F^2 + K|Y_s - Y_{s,\{i\}}|^2 + |\Theta_{s,\{i\}}(I - J)|_F^2] \right. \\
&\quad \left. + 2\mathbb{E}[|\Theta_{i-1}(I - J)|_F^2 + K|Y_{i-1} - \theta_K^*|^2] \right) \\
&\lesssim \frac{\eta_i}{nK} + \frac{1}{n} \sum_{s=i-1}^{n-1} \eta_s^2 = O\left(\frac{i^{-\beta}}{nK} + \frac{n^{1-2\beta} - i^{1-2\beta}}{n}\right), \tag{B.24}
\end{aligned}$$

1087 where (B.24) follows from Theorem 2.(ii) and Lemma S16 of Gu and Chen [2024], and Proposition
1088 4.

1089 Combining (B.20), (B.23) and (B.24), for (B.19) we obtain

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}[|u_i| \mid \Delta - \Delta_{n,\{i\}}] &\lesssim \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{i^{-\beta/2}}{\sqrt{nK}} + \frac{i^{-\beta}}{\sqrt{K}} + \sqrt{K}(n^{1/2-2\beta} - i^{1/2-2\beta}) + \frac{n^{\frac{1}{2}-\beta} - i^{\frac{1}{2}-\beta}}{\sqrt{n}} \right) \\
&\lesssim \frac{n^{\frac{1}{2}-\beta} + n^{-\frac{\beta}{2}}}{\sqrt{K}} + \sqrt{K}n^{1-2\beta}. \tag{B.25}
\end{aligned}$$

1090 **B.1.6 Bound on $\mathbb{E}[|W| \mid \Delta_n]$**

1091 From $\mathbb{E}[|u_s|^2] \lesssim n^{-1/2}$, we have $\mathbb{E}[|W|^2] = O(1)$, where $O(\cdot)$ hides constants involving d . More-
1092 over, we also have $\mathbb{E}[|\Delta_n|^2] \lesssim \sum_{l=1}^4 \mathbb{E}[|\Delta_l|^2]$. For Δ_1 , observe that from (B.7),

$$\mathbb{E}[|D_1|^2] \lesssim \frac{K}{n} |Q_0|_F^2 \lesssim \frac{K}{n} \exp(-C_\beta n^{1-\beta}) \text{ for some constant } C_\beta > 0. \tag{B.26}$$

1093 On the other hand, for Δ_2 , we can invoke Assumption A.3 and Minkowsky's inequality to deduce

$$\begin{aligned}
\sqrt{\mathbb{E}[|D_2|^2]} &\lesssim \sqrt{\frac{K}{n}} \sum_{s=1}^n \sqrt{\mathbb{E}[|A(Y_{s-1} - \theta_K^*) - \nabla F(Y_{s-1})|^2]} \\
&\lesssim \sqrt{\frac{K}{n}} \sum_{s=1}^n \sqrt{\mathbb{E}[|Y_{s-1} - \theta_K^*|^4]} \\
&\lesssim \sqrt{\frac{K}{n}} \sum_{s=1}^n \left(\frac{\eta_s}{K} + \eta_s^2 \right) = O\left(\frac{n^{\frac{1}{2}-\beta}}{\sqrt{K}} + \sqrt{K}n^{\frac{1}{2}-2\beta}\right), \tag{B.27}
\end{aligned}$$

1094 where (B.27) follows from Proposition 7. Moving on, for Δ_3 , it is immediate that

$$\sqrt{\mathbb{E}[\Delta_3^2]} \lesssim \frac{1}{\sqrt{n}} \sum_{s=1}^n \sqrt{\mathbb{E}[|\Theta_s(I - J)|_F^2]} \lesssim \frac{1}{\sqrt{n}} \sum_{s=1}^n \eta_s \sqrt{K} = O(n^{\frac{1}{2}-\beta} \sqrt{K}). \tag{B.28}$$

1095 Finally, for Δ_4 , we recall the argument in (B.24) to provide

$$\begin{aligned}
\mathbb{E}[|D_4|^2] &\lesssim \frac{K}{n} \sum_{s=1}^n \mathbb{E}\left[\left|\sum_{k=1}^K w_k(g_k(\theta_{s-1}^k, \xi_s^k) - g_k(\theta_K^*, \xi_s^k))\right|^2\right] \\
&\lesssim \frac{1}{nK} \sum_{s=1}^n (|\Theta_{s-1}(I - J)|_F^2 + K|Y_{t-1} - \theta_K^*|^2) \\
&\lesssim \frac{1}{nK} \sum_{s=1}^n (\eta_s^2 K + \eta_s) \\
&\lesssim \frac{n^{-\beta}}{K} + n^{-2\beta}.
\end{aligned} \tag{B.29}$$

1096 Combining (B.26)-(B.29), we obtain

$$\mathbb{E}[|W||\Delta_n|] \lesssim \sqrt{\frac{K}{n}} \exp(-C_\beta n^{1-\beta}) + n^{\frac{1}{2}-\beta} \sqrt{K} + \frac{n^{-\frac{\beta}{2}}}{\sqrt{K}} + n^{-\beta}. \tag{B.30}$$

1097 B.1.7 Final Berry Esseen bound

1098 Note that, for any $t \in [n]$, Pinelis-Rosenthal inequality (Theorem 4.1 of Pinelis [1994]) applies to
1099 yield

$$\mathbb{E}[|\Sigma_n^{-1/2} \mathcal{N}_t|^3] \lesssim K^{3/2} \mathbb{E}\left[\left|\sum_{k=1}^K w_k W_t^k\right|^3\right] = O(K^{-1/2}),$$

1100 which immediately implies that

$$\Upsilon_n \lesssim \sum_{s=1}^n \mathbb{E}[n^{-3/2} |\Sigma_n^{-1/2} \mathcal{N}_s|^3] = O\left(\frac{1}{\sqrt{nK}}\right). \tag{B.31}$$

1101 Therefore, combining (B.25), (B.30) and (B.31), we have that

$$d_C(\sqrt{n} \Sigma_n^{-1/2} (\bar{Y}_n - \theta_K^*), Z) \lesssim \frac{1}{\sqrt{nK}} + n^{\frac{1}{2}-\beta} \sqrt{K} + \frac{n^{-\frac{\beta}{2}}}{\sqrt{K}},$$

1102 which completes the proof.

1103 B.2 Proof of Theorem 2.2

1104 Let $\Gamma = \text{Var}(\sum_{s=1}^n \eta_s \mathcal{A}_s^n \mathcal{N}_s) = \sum_{s=1}^n \eta_s^2 \mathcal{A}_s^n \mathcal{V}_K \mathcal{A}_s^{n\top}$. Clearly, $|\Gamma|_F \lesssim \frac{n^{-\beta}}{K}$. Define $v_s =$
1105 $\Gamma^{-1/2} \eta_s \mathcal{A}_s^n \mathcal{N}_s$. Recall (B.3), and rewrite it as

$$\Gamma^{-1/2} (Y_n - \theta_K^*) = \tilde{W} + \tilde{D}_1 + \tilde{D}_2 + \tilde{D}_3 + \tilde{D}_4, \tag{B.32}$$

1106 where

$$\begin{aligned}
\tilde{W} &= \sum_{s=1}^n v_s, \\
\tilde{D}_1 &= \Gamma^{-1/2} \mathcal{A}_0^n (Y_0 - \theta_K^*), \\
\tilde{D}_2 &= \Gamma^{-1/2} \sum_{s=1}^n \eta_s \mathcal{A}_s^n (A(Y_{s-1} - \theta_K^*) - \nabla F(Y_{s-1})), \\
\tilde{D}_3 &= \Gamma^{-1/2} \sum_{s=1}^n \eta_s \mathcal{A}_s^n \sum_{k=1}^K w_k (\nabla F_k(Y_{s-1}) - \nabla F_k(\theta_{s-1}^k)), \\
\tilde{D}_4 &= \Gamma^{-1/2} \sum_{s=1}^n \eta_s \mathcal{A}_s^n \sum_{k=1}^K w_k g_k(\theta_{s-1}^k, \xi_s^k).
\end{aligned}$$

1107 Let $\tilde{\Delta}_l = |\tilde{D}_l|_2$ for $l = 1, 2, 4$, and let

$$\tilde{\Delta}_3 := |\Gamma^{-1/2}|_F K^{-1/2} \sum_{s=1}^n \eta_s \mathcal{A}_s^n |\Theta_s(I - J)|_F.$$

1108 Note that $|\tilde{D}_3|_2 \leq \tilde{\Delta}_3$. The terms $|\tilde{\Delta}_l|$ and $|\tilde{\Delta}_{l,\{i\}}|$ are defined and controlled very similarly to
 1109 Theorem 2.1, and the details are omitted. The $\frac{n^{-\beta/2}}{\sqrt{K}}$ appears by controlling $\tilde{\Upsilon}_n := \sum_{s=1}^n \mathbb{E}[|v_s|^3]$,
 1110 which we show below. Since $|H|_F \lesssim n^{-\beta} K^{-1}$, therefore

$$\sum_{s=1}^n \mathbb{E}[|v_s|^3] \lesssim K^{3/2} n^{\frac{3\beta}{2}} \sum_{s=1}^n \eta_s^3 |\mathcal{A}_s^n|^3 \mathbb{E}[|\mathcal{N}_s^3|] \lesssim n^{-\beta/2} K^{-1/2}.$$

