Appendix

Anonymous Authors

Abstract

In this paper, we propose FGPR: a Federated Gaussian process ($\mathcal{GP}$) regression framework that uses an averaging strategy for model aggregation and stochastic gradient descent for local client computations. Notably, the resulting global model excels in personalization as FGPR jointly learns a global $\mathcal{GP}$ prior across all clients. The predictive posterior then is obtained by exploiting this prior and conditioning on local data which encodes personalized features from a specific client. Theoretically, we show that FGPR converges to a critical point of the full log-likelihood function, subject to statistical error. Through extensive case studies we show that FGPR excels in a wide range of applications and is a promising approach for privacy-preserving multi-fidelity data modeling.

1 Multi-fidelity Modeling

Example 3: CURRIN The CURRIN (Currin et al., 1991; Xiong et al., 2013) is a two-dimensional problem that is widely used for multi-fidelity computer simulation models. Given the input domain $\boldsymbol{x} \in [0,1]^2$, the high-fidelity model is

$$y_h(\boldsymbol{x}) = \left[1 - \exp\left(-\frac{1}{2x_2}\right)\right] \frac{2300x_1^3 + 1900x_1^2 + 2092x_1 + 60}{100x_1^3 + 500x_1^2 + 4x_1 + 20}$$

whereas the low-fidelity model is given by

$$\begin{aligned} y_l(\boldsymbol{x}) = & \frac{1}{4}[y_h(x_1 + 0.05, x_2 + 0.05) + y_h(x_1 + 0.05, \max(0, x_2 - 0.05))] \\ & + \frac{1}{4}[y_h(x_1 - 0.05, x_2 + 0.05) + y_h(x_1 - 0.05, \max(0, x_2 - 0.05))]. \end{aligned}$$

We collect 40 data points from the HF model and 200 data points from the LF model. The number of testing data points is 1,000.

Example 4: PARK The PARK function (Cox et al., 2001; Xiong et al., 2013) is a four-dimensional problem ($\boldsymbol{x} \in (0,1]^4$) where the high-fidelity model is given as

$$y_h(\boldsymbol{x}) = \frac{x_1}{2}\left[\sqrt{1 + (x_2 + x_3^2)\frac{x_4}{x_1^2}} - 1\right] + (x_1 + 3x_4)\exp[1 + \sin(x_3)],$$

while the low-fidelity model is

$$y_l(\boldsymbol{x}) = \left[1 + \frac{\sin(x_1)}{10}\right] y_h(\boldsymbol{x}) - 2x_1 + x_2^2 + x_3^2 + 0.5.$$

Example 5: BRANIN In this example, there are three fidelity levels (Perdikaris et al., 2017; Cutajar et al., 2019):

$$\begin{aligned}
y_h &= \left(\frac{-1.275x_1^2}{\pi^2} + \frac{5x_1}{\pi} + x_2 - 6\right)^2 + \left(10 - \frac{5}{4\pi}\right)\cos(x_1) + 10,\\
y_m &= 10\sqrt{y_h(\boldsymbol{x} - 2)} + 2(x_1 - 0.5) - 3(3x_2 - 1) - 1,\\
y_l &= y_m(1.2(\boldsymbol{x} + 2)) - 3x_2 + 1,\\
x &\in [-5, 10] \times [0, 15]
\end{aligned}$$

where $y_m(\cdot)$ represents the output from the medium-fidelity (MF) model.