1111 B.3 Application of Section 2: weighted multiplier bootstrap

1112 In the context of vanilla SGD, Fang et al. [2018], Fang [2019], Sheshukova et al. [2025] introduced a
 1113 novel multiplier bootstrap paradigm that precludes the necessity of estimating Σ_n while performing
 1114 inference. In this section, we adapt this approach for the particular decentralized setting, and hint
 1115 towards the applicability of our Berry-Esseen theorems 2.1 and 2.2. Specifically, for each client
 1116 $k \in [K]$, let \mathbb{P}_W^k be a distribution of a random variable with $\mathbb{E}[W^k] = 1$ and $\text{Var}[W^k] = \sigma_k^2$,
 1117 $W^k \sim \mathbb{P}_W^k$. For the validity of the bootstrap procedure, we assume that, for all k , $\sigma_k^2 \leq C_0$ for some
 1118 constant $C_0 > 0$. Moreover, we assume that W^k 's uniformly bounded, i.e. that there exists universal
 1119 constants $c_1, c_2 > 0$ such that $c_1 \leq W^k \leq c_2$ for all $k \in [K]$ almost surely. For $b \in [B]$ where B is
 1120 the number of bootstrap samples, consider the augmented local SGD updates

$$\Theta_t^{\{b\}} = (\Theta_{t-1}^{\{b\}} - \eta_t G_t^{\{b\}}) C_t,$$

1121 where

$$G_t^{\{b\}} = K(w_1 W_{t,1}^{\{b\}} \nabla f_1(\theta_{t-1}^1, \xi_t^1), \dots, w_K W_{t,K}^{\{b\}} \nabla f_K(\theta_{t-1}^K, \xi_t^K))$$

1122 and for each $k \in [K]$, $\{W_{t,k}^{\{b\}}\}$ are i.i.d. random variables from \mathbb{P}_W^k , $t \in [n], b \in [B]$. For each
 1123 $b \in [B]$, define $\bar{Y}_n^{\{b\}} = n^{-1} K^{-1} \sum_{t,k} \theta_t^{k\{b\}}$. Suppose $\mathcal{F}_n := \sigma(\xi_s^k : s \in [n], k \in [K])$. Following
 1124 standard arguments (see Theorem 3 of Sheshukova et al. [2025]), adapting the proof of Theorem 2.1
 1125 as well as off-the-self Gaussian comparison results Chernozhukov et al. [2017], Devroye et al. [2018],
 1126 it is possible to show that

$$\sup_{A \in \mathcal{B}(\mathbb{R}^d) : A \text{ convex}} \left| \mathbb{P}(\sqrt{n}(\bar{Y}_n^{\{b\}} - \bar{Y}_n) \in A | \mathcal{F}_n) - \mathbb{P}(\sqrt{n}(\bar{Y}_n - \theta_K^*) \in A) \right| \lesssim n^{1/2-\beta} \sqrt{K},$$

1127 modulo logarithmic factors, with high probability with respect to \mathcal{F}_n . This result enables one to
 1128 approximate the distribution of $\bar{Y}_n - \theta_K^*$ via the bootstrap samples $\bar{Y}_n^{\{b\}}$. We remark that this approach
 1129 works when our focus is on \bar{Y}_n ; we do not expect this multiplier bootstrap to approximate the entire
 1130 process $\{Y_t\}$. We leave the detailed derivations to future work, since the focus of this paper is on
 1131 establishing the fundamental Gaussian approximation theorems.

1132 C Proof of Theorem 2.3

1133 Recall that $\Sigma_n = n^{-1} \sum_{s=1}^n Q_s \mathcal{V}_K Q_s^\top$, and $\Sigma = K A^{-1} \mathcal{V}_K A^{-\top}$. We aim to decompose $\Sigma_n -$
 1134 $K^{-1} \Sigma$ into manageable terms, and then control them piecemeal. To be precise, write

$$\Sigma_n - K^{-1} \Sigma = \frac{1}{n} \sum_{s=1}^n ((Q_s - A^{-1}) \mathcal{V}_K A^{-\top} + A^{-1} \mathcal{V}_K (Q_s - A^{-1})^\top + (Q_s - A^{-1}) \mathcal{V}_K (Q_s - A^{-1})^\top).$$

1135 Crucial to our proof is the observation that $\mathcal{A}_s^t - \mathcal{A}_{s-1}^t = \eta_s A \mathcal{A}_s^t$ for all $s, t \in [n]$. Therefore,

$$\sum_{s=1}^n Q_s = \sum_{s=1}^n \sum_{j=s}^n \eta_s \mathcal{A}_s^j = \sum_{j=1}^n \sum_{s=1}^j \eta_s \mathcal{A}_s^j = \sum_{j=1}^n \sum_{s=1}^j A^{-1} (\mathcal{A}_s^j - \mathcal{A}_{s-1}^j) = -A^{-1} \sum_{j=1}^n (I - \mathcal{A}_0^j), \quad (\text{C.1})$$

where the last equality is via a telescoping argument. From (C.1), we obtain $\sum_{s=1}^n (Q_s - A^{-1}) = -A^{-1} \sum_{j=1}^n \mathcal{A}_0^j$. Consequently, recalling that $|\mathcal{V}_K|_F = K^{-1/2}$, it is immediate that

$$n^{-1} |A^{-1} \mathcal{V}_K|_F \left| \sum_{s=1}^n (Q_s - A^{-1}) \right| \lesssim \frac{1}{n\sqrt{K}} \sum_{j=1}^n |\mathcal{A}_0^j| \lesssim \frac{1}{n\sqrt{K}} \int_1^n \exp(-x^{1-\beta}) dx = O(K^{-1/2} n^{\beta-1}).$$

Moving on, the term $(Q_s - A^{-1}) \mathcal{V}_K (Q_s - A^{-1})^\top$ can be similarly controlled by $K^{-1/2} n^{\beta-1}$ from Lemma A.5 of Wu et al. [2024] (also see Lemma 11 and 12 of Sheshukova et al. [2025]). This completes the proof of (2.7). Finally, (2.8) follows from (2.7) on the account of Proposition 9, and the fact that Σ_n is positive-definite, and hence maps a convex set to a convex set.

D Proofs of Section 3

In this section, we derive the time-uniform Gaussian approximation results Theorem 3.1 and 3.2. Our proofs are divided into four successive approximation steps. We summarize our arguments in the following. In the step I, we control the difference between the aggregated and the local client-level local SGD updates. In step II, we replace the martingale structure of the gradient noise by i.i.d. mean zero noise. In step III, we further linearize the local SGD updates, which we finally approximate by a stochastically linear Gaussian process such as (3.3) or (3.5) in Step IV.

D.1 Proof of Theorem 3.1

D.1.1 Step I

Consider $\Theta_0^\circ = (\theta_K^*, \dots, \theta_K^*) \in \mathbb{R}^{d \times K}$, and let

$$\Theta_t^\circ = (\Theta_{t-1}^\circ - \eta_t \mathbf{G}_t^\circ) C_t, \quad (\text{D.1})$$

where \mathbf{G}_t° is defined similar to \mathbf{G}_t in (2.2), but with $\theta_t^{k^\circ}$ instead of θ_t^k . Moreover, let $Y_t^\circ = K^{-1} \Theta_t^\circ \mathbf{1} \in \mathbb{R}^K$. Suppose $R_t = (r_t^1, \dots, r_t^K) = \Theta_{t-1} - \eta_t \mathbf{G}_t$, and R_t° is defined likewise. Recall (B.1) and (B.2). Define two more intermediate oracle processes:

$$\tilde{Y}_t = \tilde{Y}_{t-1} - \eta_t \nabla F(\tilde{Y}_{t-1}) + \eta_t \sum_{k=1}^K w_k g_k(\theta_{t-1}^k, \xi_t^k), \quad t \in [n], \quad \tilde{Y}_0 = Y_0 \quad (\text{D.2})$$

$$\tilde{Y}_t^\circ = \tilde{Y}_{t-1}^\circ - \eta_t \nabla F(\tilde{Y}_{t-1}^\circ) + \eta_t \sum_{k=1}^K w_k g_k(\theta_{t-1}^{k^\circ}, \xi_t^k), \quad t \in [n], \quad \tilde{Y}_0^\circ = Y_0. \quad (\text{D.3})$$

For a random variable X , let $\|X\| = (\mathbb{E}[|X|^2])^{1/2}$ be the random variable \mathcal{L}_2 -norm. Then,

$$\|Y_t - \tilde{Y}_t\| \leq \|(Y_{t-1} - \tilde{Y}_{t-1}) - \eta_t (\nabla F(Y_{t-1}) - \nabla F(\tilde{Y}_{t-1}))\| + \eta_t \left\| \sum_{k=1}^K w_k (F_k(Y_{t-1}) - F_k(\theta_{t-1}^k)) \right\| := A + B. \quad (\text{D.4})$$

Now, for the term A in (D.4), invoking Assumptions A.1 and A.2, it is easy to observe that

$$\begin{aligned} A^2 &= \|Y_{t-1} - \tilde{Y}_{t-1}\|^2 + \eta_t^2 \|\nabla F(Y_{t-1}) - \nabla F(\tilde{Y}_{t-1})\|^2 - 2\eta_t \mathbb{E}[(Y_{t-1} - \tilde{Y}_{t-1})^\top (\nabla F(Y_{t-1}) - \nabla F(\tilde{Y}_{t-1}))] \\ &\leq (1 - 2\eta_t \mu + \eta_t^2 L^2) \|Y_{t-1} - \tilde{Y}_{t-1}\|^2. \end{aligned} \quad (\text{D.5})$$

On the other hand, for B in (D.4), Assumption A.2 entails,

$$B^2 \leq \eta_t^2 K \sum_{k=1}^K w_k^2 \|F_k(Y_{t-1}) - F_k(\theta_{t-1}^k)\|^2 \leq \eta_t^2 b_2^2 L^2 K^{-1} \mathbb{E}[\|\Theta_t(I - J)\|_F^2] = O(\eta_t^4), \quad (\text{D.6})$$

through an application of Lemma S.16 of Supplement of Gu and Chen [2024]. Combining (D.5) and (D.6) and choosing a $c > \mu \vee L$, it must hold that

$$\|Y_t - \tilde{Y}_t\| \leq (1 - \eta_t c) \|Y_{t-1} - \tilde{Y}_{t-1}\| + O(\eta_t^2),$$