Example 6: Hartmann-3D Similar to Example 5, this is a 3-level multi-fidelity dataset where the input space is $[0,1]^3$. The evaluation of observations with fidelity t is defined as (Cutajar et al., 2019)

$$y_t(\boldsymbol{x}) = \sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{3} A_{ij}(x_j - P_{ij})^2\right)$$

where

$$A = \begin{bmatrix} 3 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3 & 10 & 30 \\ 0.1 & 10 & 35 \end{bmatrix}, P = \begin{bmatrix} 0.3689 & 0.1170 & 0.2673 \\ 0.4699 & 0.4387 & 0.7470 \\ 0.1091 & 0.8732 & 0.5547 \\ 0.0381 & 0.5743 & 0.8828 \end{bmatrix},$$

$$\boldsymbol{\alpha} = (1.0, 1.2, 3.0, 3.2)^\intercal, \boldsymbol{\alpha}_t = \boldsymbol{\alpha} + (3 - t)\boldsymbol{\delta}, \boldsymbol{\delta} = (0.01, -0.01, -0.1, 0.1)^\intercal.$$

Example 7: Borehole Model The Borehole model is an 8-dimensional physical model that simulates water flow through a borehole (Moon et al., 2012; Gramacy and Lian, 2012; Xiong et al., 2013). The high-fidelity model is given as

$$y_h(\boldsymbol{x}) = \frac{2\pi x_3(x_4 - x_6)}{\ln(x_2/x_1)[1 + 2x_7x_3/(\ln(x_2/x_1)x_1^2x_8) + x_3/x_5]}$$

where $x_1 \in [0.05, 0.15], x_2 \in [100, 50000], x_3 \in [63070, 115600], x_4 \in [990, 1110], x_5 \in [63.1, 115], x_6 \in [700, 820], x_7 \in [1120, 1680], x_8 \in [9855, 12045]$. The low-fidelity model is

$$y_l(\boldsymbol{x}) = \frac{5\pi x_3(x_4 - x_6)}{\ln(x_2/x_1)[1.5 + 2x_7x_3/(\ln(x_2/x_1)x_1^2x_8) + x_3/x_5]}.$$

2 Additional Application: NASA Aircraft Gas Turbine Engines

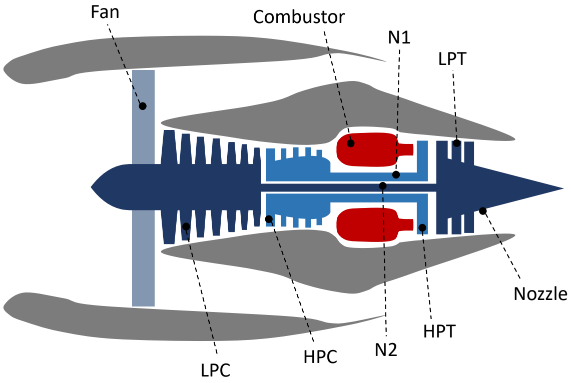

Figure 1: The engine diagram in C-MAPSS.

In this case study, we consider degradation signals generated from aircraft gas turbine engines using the NASA Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) tools (NASA dataset Link). The dataset consists of 100 engines and contains time-series degradation signals collected from multiple sensors installed on the engines. Figure 1 illustrates the engine diagram in C-MAPSS. The goal of the experiment is to predict the degradation signals for test engines in a federated paradigm. To do so, we assume that each client/device is a single engine and all engines are aiming to collaboratively learn a predictive degradation model.

We briefly describe our training procedures. We randomly divide the 100 engines into 60 training engines and 40 testing engines. For each testing unit k, we randomly split the data on each device into a 50% training dataset $D_{k,\text{train}} := (\boldsymbol{X}_{k,\text{train}}, \boldsymbol{y}_{k,\text{train}})$ and a 50% testing dataset $D_{k,\text{test}} := (\boldsymbol{X}_k^*, \boldsymbol{y}_k^*)$, where

Table 1: Averaged RMSE (line 1 in each cell) and standard deviation (std) of RMSE (line 2 in each cell) across all testing devices for the NASA data. Each experiment is repeated 30 times.

Averaged RMSE ×10 std of RMSE ×10	`FGPR`	`Polynomial`	`Neural`
Sensor 2	$\mathbf{5.45}$ $\mathbf{(0.01)}$	6.79 (0.01)	6.47 (0.05)
	$\mathbf{0.87}$ $\mathbf{(0.02)}$	0.98 (0.02)	1.02 (0.01)
Sensor 7	$\mathbf{5.76}$ $\mathbf{(0.03)}$	6.55 (0.02)	6.71 (0.02)
	$\mathbf{0.76}$ $\mathbf{(0.01)}$	0.89 (0.04)	0.85 (0.03)

$\boldsymbol{y}_k^* = \left[y_{k,1}^*, ...y_{k,|D_{k,\text{test}}|}^*\right]^\intercal$, $\boldsymbol{X}_k^* = \left[{x_{k,1}^*}^\intercal, ..., {x_{k,|D_{k,\text{test}}|}^*}^\intercal\right]$. Recall that in the main paper, we define $|D_{k,\text{test}}|$ as the number of data points in the set $D_{k,\text{test}}$. We first train `FGPR` using the 60 training units and obtain a final aggregated global model parameter $\boldsymbol{\theta}$. The testing unit k then directly uses this global parameter $\boldsymbol{\theta}$ and $D_{k,\text{train}}$ to predict outputs $[f({x_{k,1}^*}^\intercal), \cdots, f({x_{k,|D_{k,\text{test}}|}^*}^\intercal)]$ at testing locations $\boldsymbol{X}_k^*$ without any additional training.

We benchmark `FGPR` with the following models.

1. `Polynomial`: All signal trajectories exhibit polynomial patterns and therefore a polynomial regression is often employed to analyze this dataset (Liu et al., 2013; Yan et al., 2016; Song and Liu, 2018). More specifically, we train a polynomial regression using `FedAvg`. During the training process, each device updates the coefficients of a polynomial regression in the form of $y_k(x) = \sum_{i=0}^{p} \beta_{ik} x^i + \epsilon_k(x)$, where $\{\beta_{ik}\}_{i=0}^{p}$ are model parameters. This update is done by running gradient descent to minimize the local sum squared error. The central server aggregates the parameters using `FedAvg` and broadcasts the aggregated parameter to all devices in the following communication round. Here, we conduct experiments with different $p \in \{1, \ldots, 20\}$ and select the best p with the smallest averaged testing RMSE (will be defined shortly). Our empirical study finds that $p = 10$ provides the best performance.

2. `Neural`: we train a q-layer neural network using `FedAvg` (McMahan et al., 2017). Similar to `Polynomial`, we test the performance of the neural network with different $q \in \{1, \ldots, 20\}$. The best value is 2.

The performance of each model is measured by the averaged RMSE across all 40 testing devices defined as follows:

$$\text{RMSE} = \frac{1}{40} \sum_{k=1}^{40} \sqrt{\frac{\sum_{i=1}^{|D_{k,\text{test}}|} (f({x_{k,i}^*}^\intercal) - y_{k,i})^2}{|D_{k,\text{test}}|}}.$$

The averaged RMSE and the standard deviation of RMSE across all testing devices are reported in Table 1. Each experiment is repeated 30 times. The outputs on each device are scaled to be a mean 0 and variance 1 sequence.

From Table 1, we can obtain some important insights. First, `FGPR` consistently yields lower averaged RMSE than other benchmark models. This illustrates the good transferability of `FGPR`. More concretely, a shared global model can provide accurate surrogates even on untrained devices. This feature is in fact very helpful in transfer learning or online learning. For instance, the shared global model can be used as an initial parameter for fine-tuning on streaming data. Second, `FGPR` also provides smaller standard deviation of RMSE across all devices. This credits to the automatic personalization feature encoded in $\mathcal{GP}$. In the next section, we will compare our model with a state-of-the-art personalized FL framework to further demonstrate the advantage of `FGPR`.

3 Important Lemmas

In this section, we present some key lemmas used in our theoretical analysis. We defer the proofs of those Lemmas into Section 5.