1160 which readily yields

$$\|Y_t - \tilde{Y}_t\| = O(\eta_t). \quad (\text{D.7})$$

1161 Very similarly, one can show that $\|Y_t^\circ - \tilde{Y}_t^\circ\| = O(\eta_t)$. Finally it remains to show that \tilde{Y}_t and \tilde{Y}_t° is
1162 approximately close. We show it as follows.

$$\begin{aligned} & \|\tilde{Y}_t - \tilde{Y}_t^\circ\|^2 \\ &= \|\tilde{Y}_{t-1} - \tilde{Y}_{t-1}^\circ - \eta_t(\nabla F(\tilde{Y}_{t-1}) - \nabla F(\tilde{Y}_{t-1}^\circ))\|^2 + \eta_t^2 \sum_{k=1}^K w_k^2 \|g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1}^{k^\circ}, \xi_t^k)\|^2 \\ &\leq (1 - \eta_t c) \|\tilde{Y}_t - \tilde{Y}_t^\circ\|^2 + 4\eta_t^2 b_2^2 K^{-2} \sum_{k=1}^K (\|\theta_{t-1}^k - Y_{t-1}\|^2 + \|Y_{t-1} - \theta_K^*\|^2 + \|\theta_{t-1}^{k^\circ} - Y_{t-1}^\circ\|^2 + \|Y_{t-1}^\circ - \theta_K^*\|^2) \end{aligned} \quad (\text{D.8})$$

$$\leq (1 - \eta_t c) \|\tilde{Y}_t - \tilde{Y}_t^\circ\|^2 + 4\eta_t^2 b_2^2 (2 \frac{\eta_t}{K^2} + 2 \frac{\eta_t^2}{K}) \quad (\text{D.10})$$

$$\leq (1 - \eta_t c) \|\tilde{Y}_t - \tilde{Y}_t^\circ\|^2 + O(\frac{\eta_t^3}{K^2} + \frac{\eta_t^4}{K}). \quad (\text{D.11})$$

1163 Here, (D.8) employs Assumption A.3 to deduce that $g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1}^{k^\circ}, \xi_t^k)$ are mean-zero
1164 martingale differences adapted to $\mathcal{F}_t := \sigma(\xi_s^k, s \leq t, k \in [K])$; moreover, since $\{\xi_t^k\}_{k=1}^K$ are
1165 independent for a fixed t , hence $g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1}^{k^\circ}, \xi_t^k)$ are also uncorrelated. Additionally,
1166 (D.9) uses a treatment analogous to (D.5) along with applying Assumption A.2 to the g_k terms; and
1167 (D.10) involves applications of Lemmas S.16 and Theorem 2(ii) from Gu and Chen [2024]. Finally,
1168 (D.11) immediately implies that

$$\|\tilde{Y}_t - \tilde{Y}_t^\circ\|^2 = O(\eta_t^2 K^{-2} + \eta_t^3 K^{-1}),$$

1169 which, coupled with (D.7), yields,

$$\max_{1 \leq t \leq n} \left| \sum_{s=1}^t (Y_s - \tilde{Y}_s^\circ) \right| \leq \sum_{t=1}^n |Y_t - \tilde{Y}_t^\circ| = O_{\mathbb{P}}(n^{1-\beta}). \quad (\text{D.12})$$

1170 D.1.2 Step II

1171 Moving on, we approximate \tilde{Y}_t° by another oracle descent sequence, given by

$$Y_t^\dagger = Y_{t-1}^\dagger - \eta_t \nabla F(Y_{t-1}^\dagger) + \eta_t \sum_{k=1}^K w_k g_k(\theta_K^*, \xi_t^k), \quad Y_0^\dagger = Y_0. \quad (\text{D.13})$$

1172 Importantly, (D.13) can be leveraged to linearize the original sequence Y_{t-1} in (B.1). Before we
1173 proceed in that direction, we still need to approximate \tilde{Y}_t° by Y_t^\dagger . From (D.3) and (D.13), it follows
1174 very similarly to (D.8)-(D.11), that,

$$\begin{aligned} & \|\tilde{Y}_t^\circ - Y_t^\dagger\|^2 \\ &= \|\tilde{Y}_{t-1}^\circ - Y_{t-1}^\dagger - \eta_t(\nabla F(\tilde{Y}_{t-1}^\circ) - \nabla F(Y_{t-1}^\dagger))\|^2 + \eta_t^2 \sum_{k=1}^K w_k^2 \|g_k(\theta_{t-1}^{k^\circ}, \xi_t^k) - g_k(\theta_K^*, \xi_t^k)\|^2 \\ &\leq (1 - \eta_t c) \|\tilde{Y}_{t-1}^\circ - Y_{t-1}^\dagger\|^2 + \eta_t^2 L^2 b_2^2 K^{-2} (\mathbb{E}[\|\Theta_t^\circ(I - J)|_F^2] + C\eta_t + C\eta_t^2 K) \\ &\leq (1 - \eta_t c) \|\tilde{Y}_{t-1}^\circ - Y_{t-1}^\dagger\|^2 + O(\eta_t^3 K^{-2} + \eta_t^4 K^{-1}), \end{aligned}$$

1175 which immediately yields $\|\tilde{Y}_t^\circ - Y_t^\dagger\| = O(\eta_t K^{-1} + \eta_t^{3/2} K^{-1/2})$. Similar to (D.12), here too we
1176 finally obtain

$$\max_{1 \leq t \leq n} \left| \sum_{s=1}^t (\tilde{Y}_s^\circ - Y_s^\dagger) \right| = O_{\mathbb{P}}(n^{1-\beta} K^{-1} + n^{1-3\beta/2} K^{-1/2}). \quad (\text{D.14})$$

1177 D.1.3 Step III

1178 Define \tilde{Y}_t^\dagger as

$$\tilde{Y}_t^\dagger = Y_{t-1}^\dagger - \eta_t \nabla F(\tilde{Y}_{t-1}^\dagger) + \eta_t \sum_{k=1}^K w_k g_k(\theta_K^*, \xi_t^k), \quad \tilde{Y}_0 = \theta_K^*.$$

1179 Then it trivially follows that

$$\begin{aligned} \|Y_t^\dagger - \tilde{Y}_t^\dagger\|^2 &= \|(Y_{t-1}^\dagger - \tilde{Y}_{t-1}^\dagger) - \eta_t(\nabla F(Y_{t-1}^\dagger) - \nabla F(\tilde{Y}_{t-1}^\dagger))\|^2 \\ &\leq (1 - \eta_t c) \|Y_{t-1}^\dagger - \tilde{Y}_{t-1}^\dagger\|^2 \lesssim \exp(-t^{1-\beta}) |Y_0 - \theta_K^*|^2, \end{aligned} \quad (\text{D.15})$$

1180 which implies $\max_{1 \leq t \leq n} |\sum_{s=1}^t (Y_s^\dagger - \tilde{Y}_s^\dagger)| = O_{\mathbb{P}}(1)$, since $\int_1^n \exp(-t^{1-\beta}) dt = O(1)$. Moving
1181 on, to linearize \tilde{Y}_t^\dagger , write (D.13) as

$$\tilde{Y}_t^\dagger - \theta_K^* = (I - \eta_t A)(\tilde{Y}_{t-1}^\dagger - \theta_K^*) - \eta_t(\nabla F(\tilde{Y}_{t-1}^\dagger) - A(\tilde{Y}_{t-1}^\dagger - \theta_K^*)) + \eta_t \sum_{k=1}^K w_k g_k(\theta_K^*, \xi_t^k), \quad (\text{D.16})$$

1182 where $A = \nabla_2 F(\theta_K^*)$. Note that, Assumption A.1 along with $\sum w_k = 1$ implies that $A \succeq \mu I$.
1183 Mimicking (D.16), define

$$Y_t^\diamond = (I - \eta_t A)Y_{t-1}^\diamond + \eta_t \sum_{k=1}^K w_k g_k(\theta_K^*, \xi_t^k), \quad Y_0^\diamond = 0. \quad (\text{D.17})$$

1184 Clearly, it follows that

$$\begin{aligned} \mathbb{E}[\|\tilde{Y}_t^\dagger - \theta_K^* - Y_t^\diamond\|] &\leq (1 - \eta_t \mu) \mathbb{E}[\|\tilde{Y}_{t-1}^\dagger - \theta_K^* - Y_{t-1}^\diamond\|] + \eta_t \mathbb{E}[\|\nabla F(\tilde{Y}_{t-1}^\dagger) - A(\tilde{Y}_{t-1}^\dagger - \theta_K^*)\|] \\ &\leq (1 - \eta_t \mu) \mathbb{E}[\|\tilde{Y}_{t-1}^\dagger - \theta_K^* - Y_{t-1}^\diamond\|] + L_Q \eta_t \mathbb{E}[\|\tilde{Y}_{t-1}^\dagger - \theta_K^*\|^2] \end{aligned} \quad (\text{D.18})$$

$$\leq (1 - \eta_t \mu) \mathbb{E}[\|\tilde{Y}_{t-1}^\dagger - \theta_K^* - Y_{t-1}^\diamond\|] + O(\eta_t^2 K^{-1} + \eta_t^3), \quad (\text{D.19})$$

1185 where, (D.18) follows from Assumption A.3, and (D.19) is a trivial consequence of Theorem 2.(ii)
1186 of Gu and Chen [2024]. Finally, (D.19) yields that

$$\max_{1 \leq t \leq n} \left| \sum_{s=1}^t (Y_s^\dagger - \theta_K^* - Y_s^\diamond) \right| = O_{\mathbb{P}}(n^{1-\beta} K^{-1/2} + n^{1-2\beta}). \quad (\text{D.20})$$

1187 D.1.4 Step IV

1188 Note that Y_t^\diamond is a linear process, and thus we can hope to bear down standard strong invariance
1189 principle results Komlós et al. [1975], Sakhanenko [2006], Göttsche and Zaitsev [2009] on it to yield an
1190 asymptotically optimal Gaussian approximation. In particular, let $\mathcal{V}_K = \text{Var}(\sum_{k=1}^K w_k g_k(\theta_K^*, \xi^k))$,
1191 $\xi^k \sim \mathcal{P}_k, k \in [K]$. Note that, Assumption 4.2 in Gu and Chen [2024] can also be summarized as
1192 $\|K \mathcal{V}_K\|_F \asymp 1$. We pursue two different type of Gaussian approximation. Let $W_t^k = g_k(\theta_K^*, \xi_t^k)$, and
1193 $\mathcal{N}_t = \sum_{k=1}^K w_k W_t^k$. By Göttsche and Zaitsev [2009], there exists i.i.d. $Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} N(0, K \mathcal{V}_K)$,
1194 such that $\max_{1 \leq t \leq n} |\sum_{s=1}^t (\sqrt{K} \mathcal{N}_s - Z_s)| = o_{\mathbb{P}}(n^{1/p})$. Write (D.17) as

$$Y_t^\diamond = (I - \eta_t A)Y_{t-1}^\diamond + \eta_t \mathcal{N}_t,$$

1195 which immediately yields

$$\sum_{s=1}^t Y_s^\diamond = \sum_{s=1}^t \eta_s \mathcal{N}_s B_{s,t}, \quad B_{s,t} = \sum_{j=s}^t \mathcal{A}_s^j, \quad (\text{D.21})$$

1196 where $\mathcal{A}_s^t = \prod_{j=s+1}^t (I - \eta_j A)$, $\mathcal{A}_t^t = I$. Mimicking Y_t^\diamond , define $Y_{t,1}^G$ as in (3.3), to which we can
1197 simplify $\sum_{s=1}^t Y_{s,1}^G = K^{-1/2} \sum_{s=1}^t \eta_s Z_s \sum_{j=s}^t \mathcal{A}_s^j$. Note that,

$$\max_{1 \leq t \leq n} \left| \sum_{s=1}^t (Y_s^\diamond - Y_{s,1}^G) \right| \leq \max_{1 \leq t \leq n} \Omega_t \max_{1 \leq t \leq n} \left| \sum_{s=1}^t (\mathcal{N}_s - K^{-1/2} Z_s) \right| = o_{\mathbb{P}}\left(\max_{1 \leq t \leq n} \Omega_t n^{1/p} K^{-1/2}\right). \quad (\text{D.22})$$

1198 where $\Omega_t := |B_{1,t}|_F + \sum_{s=2}^t |B_{s,t} - B_{s-1,t}|_F$. The proof of (3.4) is completed after combining
1199 (D.12), (D.14), (D.20) and (D.22) in view of Proposition 2.

1200 D.2 Proof of Theorem 3.2 and Proposition 1

1201 In this subsection, we pursue a finer, client-level Gaussian approximation, with slight sacrifice to the
 1202 optimality in terms of error rate. In particular, the steps I, II and II from the proof of Theorem 3.1
 1203 carry forward verbatim. Consequently, it enables us to invoke from Theorem 2.1 of Mies and Steland
 1204 [2023] so that for each $k \in [K]$, there exists $Z_1^k, \dots, Z_n^k \sim N(0, \text{Var}(W^k))$, such that

$$\max_{1 \leq t \leq n} \sum_{k=1}^K \left| \sum_{s=1}^t (W_s^k - Z_s^k) \right| = O_{\mathbb{P}}(K^{\frac{3}{4} - \frac{1}{2p}} n^{\frac{1}{2p} + \frac{1}{4}} \sqrt{\log n}). \quad (\text{D.23})$$

1205 Here W^k denotes a generic $g_k(\theta_K^*, \xi^k)$. For $\tilde{\theta}_t^{1^G}, \dots, \tilde{\theta}_t^{K^G} \in \mathbb{R}^d$, define $\tilde{\Theta}_t^G = (\tilde{\theta}_t^{1^G} \dots \tilde{\theta}_t^{K^G}) \in$
 1206 $\mathbb{R}^{d \times K}$, and simultaneously define the recursion (3.5). Letting $Y_{t,2}^G = K^{-1} \tilde{\Theta}_t^G \mathbf{1}$, one arrives at the
 1207 recursion

$$Y_{t,2}^G = (I - \eta_t A) Y_{t-1,2}^G + \eta_t \sum_{k=1}^K w_k Z_t^k, \quad (\text{D.24})$$

1208 to which, from (D.23), one has

$$\begin{aligned} \max_{1 \leq t \leq n} \left| \sum_{s=1}^t (Y_t^\diamond - Y_{t,2}^G) \right| &\leq \max_{1 \leq t \leq n} \Omega_t \left| \sum_{s=1}^t \sum_{k=1}^K w_k (W_t^k - Z_t^k) \right| \leq b_2 \log n \max_{1 \leq t \leq n} K^{-1} \sum_{k=1}^K \left| \sum_{s=1}^t (W_t^k - Z_t^k) \right| \\ &= O_{\mathbb{P}}((n/K)^{\frac{1}{4} + \frac{1}{2p}} (\log n)^{3/2}), \end{aligned} \quad (\text{D.25})$$

1209 where the second inequality is due to Proposition 2 and $\max_k w_k \leq b_2 K^{-1}$. Again, we conclude
 1210 (3.6) in light of (D.12), (D.14), (D.20) and (D.25). Finally, Proposition 1 follows trivially from
 1211 Theorems 3.1 and 3.2.

1212 E Auxiliary propositions

1213 In this section, we present some technical results required to prove our main theorems. Propositions 2
 1214 and 3 relates the local SGD updates to its asymptotic covariance matrices. In particular, Proposition
 1215 2 controls the implicit total variation between the linearized local SGD updates, and as such, is
 1216 crucial in deriving the time-uniform approximations Aggr-GA and Client-GA.

1217 **Proposition 2.** Let $A \in \mathbb{R}^{d \times d}$ be a positive definite matrix with smallest eigen value $\lambda_{\min} > 0$, and
 1218 define $\mathcal{A}_s^t = \prod_{j=s+1}^t (I - \eta_j A)$, $\mathcal{A}_t^t = I$. If

$$\Omega_t := |B_{1,t}|_F + \sum_{s=2}^t |B_{s,t} - B_{s-1,t}|_F,$$

1219 then it holds that $\max_{1 \leq t \leq n} \Omega_t = O(\log n)$.

1220 **Proposition 3.** Let $B_{s,t}$ be as in Proposition 2. Then, for all $s \geq 1, t \geq s$, it holds that

$$|B_{s,t} - A^{-1}|_F \lesssim s^{-1} + \exp(-c_\beta(t^{1-\beta} + s^{1-\beta})),$$

1221 where c_β is some constant depending on β, λ_{\min} .

1222 Propositions 4-8 characterizes the various properties of the local SGD updates and its difference
 1223 with its corresponding Lindeberg coupling. These results hold under the conditions of Theorem 2.1,
 1224 and can be considered as its building blocks.

1225 **Proposition 4.** Under the assumptions of Theorem 2.1, it holds that for all $i \in [n], t \geq i$, it holds that

$$\mathbb{E}[|Y_t - Y_{t,\{i\}}|^2] = O(\eta_t^2). \quad (\text{E.1})$$

1226 **Proposition 5.** Under the conditions of Theorem 2.1, for all $i \in [n], t > i$, it holds that

$$\mathbb{E}[|(\Theta_t - \Theta_{t,\{i\}})(I - J)|_F^2] = O(\eta_t^4 K).$$

1227 **Proposition 6.** Under the conditions of Theorem 2.1, it holds that

$$\mathbb{E}[|\Theta_t(I - J)|_F^4] = O(\eta_t^4 K^2). \quad (\text{E.2})$$

1228 **Proposition 7.** Under the conditions of Theorem 2.1, it holds that

$$\mathbb{E}[|Y_t - \theta_K^*|^4] = O\left(\frac{\eta_t^2}{K^2} + \eta_t^4\right). \quad (\text{E.3})$$

1229 **Proposition 8.** Grant the assumptions of Theorem 2.1. Then, for $t \geq i$, it holds that

$$\mathbb{E}[|Y_t - Y_{t,\{i\}}|^4] = O(\eta_t^4). \quad (\text{E.4})$$

1230 Proposition 9 is a typical Gaussian comparison results that relates the finite-sample covariance Σ_n to
 1231 the asymptotic covariance Σ in terms of the corresponding normal distributions. This result enables
 1232 theorem 2.3 to reflect the computation-communication trade-off of Remark 2.2.