Lemma 1. *(Theorem 4 in Braun (2006)) Let Ker be a Mercer kernel on a probability space $\mathcal{X}$ with probability measure μ, satisfying Ker$(x, x) \leq 1$ for all $x \in \mathcal{X}$, with eigenvalues $\{\lambda_i^*\}_{i=1}^{\infty}$. Let $\boldsymbol{K}_{f,N}$ be the empirical kernel matrix evaluated on data $\boldsymbol{X}$ $i.i.d.$ sampled from μ, then with probability at least $1-\delta$, the eigenvalues of $\lambda_j(\boldsymbol{K}_{f,N})$ satisfies the following bound for $1 \leq j \leq N$ and $1 \leq r \leq N$:*

$$\left|\frac{\lambda_j(\boldsymbol{K}_{f,N})}{N} - \lambda_j^*\right| \leq \lambda_j^* C(r,N) + H(r,N),$$

where

$$C(r,N) < r\sqrt{\frac{2}{N\lambda_r^*}\log\frac{2r(r+1)}{\delta}} + \frac{4r}{3N\lambda_r^*}\log\frac{2r(r+1)}{\delta},$$

$$H(r,N) < \lambda_r^* + \sum_{i=r+1}^{\infty}\lambda_i^* + \sqrt{\frac{2\sum_{i=r+1}^{\infty}\lambda_i^*}{N}\log\frac{2}{\delta}} + \frac{2}{3N}\log\frac{2}{\delta}.$$

Alternatively, $C(r,N)$ and $H(r,N)$ can also be bounded as follows:

$$
\begin{aligned}
C(r,N) &< r\sqrt{\frac{r(r+1)}{N\delta\lambda_r^*}}, \\
H(R,N) &< \lambda_r^* + \sum_{i=r+1}^{\infty} \lambda_i^* + \sqrt{\frac{2\sum_{i=r+1}^{\infty}\lambda_i^*}{N\delta}}.
\end{aligned}
$$

This Lemma is proved in Braun (2006).

Lemma 2. *(Chen et al., 2020) Under Assumptions 1-3a, in device k, for any $0 < \epsilon_k, \alpha_k < 1$, $C_{1k}(\alpha_k, b_k) > 0$ and $N_k > C_{2k}(\epsilon_k, b_k)$, then with probability at least $1 - \frac{2}{N_k^{\alpha_k}}$, we have*

$$
\begin{aligned}
\frac{\epsilon_k \log N_k}{8b_k\theta_{max}^2} \leq \sum_{j=1}^{N_k} \frac{\lambda_{1j}^2}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} &\leq \frac{4+2\alpha_k}{b_k\theta_{min}^2}\log N_k \\
\frac{N_k - C_{1k}(\alpha_k, b_k)\log N_k}{4\theta_{max}^2} \leq \sum_{j=1}^{N_k} \frac{\lambda_{2j}^2}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} &\leq \frac{N_k}{\theta_{min}^2} \\
\sum_{j=1}^{N_k} \frac{\lambda_{1j}\lambda_{2j}}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} &\leq \frac{5+2\alpha_k}{7b_k\theta_{min}^2}\log N_k.
\end{aligned}
$$

This Lemma is proved in Chen et al. (2020). Here note that we omit the subscript k in the eigenvalues λ for simplicity. The full notation should be, for example, λ_{1jk}^2 for device k.

Lemma 3. *Under Assumption 1-2 and 3b, for any $0 < \alpha_k < \frac{8b_k^2 - 12b_k - 6}{4b_k + 3}$, with probability at least $1 - \frac{1}{N_k^{1+\alpha_k}}$, the following inequalities hold:*

$$
\begin{aligned}
\sum_{j=1}^{N_k} \frac{\lambda_{1j}^2}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} &\leq N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}} \left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}^2(4b_k+3)}{\theta_{min}^2(8b_k^2 - 8b_k - 3)}\right), \\
\sum_{j=1}^{N_k} \frac{\lambda_{1j}\lambda_{2j}}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} &\leq N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}} \left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2 - 6b_k - 3)}\right), \\
\frac{N_k - C_{mat,k}N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}}{4\theta_{max}} \leq \sum_{j=1}^{N_k} \frac{\lambda_{2j}^2}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} &\leq \frac{N_k}{\theta_{min}^2}.
\end{aligned}
$$

Lemma 3 provides several bounds to constrain the eigenvalues of Matérn kernel.