1233 **Proposition 9** (Gaussian comparison lemma; Theorem 1.1, Devroye et al. [2018]). Let Σ_1 and Σ_2
 1234 be positive definite covariance matrices in $\mathbb{R}^{p \times p}$. Let $X \sim \mathcal{N}(0, \Sigma_1)$ and $Y \sim \mathcal{N}(0, \Sigma_2)$. Then

$$d_{\text{TV}}(X, Y) \leq \frac{3}{2} \left\| \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} - I_p \right\|_F.$$

1235 E.1 Proofs of the auxiliary results

1236 *Proof of Proposition 2.* In the following, all \lesssim solely depend on β and λ_{\min} . Observe that for $s < t$,
 1237 $B_{s,t} = \eta_s(I + \eta_{s+1}^{-1}(I - \eta_s A)B_{s+1,t})$. Therefore, it can be written that

$$B_{s,t} - B_{s-1,t} = \frac{\eta_s - \eta_{s-1}}{\eta_s} B_{s,t} + \eta_{s-1}(B_{s,t}A - I) := I_1 + I_2. \quad (\text{E.5})$$

1238 The I_1 term is relatively straightforward by noting that $\max_{s,t} |B_{s,t}|_F = O(1)$, and $|\frac{\eta_s - \eta_{s-1}}{\eta_s}| =$
 1239 $O(s^{-1})$. On the other hand, for I_2 , Proposition 3 instructs that

$$\eta_{s-1}|B_{s,t}A - I|_F \lesssim s^{-\beta-1} + s^{-\beta} \exp(-c_\beta(t^{1-\beta} + s^{1-\beta})). \quad (\text{E.6})$$

1240 Combining (E.5) and (E.6), we obtain

$$|B_{s,t} - B_{s-1,t}|_F \lesssim s^{-1} + s^{-\beta} \exp(-c_\beta(t^{1-\beta} + s^{1-\beta})),$$

1241 which immediately shows

$$\Omega_t \lesssim \sum_{s=1}^t s^{-1} + \exp(-c_\beta t^{1-\beta}) \int_1^t s^{-\beta} \exp(c_\beta s^{1-\beta}) \lesssim \log t,$$

1242 which completes the proof. \square

1243 *Proof of Proposition 3.* Decompose $B_{s,t} = \eta_s \sum_{j=s}^t \mathcal{A}_j^s$ as

$$B_{s,t} - A^{-1} = -A^{-1} \mathcal{A}_s^t + \sum_{j=s}^{t-1} (\eta_{j+1} - \eta_s) \mathcal{A}_j^s + \eta_s \mathcal{A}_s^t, \quad (\text{E.7})$$

1244 where the sum $\sum_{j=s}^{t-1}$ is interpreted as 0 if $s = t$. For the term $A^{-1} \mathcal{A}_s^t$ in (E.7), we deduce
 1245 $|\mathcal{A}_s^t| \lesssim \exp(-c_\beta(t^{1-\beta} + s^{1-\beta}))$. On the other hand,

$$\sum_{j=s}^{t-1} (\eta_{j+1} - \eta_s) |\mathcal{A}_j^s|_F \lesssim s^{-\beta-1} \exp(c_\beta s^{1-\beta}) \sum_{j=s}^{t-1} (j - s) \exp(-j^{1-\beta}) \lesssim s^{-1}.$$

1246 This completes the proof. \square

1247 *Proof of Proposition 4.* From (B.2) we write

$$Y_t - Y_{t,\{i\}} = \begin{cases} \eta_i \sum_{k=1}^K (g_k(\theta_{i-1}^k, \xi_i^k) - g_k(\theta_{i-1}^k, \xi_i^{k'})), & t = i, \\ (Y_{t-1} - Y_{t-1,\{i\}}) - \eta_t (\nabla F(Y_{t-1}) - \nabla F(Y_{t-1,\{i\}})) \\ \quad + \eta_t \sum_{k=1}^K w_k (\nabla F_k(Y_{t-1}) - \nabla F(Y_{t-1,\{i\}}) - \nabla F_k(\theta_{t-1}^k) + \nabla F_k(\theta_{t-1,\{i\}}^k)) \\ \quad + \eta_t \sum_{k=1}^K w_k (g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1,\{i\}}^k, \xi_t^k)), & t > i. \end{cases} \quad (\text{E.8})$$

1248 Clearly, when $t = i$, we have trivially that $\mathbb{E}[|Y_i - Y_{i,\{i\}}|^2] = O(\eta_i^2 K^{-1})$. Hence, we focus on
 1249 $t > i$. Consider the observation that $\sum_{k=1}^K w_k (g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1,\{i\}}^k, \xi_t^k))$ is a martingale
 1250 difference sequence adapted to the filtration $\mathcal{F}_t = \sigma(\Xi_s : s \leq t) \vee \sigma(\Xi'_i)$. Moreover, for a fixed t ,
 1251 $g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1,\{i\}}^k, \xi_t^k)$ are independent conditional on \mathcal{F}_{t-1} . Therefore, rewriting (E.8) as

$$Y_t - Y_{t,\{i\}} = T_1 + T_2 + T_3 \quad (\text{E.9})$$

1252 with

$$\begin{aligned} T_1 &= (Y_{t-1} - Y_{t-1,\{i\}}) - \eta_t (\nabla F(Y_{t-1}) - \nabla F(Y_{t-1,\{i\}})), \\ T_2 &= \eta_t \sum_{k=1}^K w_k (\nabla F_k(Y_{t-1}) - \nabla F(Y_{t-1,\{i\}}) - \nabla F_k(\theta_{t-1}^k) + \nabla F_k(\theta_{t-1,\{i\}}^k)), \text{ and,} \\ T_3 &= \eta_t \sum_{k=1}^K w_k (g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1,\{i\}}^k, \xi_t^k)), \end{aligned}$$

1253 it is easy to see that $\mathbb{E}[T_1^\top T_3] = \mathbb{E}[T_2^\top T_3] = 0$. Consequently, from (E.8), one computes

$$\mathbb{E}[|Y_t - Y_{t,\{i\}}|^2] = \mathbb{E}[|T_1|^2] + \mathbb{E}[|T_2|^2] + \mathbb{E}[|T_3|^2] + 2\mathbb{E}[T_1^\top T_2]. \quad (\text{E.10})$$

1254 Now all that is required is to build a recursion by analyzing (E.10) term-by-term. Note that standard
 1255 arguments invoking Assumptions A.2 and A.1 yields

$$\mathbb{E}[|T_1|^2] \leq (1 - \eta_t c) \mathbb{E}[|Y_{t-1} - Y_{t-1,\{i\}}|^2]. \quad (\text{E.11})$$

1256 On the other hand, for T_3 , we proceed as follows:

$$\begin{aligned} \mathbb{E}[|T_3|^2] &= \eta_t^2 \sum_{k=1}^K w_k^2 \mathbb{E}[|g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1,\{i\}}^k, \xi_t^k)|^2] \\ &\lesssim \eta_t^2 \sum_{k=1}^K w_k^2 \left(\mathbb{E}[|\theta_{t-1}^k - Y_{t-1}|^2] + \mathbb{E}[|\theta_{t-1,\{i\}}^k - Y_{t-1,\{i\}}|^2] + \mathbb{E}[|Y_{t-1} - Y_{t-1,\{i\}}|^2] \right) \\ &\lesssim \frac{\eta_t^2}{K} \mathbb{E}[|Y_{t-1} - Y_{t-1,\{i\}}|^2] + O\left(\frac{\eta_t^4}{K}\right), \end{aligned} \quad (\text{E.12})$$

1257 where $O(\eta_t^4 K^{-1})$ bound in (E.12) is derived upon applying Lemma S16 of Gu and Chen [2024].

1258 Very similarly, one can bound T_2 as

$$\mathbb{E}[|T_2|^2] \lesssim \eta_t^2 K \sum_{k=1}^K w_k^2 \mathbb{E}[|Y_{t-1} - \theta_{t-1}^k|^2 + |Y_{t-1,\{i\}} - \theta_{t-1,\{i\}}^k|^2] = O(\eta_t^4). \quad (\text{E.13})$$

1259 Finally we tackle the cross-product term in (E.10). Again, Assumption A.2 and yet another application
 1260 of Lemma S16 of Gu and Chen [2024] produces

$$\begin{aligned} \mathbb{E}[T_1^\top T_2] &\leq \eta_t \sqrt{\mathbb{E}[|T_1|^2]} \sqrt{\mathbb{E}\left[\left|\sum_{k=1}^K w_k (\nabla F_k(Y_{t-1}) - \nabla F(Y_{t-1,\{i\}}) - \nabla F_k(\theta_{t-1}^k) + \nabla F_k(\theta_{t-1,\{i\}}^k))\right|^2\right]} \\ &\leq \eta_t \sqrt{\mathbb{E}[|T_1|^2]} \frac{b_2 \sqrt{L}}{\sqrt{K}} \sqrt{\sum_{k=1}^K \mathbb{E}[|Y_{t-1} - \theta_{t-1}^k|^2 + |Y_{t-1,\{i\}} - \theta_{t-1,\{i\}}^k|^2]} \\ &\lesssim \eta_t^2 \sqrt{\mathbb{E}[|T_1|^2]} \\ &\leq \eta_t \left(\frac{1}{4c} \eta_t^2 + c \mathbb{E}[|T_1|^2] \right) \end{aligned} \quad (\text{E.14})$$

$$\leq \eta_t \frac{c}{4} \mathbb{E}[|T_1|^2] + O(\eta_t^3), \quad (\text{E.15})$$

1261 where (E.14) involves an application of Young's inequality $xy \leq \epsilon x^2 + (4\epsilon)^{-1}y^2$ with $\epsilon = (4c)^{-1}$,
 1262 where c is as in (E.11). Therefore, in view of $(1 + \eta_t \frac{c}{2})(1 - \eta_t c) \leq 1 - \eta_t \frac{c}{2}$, we combine (E.11)
 1263 -(E.15) into (E.10) to obtain

$$\mathbb{E}[|Y_t - Y_{t,\{i\}}|^2] \leq (1 - \eta_t \frac{c}{2} + \frac{\eta_t^2}{K})\mathbb{E}[|Y_{t-1} - Y_{t-1,\{i\}}|^2] + O(\eta_t^3 + \frac{\eta_t^4}{K}), \quad t > i,$$

1264 which immediately shows (E.1) with standard manipulations (see Lemma A.1 and A.2 of [Zhu et al. \[2023\]](#);
 1265 [Polyak and Juditsky \[1992\]](#)) \square