Lemma 4. *Under Assumption 1-3a, with probability at least $1 - 2TM^{-c}$, the following inequality holds for* *any $k \in [K]$ and $0 \leq t < T$:*

$$\langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k^*(\boldsymbol{\theta}_k^{(t)}) \rangle \geq \frac{\gamma_k}{2} \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}_k^* \right\|_2^2 - C_{3k}(\alpha_k, b_k) \frac{\log M_k}{M_k},$$

where $\gamma_k = \min \left\{ \frac{1}{32\tau b_k \theta_{max}^2}, \frac{1}{4\theta_{max}^2} - \frac{8\theta_{max}^2}{\tau b_k \theta_{min}^4} \right\}$ and $C_{3k}(\alpha_k, b_k) = \frac{1}{64 b_k} + \frac{C_{1k}(\alpha_k, b_k)}{8} - \frac{4\theta_{max}^2}{b\theta_{min}^2}$.

Lemma 5. *Under Assumption 1-2 and 3b, with probability at least $1 - \frac{1}{M_k^{1+\alpha_k}}$, the following inequality holds:*

$$\left[g_k^*(\boldsymbol{\theta}_k^{(t)})\right]_2 (\theta_{2k}^{(t)} - \theta_{2k}^*) \geq \frac{\gamma_k}{2} (\theta_{2k}^{(t)} - \theta_{2k}^*)^2 - (\theta_{max} - \theta_{min})^2 M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)} - 1} \left(\frac{1}{2\theta_{min}^2} + \frac{C_{mat,k}(4b_k + 3)}{2\theta_{min}^2 (4b_k^2 - 6b_k - 3)} \right),$$

where $\gamma_k := \frac{1}{2M_k} \frac{M_k - C_{mat,k} M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}}{4\theta_{max}}$.

Lemma 6. *(Chen et al., 2020) Under Assumptions 1-2, for any $\phi > 0$, we have*

$$P\left(\sup_{\boldsymbol{\theta}} \frac{N}{s_i(N)} \left| [\nabla L(\boldsymbol{\theta})]_i - [\nabla L^*(\boldsymbol{\theta})]_i \right| > C_{\boldsymbol{\theta}} \phi \right) \leq \delta(\phi), i = 1, 2.$$

Furthermore, if assumption 3a holds and $s_i(N) = \tau \log N$, then for $N > C_{\boldsymbol{\theta}}, c_{\boldsymbol{\theta}} > 0$, we have

$$\delta(\phi) \leq \frac{C_{\boldsymbol{\theta}}}{N^{c_{\boldsymbol{\theta}}}} + C_{\boldsymbol{\theta}} (\log \phi)^4 \exp\{-c_{\boldsymbol{\theta}} \log N \min\{\phi^2, \phi\}\}.$$

If assumption 3a or 3b holds and $s_i(N) = N$, then

$$\delta(\phi) \leq C_{\boldsymbol{\theta}} (\log \phi)^4 \exp\{-c_{\boldsymbol{\theta}} N \min\{\phi^2, \phi\}\}.$$