1266 *Proof of Proposition 5.* Recall C_t from (2.2). Let $r_{t,s}$ be the number of synchronization steps
 1267 between $s - 1$ and t , satisfying $\lfloor \frac{t-s}{\tau} \rfloor + 1 \geq r_{t,s} \geq \lfloor \frac{t-s}{\tau} \rfloor$. Further note that $\mathbf{C}^{r_{t,s}} = \prod_{j=s}^t C_j$.
 1268 From (2.2) and (B.14), it is easy to see that

$$(\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{t,\{i\}})(I - J) = - \sum_{s=i}^t \eta_s (G_s - G_{s,\{i\}})(\mathbf{C}^{r_{t,s}} - J), \quad (\text{E.16})$$

1269 where we have repeatedly used the fact that $\mathbf{C}\mathbf{1} = \mathbf{1}$. Moreover, it also holds that

$$\|\mathbf{C}^{r_{t,s}} - J\|_2 \leq \left(\rho^{\frac{1}{\tau}}\right)^{\max\{t-s-(\tau-2), 0\}} = 1_{\{t-s < \tau-1\}} + 1_{\{t-s \geq \tau-1\}} \tilde{\rho}^{t-s-(\tau-1)} := \kappa_{\rho,\tau}(t, s), \quad (\text{E.17})$$

1270 where $\tilde{\rho} = \rho^{1/\tau}$. Equation E.17 also appears as (S7) in [Gu and Chen \[2024\]](#). In view of (E.17), one
 1271 can expand (E.16) as follows:

$$\begin{aligned} & \mathbb{E}\left[\left\|\sum_{s=i}^t \eta_s (G_s - G_{s,\{i\}})(\mathbf{C}^{r_{t,s}} - J)\right\|_F^2\right] \\ & \leq \sum_{s=i}^t \kappa_{\rho,\tau}^2(t, s) \eta_s^2 \mathbb{E}[|G_s - G_{s,\{i\}}|_F^2] \\ & \quad + \sum_{s=i}^t \sum_{l=i, l \neq s}^t \kappa_{\rho,\tau}(t, s) \kappa_{\rho,\tau}(t, l) \eta_s \eta_l \mathbb{E}\left[\text{Tr}[(G_s - G_{s,\{i\}})^\top (G_l - G_{l,\{i\}})]\right] \end{aligned} \quad (\text{E.18})$$

$$\begin{aligned} & \leq \sum_{s=i}^t \kappa_{\rho,\tau}^2(t, s) \eta_s^2 \mathbb{E}[|G_s - G_{s,\{i\}}|_F^2] \\ & \quad + \sum_{s=i}^t \sum_{l=i, l \neq s}^t \kappa_{\rho,\tau}(t, s) \kappa_{\rho,\tau}(t, l) 2^{-1} \mathbb{E}\left(\eta_s^2 |G_s - G_{s,\{i\}}|_F^2 + \eta_l^2 |G_l - G_{l,\{i\}}|_F^2\right) \\ & \leq \sum_{s=i}^t \kappa_{\rho,\tau}^2(t, s) \eta_s^2 \mathbb{E}[|G_s - G_{s,\{i\}}|_F^2] + \sum_{s=i}^t \kappa_{\rho,\tau}(t, s) \eta_s^2 \mathbb{E}[|G_s - G_{s,\{i\}}|_F^2] \left(\sum_{l=i, l \neq s}^t \kappa_{\rho,\tau}(t, l)\right). \end{aligned} \quad (\text{E.19})$$

1272 Now we are required to tackle $\mathbb{E}[|G_s - G_{s,\{i\}}|_F^2]$. To that end, observe that for $s > i$

$$\begin{aligned} \mathbb{E}[|G_s - G_{s,\{i\}}|_F^2] &= K^2 \sum_{k=1}^K w_k^2 \mathbb{E}[|\nabla f_k(\theta_{s-1}^k, \xi_s^k) - \nabla f_k(\theta_{s-1,\{i\}}^k, \xi_s^k)|^2] \\ &\leq 2b_2^2 L \sum_{k=1}^K \mathbb{E}[|\theta_{s-1}^k - \theta_{s-1,\{i\}}^k|^2] \\ &\lesssim \sum_{k=1}^K \mathbb{E}[|\theta_{s-1}^k - Y_{s-1}|^2] + \sum_{k=1}^K \mathbb{E}[|\theta_{s-1,\{i\}}^k - Y_{s-1,\{i\}}|^2] + \sum_{k=1}^K \mathbb{E}[|Y_{s-1} - Y_{s-1,\{i\}}|^2] \\ &= O(\eta_s^2 K), \end{aligned} \quad (\text{E.20})$$

1273 where (E.20) follows from Lemma S16 of Gu and Chen [2024] and Proposition 4 respectively. Putting
 1274 (E.20) back into (E.19), we obtain

$$\mathbb{E}[|(\Theta_t - \Theta_{t,\{i\}})(I - J)|_F^2] \lesssim \sum_{s=i}^t \eta_s^4 K \kappa_{\rho,\tau}^2(t, s) \leq \sum_{s=i}^t \eta_s^4 \tilde{\rho}^{t-s} = O(\eta_t^4 K),$$

1275 where the last assertion uses $\int_1^n x^{-a} e^{yx} dx \lesssim n^{-a} e^{ny}$ for $a, y > 0$, where \lesssim is independent of n .
 1276 This completes the proof. \square

1277 *Proof of Proposition 6.* We can re-purpose significant portions of the proof of Lemma S16 of Gu
 1278 and Chen [2024] to prove (E.2). Indeed, writing $\Theta_t = \sum_{s=1}^t \eta_s G_s C_s$, we have from the referenced
 1279 proof that

$$\begin{aligned} \mathbb{E}[|\Theta_t(I - J)|_F^4] &= \mathbb{E}\left[\left|\sum_{s=1}^t \eta_s G_s (\mathbf{C}^{r_{t,s}} - J)\right|_F^4\right] \\ &\leq 2\mathbb{E}\left[\left(\sum_{s=1}^t \eta_s^2 \kappa_{\rho,\tau}^2(t, s) |G_s|^2\right)^2\right] + 2\mathbb{E}\left[\left(\sum_{s=1}^t \kappa_{\rho,\tau}(t, s) \sum_{l=1, l \neq s}^t \kappa_{\rho,\tau}(t, l) \eta_s \eta_l |G_s^\top G_l|\right)^2\right] \\ &:= S_1 + S_2. \end{aligned} \quad (\text{E.21})$$

1280 For S_1 in (E.21), it is straightforward to obtain

$$\begin{aligned} S_1 &\lesssim \sum_{s=1}^t \eta_s^4 \kappa_{\rho,\tau}^4(t, s) \mathbb{E}[|G_s|^4] + \sum_{s=1}^t \sum_{l=1, l \neq s}^t \eta_s^2 \eta_l^2 \kappa_{\rho,\tau}^2(t, s) \kappa_{\rho,\tau}^2(t, l) \mathbb{E}[|G_s|^2 |G_l|^2] \\ &\lesssim \sum_{s=1}^t \eta_s^4 \kappa_{\rho,\tau}^4(t, s) \mathbb{E}[|G_s|^4] + \sum_{s=1}^t \kappa_{\rho,\tau}^2(t, s) \eta_s^4 \mathbb{E}[|G_s|^4] \max_s \sum_{l=1, l \neq s}^t \kappa_{\rho,\tau}^2(t, l) \end{aligned} \quad (\text{E.22})$$

$$\lesssim \sum_{s=1}^t K^2 \eta_s^4 (\kappa_{\rho,\tau}^4(t, s) + \kappa_{\rho,\tau}^2(t, s)) = O(\eta_t^4 K^2), \quad (\text{E.23})$$

1281 where, in (E.22) we apply AM-GM inequality to derive

$$\eta_s^2 \eta_l^2 \mathbb{E}[|G_s|^2 |G_l|^2] \leq \frac{\eta_s^4 \mathbb{E}[|G_s|^4] + \eta_l^4 \mathbb{E}[|G_l|^4]}{2}.$$

1282 A very similar treatment yields the same bound on S_2 , completing the proof of (E.2). \square

1283 *Proof of Proposition 7.* Write

$$\begin{aligned} R_t &:= Y_t - \theta_K^* = E_1 + E_2 + E_3, \text{ where,} \\ E_1 &= R_{t-1} - \eta_t \nabla F(Y_{t-1}), \\ E_2 &= \eta_t \sum_{k=1}^K w_k (\nabla F_k(Y_{t-1}) - \nabla F_k(\theta_{t-1}^k)), \text{ and} \\ E_3 &= \eta_t \sum_{k=1}^K w_k g_k(\theta_{t-1}^k, \xi_t^k). \end{aligned} \quad (\text{E.24})$$

1284 Note that trivially, Assumptions A.1 and A.2 imply that

$$\mathbb{E}[|E_1|^4] \leq (1 - \eta_t c) \mathbb{E}[|R_{t-1}|^4]. \quad (\text{E.25})$$

1285 Moving on, for E_2 we proceed as follows:

$$\begin{aligned} \mathbb{E}[|E_2|^4] &= \eta_t^4 \mathbb{E}\left[\left|\sum_{k=1}^K w_k (\nabla F_k(Y_{t-1}) - \nabla F_k(\theta_{t-1}^k))\right|^4\right] \\ &\leq C_2 \frac{\eta_t^4}{K^2} \mathbb{E}\left[\left(\sum_{k=1}^K |Y_{t-1} - \theta_{t-1}^k|^2\right)^2\right] \\ &\leq C_2 \frac{\eta_t^4}{K^2} \mathbb{E}[|\Theta_{t-1}(I - J)|_F^4] \leq C_2' \eta_t^8, \end{aligned} \quad (\text{E.26})$$

1286 for some constants $C_2, C'_2 > 0$, where the final assertion is drawn from Proposition 6. Finally, for
 1287 E_3 , we obtain,

$$\mathbb{E}[|E_3|^4] = \eta_t^4 \mathbb{E}\left[\left|\sum_{k=1}^K w_k g_k(\theta_{t-1}^k, \xi_t^k)\right|^4\right] \leq C_3 \frac{\eta_t^4}{K^2}, \quad (\text{E.27})$$

1288 for some constant $C_3 > 0$, where we have used the fact that $g_k(\theta_{t-1}^k, \xi_t^k)$ are mean-zero and
 1289 independent random vectors conditional on \mathcal{F}_t . Now, we will leverage (E.25)-(E.27) to develop
 1290 bounds on the cross-product terms. In particular,

$$\begin{aligned} \mathbb{E}[|E_1|^2 |E_2|^2] &\leq \sqrt{\mathbb{E}[|E_1|^4]} \sqrt{\mathbb{E}[|E_l|^4]} \\ &\leq C_3 \eta_t^4 \sqrt{\mathbb{E}[|E_1|^4]} \\ &\leq C_3 \min\{\eta_t \varepsilon \mathbb{E}[|E_1|^4] + (4\varepsilon)^{-1} \eta_t^7, \eta_t^2 \varepsilon \mathbb{E}[|E_1|^4] + (4\varepsilon)^{-1} \eta_t^6\}, \end{aligned} \quad (\text{E.28})$$

1291 and similarly

$$\mathbb{E}[|E_1|^2 |E_l|^2] \leq C_3 \min\{\eta_t \varepsilon \mathbb{E}[|E_1|^4] + (4\varepsilon)^{-1} \frac{\eta_t^3}{K^2}, \eta_t^2 \varepsilon \mathbb{E}[|E_1|^4] + (4\varepsilon)^{-1} \frac{\eta_t^2}{K^2}\}, \quad (\text{E.29})$$

1292 where the final assertions follow from Young's inequality. Here, ε is chosen to be small enough,
 1293 however it remains a constant; the explicit choice of ε will be indicated towards the end of the proof,
 1294 when we collect terms to establish the recursion. Note that, quite trivially, from (E.26) and (E.27),
 1295 one has

$$\mathbb{E}[|E_2|^2 |E_3|^2] \leq C_4 \frac{\eta_t^6}{K} \text{ for some constant } C_4 > 0. \quad (\text{E.30})$$

1296 Rest of the cross-products are strictly dominated by some combinations of the terms analyzed till
 1297 now. For example, for $l, r, q \in \{1, 2, 3\}$, Cauchy-Schwarz and AM-GM inequalities implies that

$$\begin{aligned} \mathbb{E}[(E_l^\top E_q)^2] &\leq \mathbb{E}[|E_l|^2 |E_q|^2], \\ \mathbb{E}[|E_l|^2 (E_r^\top E_q)] &\leq \sqrt{\mathbb{E}[|E_l|^4]} \sqrt{\mathbb{E}[(E_r^\top E_q)^2]}, \\ \mathbb{E}[(E_l^\top E_r)(E_l^\top E_q)] &\leq 2^{-1} (\mathbb{E}[(E_l^\top E_r)^2] + \mathbb{E}[(E_l^\top E_q)^2]). \end{aligned} \quad (\text{E.31})$$

1298 A careful collection of terms from (E.25)-(E.31) yields

$$\mathbb{E}[|R_t|^4] \leq (1 - \eta_t c)(1 + \eta_t C_0 \varepsilon) \mathbb{E}[|R_{t-1}|^4] + O\left(\frac{\eta_t^3}{K^2} + \eta_t^5\right), \quad (\text{E.32})$$

1299 where C_0 is a large constant depending upon C_2, C_3 and C_4 . Now, choose $\varepsilon > 0$ so that $C_0 \varepsilon < c/2$,
 1300 upon which we immediately obtain $(1 - \eta_t c)(1 + \eta_t C_0 \varepsilon) < 1 - \eta_t c/2$. Therefore, (E.32) immediately
 1301 yields (E.3). \square

1302 *Proof of Proposition 8.* Recall (E.8). Clearly, for $t = i$, the result is trivial. For $t > i$, we leverage
 1303 (E.9). A proof very similar to (E.3), which uses a similar decomposition (E.24), can then be followed.
 1304 The crucial term is $\mathbb{E}[|T_1|^2 |T_3|^2]$, which is computed below. Note that

$$\begin{aligned} \mathbb{E}[|T_3|^4] &\leq \frac{\eta_t^4}{K^2} \mathbb{E}\left[\sum_{k=1}^K |g_k(\theta_{t-1}^k, \xi_t^k) - g_k(\theta_{t-1, \{i\}}^k, \xi_t^k)|^4\right] \\ &\leq L' \frac{\eta_t^4}{K} \mathbb{E}\left[\sum_{k=1}^K |\theta_{t-1}^k - \theta_{t-1, \{i\}}^k|^4\right] \\ &\leq 27L' \frac{\eta_t^4}{K} \mathbb{E}\left[\sum_{k=1}^K |\theta_{t-1}^k - Y_{t-1}|^4 + K |Y_{t-1} - Y_{t-1, \{i\}}|^4 + \sum_{k=1}^K |\theta_{t-1, \{i\}}^k - Y_{t-1, \{i\}}|^4\right] \\ &\lesssim \frac{\eta_t^4}{K} \mathbb{E}[|\Theta_{t-1}(I - J)|_F^4 + |\Theta_{t-1, \{i\}}(I - J)|_F^4 + K |Y_{t-1} - Y_{t-1, \{i\}}|^4] \\ &\lesssim \eta_t^4 \mathbb{E}[|Y_{t-1} - Y_{t-1, \{i\}}|^4] + O(\eta_t^8). \end{aligned} \quad (\text{E.33})$$

Therefore, from (E.33), it follows

$$\begin{aligned}
\mathbb{E}[|T_1|^2|T_3|^2] &\leq \sqrt{\mathbb{E}[|T_1|^4]}\sqrt{\mathbb{E}[|T_3|^4]} \\
&\leq \sqrt{\mathbb{E}[|T_1|^4]}(\sqrt{L}\eta_t^2\sqrt{\mathbb{E}[|Y_{t-1} - Y_{t-1,\{i\}}|^4]} + O(\eta_t^4)) \\
&\lesssim \eta_t^2\mathbb{E}[|Y_{t-1} - Y_{t-1,\{i\}}|^4] + \eta_t^4\sqrt{\mathbb{E}[|T_1|^4]} \\
&\lesssim \eta_t^2\mathbb{E}[|Y_{t-1} - Y_{t-1,\{i\}}|^4] + \eta_t^3\mathbb{E}[|T_1|^4] + O(\eta_t^5).
\end{aligned} \tag{E.34}$$

Rest of the terms are computed similar to Proposition 7, and the details are omitted. The final recursion can be derived to be

$$\mathbb{E}[|Y_t - Y_{t,\{i\}}|^4] \leq (1 - \eta_t c)\mathbb{E}[|Y_{t-1} - Y_{t-1,\{i\}}|^4] + O(\eta_t^5) \text{ for some small constant } c > 0,$$

which immediately yields (E.4). \square

F Additional Simulations

F.1 Effect of n and K on the Berry-Esseen rate

In this subsection, we empirically investigate the behavior of the Berry-Esseen error $d_C(\sqrt{n}(\bar{Y}_n - \theta_K^*), Z)$ for $Z \sim N(0, \Sigma_n)$ with varying choices of the number of iterations N and the number of clients K . If the bound (2.5) is sharp, we expect the Berry-Esseen error to decay with increasing N , and increase with an increasing number of clients. Since the distance d_C involves taking a supremum over all convex sets, which is computationally infeasible, we restrict ourselves to the following measure of the approximation error:

$$\tilde{d}_c = \sup_{x \in [0, c]} |\mathbb{P}(|\sqrt{n}\Sigma_n^{-1/2}(\bar{Y}_n - \theta_K^*)| \leq x) - \mathbb{P}(|Z| \leq x)|, \quad Z \sim N(0, I).$$

For large enough $c > 0$, we expect \tilde{d}_c to be a reasonable proxy for d_C . For our numerical exercises to quantify \tilde{d}_c , we analyze the output \bar{Y}_n of the `local SGD` algorithm under a federated random effects model, hereafter denoted as `FRand-eff`. We describe the set-up below.

F.1.1 FRand-eff formulation

Consider a positive definite matrix $\Gamma \in \mathbb{R}^{d \times d}$ and $\beta_0 \in \mathbb{R}^d$, and let $\mathcal{D}^K := \{\beta_1, \dots, \beta_K\}$ i.i.d. $N_d(\beta_0, \Gamma)$. Moreover, consider $\Sigma^K := \{\sigma_1^2, \dots, \sigma_K^2\} \subset \mathbb{R}_{>0}^K$. For $k \in [K]$ and at $t \in [n]$ -th iteration, suppose that the k -th client has access to data $(y_{tk}, x_{tk}) \in \mathbb{R} \times \mathbb{R}^d$ generated from the linear model $y_{tk} \sim N(x_{tk}^\top \beta_k, \sigma_k^2)$. If the weights are chosen such that $w_1 = \dots = w_K = K^{-1}$, then clearly $\theta_K^* = \sum_{k=1}^K w_k \beta_k \rightarrow \beta_0$ as $K \rightarrow \infty$. Therefore, `local SGD` can be employed, and we expect \bar{Y}_n to consistently estimate β_0 as n and K grow. This model highlights the need for information-sharing across client, since unless $\Gamma = 0$, the output of local vanilla SGD for any particular client is inconsistent for β_0 .