4 Proof of Theorems

4.1 Detailed Notations

Let $\boldsymbol{\theta}_k^{(t)}$ be the model parameter maintained in the k^{th} device at the t^{th} step. Let $\mathcal{I}_E = \{cE \mid c = 1, 2, \ldots, R\}$ be the set of global aggregation steps. If $t + 1 \in \mathcal{I}_E$, then the central server collects model parameters from active devices and aggregates all of those model parameters. Motivated by (Li et al., 2019), we introduce an intermediate parameter $\boldsymbol{v}_k^{(t+1)} := \boldsymbol{\theta}_k^{(t)} - \eta^{(t)} g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})$. It can be seen that $\boldsymbol{\theta}_k^{(t+1)} = \boldsymbol{v}_k^{(t+1)}$ if $t + 1 \notin \mathcal{I}_E$ and $\boldsymbol{\theta}_k^{(t+1)} = \sum_{k=1}^K p_k \boldsymbol{v}_k^{(t+1)}$ otherwise. Let $\bar{\boldsymbol{v}}^{(t)} = \sum_{k=1}^K p_k \boldsymbol{v}_k^{(t)}$ and $\bar{\boldsymbol{\theta}}^{(t)} = \sum_{k=1}^K p_k \, \boldsymbol{\theta}_k^{(t)}$. The central server can only obtain $\bar{\boldsymbol{\theta}}^{(t)}$ when $t+1 \in \mathcal{I}_E$. The term $\bar{\boldsymbol{v}}^{(t)}$ is introduced for the purpose of proof and is inaccessible in practice. We further define $g^{(t)} = \sum_{k=1}^{K} p_k g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})$.

4.2 Proof of Theorem 1

Under the scenario of full device participation, we have $\bar{\boldsymbol{\theta}}^{(t+1)} = \bar{\boldsymbol{v}}^{(t+1)}$ for all t. By definition of $\bar{\boldsymbol{v}}^{(t)}$, we have

$$\begin{aligned}
\left\|\bar{\boldsymbol{v}}^{(t+1)} - \boldsymbol{\theta}^*\right\|_2^2 &= \left\|\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)} g^{(t)} - \boldsymbol{\theta}^*\right\|_2^2 \\
&= \underbrace{\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|_2^2}_{\text{A}} \underbrace{-2\eta^{(t)}\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, g^{(t)}\rangle}_{\text{B}} + \eta^{(t)2} \underbrace{\left\|g^{(t)}\right\|_2^2}_{\text{C}}.
\end{aligned}$$

We can write B as

$$\begin{aligned}
\text{B} &= -2\eta^{(t)}\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, g^{(t)}\rangle = -2\eta^{(t)}\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, \sum_{k=1}^{K} p_k g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\rangle \\
&= -2\eta^{(t)} \sum_{k=1}^{K} p_k \langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\rangle \\
&= -2\eta^{(t)} \sum_{k=1}^{K} p_k \langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\rangle - 2\eta^{(t)} \sum_{k=1}^{K} p_k \langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\rangle.
\end{aligned}$$

By Cauchy-Schwarz inequality and inequality of arithmetic and geometric means, we can simplify the first term in B as

$$\begin{aligned}
-2\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\rangle &\leq 2\frac{\sqrt{\eta^{(t)}}}{\sqrt{\eta^{(t)}}} \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\| \left\|g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\right\| \\
&\leq 2\frac{\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + \eta^{(t)}\left\|g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\right\|^2}{2} \\
&\leq \left(\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|_2^2 + \eta^{(t)}\left\|g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\right\|_2^2\right).
\end{aligned}$$

By Lemma 4, we can simplify the second term in B as

$$\begin{aligned}
&-2\eta^{(t)}\langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)})\rangle = -2\eta^{(t)}\langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) + g_k^*(\boldsymbol{\theta}_k^{(t)}) - g_k^*(\boldsymbol{\theta}_k^{(t)})\rangle \\
&\leq -2\eta^{(t)}\frac{\gamma_k}{2}\left\|\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*\right\|_2^2 + 2\eta^{(t)} C_{3k}(\alpha_k, b_k)\frac{\log M_k}{M_k} - 2\eta^{(t)}\langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) - g_k^*(\boldsymbol{\theta}_k^{(t)})\rangle.
\end{aligned}$$

By Assumption 2,

$$\mathrm{C} = \left\| g^{(t)} \right\|_2^2 = \left\| \sum_{k=1}^{K} p_k g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) \right\|^2 \leq \left(\sum_{k=1}^{K} \left\| p_k g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) \right\| \right)^2 \leq \left(\sum_{k=1}^{K} p_k G \right)^2 = G^2.$$

Combining A, B and C together, we obtain

$$\begin{aligned}
&\left\| \bar{\boldsymbol{v}}^{(t+1)} - \boldsymbol{\theta}^* \right\|_2^2 \\
&\leq \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|_2^2 + \eta^{(t)} \sum_{k=1}^{K} p_k \left(\frac{1}{\eta^{(t)}} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|_2^2 + \eta^{(t)} G^2 \right) \\
&\quad - 2\eta^{(t)} \sum_{k=1}^{K} p_k \frac{\gamma_k}{2} \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* \right\|_2^2 + 2\eta^{(t)} \sum_{k=1}^{K} p_k C_{3k}(\alpha_k, b_k) \frac{\log M_k}{M_k} + \eta^{(t)2} G^2 \\
&\quad - 2\eta^{(t)} \sum_{k=1}^{K} p_k \langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) - g_k^*(\boldsymbol{\theta}_k^{(t)}) \rangle \\
&= \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|_2^2 + \underbrace{\sum_{k=1}^{K} p_k \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|_2^2}_{\mathrm{D}} + \eta^{(t)2} G^2 \\
&\quad - 2\eta^{(t)} \underbrace{\sum_{k=1}^{K} p_k \frac{\gamma_k}{2} \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* \right\|_2^2}_{\mathrm{E}} + 2\eta^{(t)} \sum_{k=1}^{K} p_k C_{3k}(\alpha_k, b_k) \frac{\log M_k}{M_k} + \eta^{(t)2} G^2 \\
&\quad - 2\eta^{(t)} \sum_{k=1}^{K} p_k \langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) - g_k^*(\boldsymbol{\theta}_k^{(t)}) \rangle.
\end{aligned}$$

Since the aggregation step happens each E steps, for any $t \geq 0$, there exists a $t_0 \leq t$ such that $t - t_0 \leq E - 1$

and $\boldsymbol{\theta}_k^{(t_0)} = \bar{\boldsymbol{\theta}}^{(t_0)}$ for all $k \in [K]$. Since $\eta^{(t)}$ is non-increasing, for all $t - t_0 \leq E - 1$, we can simplify D as

$$\begin{aligned}
\mathrm{D} &= \sum_{k=1}^{K} p_k \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|_2^2 = \sum_{k=1}^{K} p_k \left\| (\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)}) - (\bar{\boldsymbol{\theta}}^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)}) \right\|_2^2 \\
&\leq \sum_{k=1}^{K} p_k \left\| \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)} \right\|_2^2 + \underbrace{\sum_{k=1}^{K} p_k \left\| \bar{\boldsymbol{\theta}}^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)} \right\|_2^2}_{\sum p_k = 1} \\
&= \sum_{k=1}^{K} p_k \left\| \sum_{t=t_0}^{t-1} \eta^{(t)} g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) \right\|_2^2 + \left\| \sum_{k=1}^{K} p_k \sum_{t=t_0}^{t-1} \eta^{(t)} g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) \right\|_2^2 \\
&\leq \sum_{k=1}^{K} p_k (t - t_0) \sum_{t=t_0}^{t-1} \eta^{(t)2} \left\| g(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) \right\|_2^2 + \sum_{k=1}^{K} p_k \left\| \sum_{t=t_0}^{t-1} \eta^{(t)} g_k(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) \right\|_2^2 \\
&\leq 2 \sum_{k=1}^{K} p_k (E-1) \sum_{t=t_0}^{t-1} \eta^{(t)2} \left\| g(\boldsymbol{\theta}_k^{(t)}; \xi_k^{(t)}) \right\|_2^2 \leq 2 \sum_{k=1}^{K} p_k (E-1) \sum_{t=t_0}^{t-1} \eta^{(t_0)2} G^2 = 2 \sum_{k=1}^{K} p_k (E-1)^2 \eta^{(t_0)2} G^2.
\end{aligned}$$