For the purpose of the numerical exercises in this section, we choose $d = 2$ and $\beta_0 = (2, -3)^\top$, and let $\Gamma = \gamma I$ with $\gamma \geq 0$. In particular, $\gamma = 0$ corresponds to a fixed effect β_0 from which each client generates their observations. For each K , we generate Σ^K uniformly from the set $\{1, \dots, 5\}$, \mathcal{D}^K from the specification above, and keep them fixed throughout the corresponding experiments as n varies. The underlying connection matrix C is taken as $C_{ij} = \frac{1}{3}I\{|i - j| \leq 1\}$, $i, j \in [K]$. In other words, every client is connected to only its two immediate neighbors.

F.1.2 \tilde{d}_c versus n and K

In this set, we analyze the behavior of \tilde{d}_c versus n for different choices of K , τ , and γ . In particular, we aim to verify the Berry-Esseen error rate of $n^{1/2-\beta}\sqrt{K}$ from Theorems 2.1 and 2.2. Let $\gamma = 1$. Consider the following two separate settings corresponding to n , K , and τ .

- **Setting 1.** Let $K = 10$, and vary $n \in \{100, 200, 300, 400, 500\}$, and $\tau \in \{10, 15, 20\}$. For each pair of (n, τ) , we plot \tilde{d}_c against n .

1341 • **Setting 2.** fix $n = 300$, and vary $K \in \{20, 40, 60, 80, 100\}$, and $\tau \in \{10, 15, 20\}$. For each
1342 pair of (K, τ) , we plot \tilde{d}_c against K .

1343 We provide the practical details behind empirically estimating \tilde{d}_c . For each of the experimental
1344 settings described above, we generate $(y_{tk}, x_{tk}) \in \mathbb{R} \times \mathbb{R}^d, k \in [K], t \in [n]$ from the `FRand-eff`
1345 specification described above, and run the `local SGD` algorithm with step size $\eta_t = 0.3t^{-0.75}$
1346 for $t \in [n]$. For the large choice of $c = 100$, \tilde{d}_c is empirically estimated by $n_{\text{sim}} = 1000$ many
1347 independent Monte-Carlo repetitions of our experiments.

1348 Figure 1 allows us to draw important practical insights from the rates of Theorems 2.1 and 2.3. Firstly,
1349 from the Settings 1 and 3, the synchronization parameter τ does not seem to have a significant effect
1350 on the behavior of \tilde{d}_c . Moreover, Figure 1(left) seems to corroborate well with the conclusion of
1351 Theorem 2.1, with \tilde{d}_c decaying with n for a fixed K . On the other hand, for Setting 2, Figure 1(right)
1352 seems to point towards a trade-off in terms of K for fixed n . This particular behavior becomes clearer
1353 as we recall (2.5). For fixed $n = 300$, the initial decay of \tilde{d}_c (and by extension, d_C) with increasing
1354 K , is caused by the $n^{-\beta/2}K^{-1/2}$ term. However, as K increases, the term $n^{1-\beta/2}\sqrt{K}$ starts to
1355 dominate, leading the error \tilde{d}_c to increase with increasing K . This numerical exercise establish the
1356 sharpness of our upper-bound (2.5), complementing the discussions in Remark 2.1.

1357 To investigate the behavior of \tilde{d}_c further, we also consider the case $\gamma = 5$. In Figure 4, due to the
1358 increased heterogeneity across β_k , the effect of synchronization becomes more pronounced; for the
1359 same values of n, K, τ , the \tilde{d}_c values are much lesser compared to that in Figure 1. In particular, in
1360 Figure 4(right), the inflection point in K beyond which $n^{1/2-\beta}\sqrt{K}$ starts to dominate, has shifted to
1361 the right. This is understandable, since increased variability among β_k means a greater reward for
1362 sharing information, and thus the effect of increasing the clients leads to lowering the error \tilde{d}_c for a
1363 longer regime, before the asymptotics of $n^{1/2-\beta}\sqrt{K}$ eventually kicks in.

1364 F.2 Computation-communication trade-off

1365 In this section, we numerically investigate the computation-communication trade-off hinted at in
1366 Remark 2.2. There, we noted that if $K \asymp n^c$ for $c > 1/2$, then, based on our upper bounds, we
1367 argued that for no $\beta \in (1/2, 1)$ does d_C converge to 0. In particular, this observation is trivial
1368 for $\beta \in (1/2, 1/2 + c/2]$ since the central limit theory itself fail to help in view of violation of
1369 $K \gtrsim n^{2\beta-1}$. Of particular interest is the range $\beta \in (1/2 + c/2, 1)$, where, as per Theorem 3 of Gu
1370 and Chen [2024], central limit theory continues to hold, but (2.8) suggests that the upper bound to d_C
1371 is no longer $o(1)$.

1372 To explore this phenomena through numerical examples, we invoke `FRand-eff` for $\gamma = 0$, and let
1373 $n \in \{100, 200, 300, 400, 500\}$, and $K = \lfloor n^r \rfloor$ for $r \in \{0.2, 0.6\}$. In conjunction with Theorem 2.3,
1374 we consider the following error:

$$d_c^\dagger = \sup_{x \in [0, c]} |\mathbb{P}(|\sqrt{n}\Sigma^{-1/2}(\bar{Y}_n - \theta_K^*)| \leq x) - \mathbb{P}(|Z| \leq x)|,$$

1375 where $\Sigma = A^{-1}SA^{-\top}$. Moreover, we consider the `local SGD` algorithm with $\tau = 5$, and $\eta_t =$
1376 $0.5t^{-\beta}$. In light of $1/2 + r/2 \in \{0.6, 0.8\}$, we ensure the validity of central limit theory by letting
1377 $\beta \in \{0.85, 0.9, 0.95\}$. Finally, for each value of r , we plot \tilde{d}_c against n for the different choices of β .
1378 For $r = 0.2$ and $r = 0.6$, the evident decreasing and increasing trends of \tilde{d}_c in Figure 2 respectively,
1379 vindicate not only the sharpness of our Berry-Esseen bounds Theorems 2.1-2.3, but also clearly
1380 highlights trade-off at the region $\sqrt{n} \ll K \ll n$.

1381 F.3 Performance of the time-uniform Gaussian approximations

1382 This section devotes itself to numerical studies to validate the efficacy of the Gaussian approximations
1383 `Aggr-GA` and `Client-GA`, discussed in Section 3. Consider the quantities

$$U_n = \max_{1 \leq t \leq n} \left| \sum_{s=1}^t (Y_s - \theta_K^*) \right|, U_n^{\text{Aggr-GA}} = \max_{1 \leq t \leq n} \left| \sum_{s=1}^t Y_{s,1}^G \right|, \text{ and } U_n^{\text{Client-GA}} = \max_{1 \leq t \leq n} \left| \sum_{s=1}^t Y_{s,2}^G \right|,$$

1384 where $Y_{s,1}^G$ and $Y_{s,2}^G$ are defined as in Theorem 3.1. Moreover, we also consider the Brownian motion
 1385 approximation by functional central limit theorem as another competitor, and as such, consider

$$U_n^{\text{f-CLT}} = \max_{1 \leq t \leq n} \left| \sum_{s=1}^t Z_s \right|, \quad Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} N(0, \Sigma),$$

1386 where Σ is as in Section 2.2.1. In order to compare the distributions of $U_n^{\text{Aggr-GA}}$, $U_n^{\text{Client-GA}}$ and
 1387 $U_n^{\text{f-CLT}}$ to that of U_n , we resort to Q-Q plots. Fix $N = 500$, $\tau = 20$, and let $K \in \{10, 25, 50\}$.
 1388 For each triplet of (N, K, γ) , we simulate $n_{\text{sim}} = 500$ parallel independent local SGD chains
 1389 with step-sizes $\eta_t = 0.7t^{-0.85}$, and observations from the FRand-eff model in order to empirically
 1390 simulate U_n . Concurrently, we also simulate n_{sim} independent observations from the distributions of
 1391 $U_n^{\text{Aggr-GA}}$, $U_n^{\text{Client-GA}}$ and $U_n^{\text{f-CLT}}$ by running the corresponding chains in parallel. The QQ-plots are
 1392 shown in Figure 3.

1393 The sub-optimality of the functional CLT as a time-uniform Gaussian approximation to $\{Y_t\}$ is
 1394 empirically evident from the QQ-plots. Both our proposals Aggr-GA and Client-GA uniformly
 1395 dominate the approximation via Brownian motion across different settings. Moreover, as K increases
 1396 from left panel to the right, Aggr-GA out-performs Client-GA. This in line with Proposition 1 (i)
 1397 and (ii) underpinning the sharper approximation rate for Aggr-GA. However, we must recall that
 1398 Client-GA requires local covariance estimation for each client, thus protecting the privacy of the
 1399 federated setting.

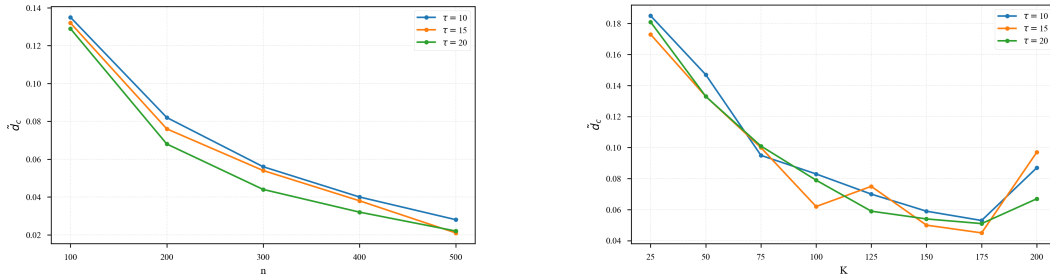


Figure 4: Plot of \tilde{d}_c against n and K for $\gamma = 5$, and Settings 1(left), and 2(right).