Without loss of generality, assume $\eta^{(t_0)} \leq 2\eta^{(t)}$ since the learning rate is decreasing. Therefore, $\mathrm{D} \leq 8(E-1)^2 \eta^{(t)2} G^2$. To simplify E, we have

$$\begin{aligned}
\mathrm{E} = \sum_{k=1}^{K} p_k \frac{\gamma_k}{2} \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* \right\|_2^2 &\geq \min_k \gamma_k \frac{1}{2} \sum_{k=1}^{K} p_k \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* \right\|_2^2 \\
&\geq \min_k \gamma_k \frac{1}{2} \left\| \sum_{k=1}^{K} p_k (\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*) \right\|_2^2 = \min_k \gamma_k \frac{1}{2} \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|_2^2,
\end{aligned}$$

using Jensen's inequality.

Therefore, we obtain

$$\begin{aligned}
&\left\|\bar{\boldsymbol{v}}^{(t+1)}-\boldsymbol{\theta}^*\right\|_2^2\\
&\leq \left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|_2^2+8(E-1)^2\eta^{(t)2}G^2+\eta^{(t)2}G^2\\
&\quad -2\eta^{(t)}\min_k \gamma_k\frac{1}{2}\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|_2^2+2\eta^{(t)}\sum_{k=1}^K p_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+\eta^{(t)2}G^2\\
&\quad -2\eta^{(t)}\sum_{k=1}^K p_k\langle\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)};\xi_k^{(t)})-g_k^*(\boldsymbol{\theta}_k^{(t)})\rangle\\
&=\left(1-2\eta^{(t)}\min_k\gamma_k\frac{1}{2}\right)\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|_2^2+(8(E-1)^2\eta^{(t)2}+2\eta^{(t)2})G^2+2\eta^{(t)}\sum_{k=1}^K p_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}\\
&\quad -2\eta^{(t)}\sum_{k=1}^K p_k\langle\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)};\xi_k^{(t)})-g_k^*(\boldsymbol{\theta}_k^{(t)})\rangle\\
&\leq\left(1-2\eta^{(t)}\min_k\gamma_k\frac{1}{2}\right)\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|_2^2+(8(E-1)^2\eta^{(t)2}+2\eta^{(t)2})G^2+2\eta^{(t)}\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}\\
&\quad -2\eta^{(t)}\sum_{k=1}^K p_k\langle\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)};\xi_k^{(t)})-g_k^*(\boldsymbol{\theta}_k^{(t)})\rangle\\
&=\left(1-2\eta^{(t)}\min_k\gamma_k\frac{1}{2}\right)\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|_2^2+\left(8(E-1)^2+2\right)\eta^{(t)2}G^2+2\eta^{(t)}\Bigg(\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}\\
&\quad -\sum_{k=1}^K p_k\langle\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)};\xi_k^{(t)})-g_k^*(\boldsymbol{\theta}_k^{(t)})\rangle\Bigg).
\end{aligned}$$

Since $\frac{3}{2\min_k\gamma_k}\leq\beta_1\leq\frac{2}{\min_k\gamma_k}$ and $\eta^{(t)}=\frac{\beta_1}{t}$ for all $t\geq 1$. Here we set $\eta^{(0)}=\beta_1$. We will show

$$\begin{aligned}
&\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|_2^2\leq\frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+1}\\
&\quad+\sum_{u=0}^{t-1}\left(2\eta^{(u+1)}\prod_{v=u+2}^{t}(1-\eta^{(v)}\min_k\gamma_k)\right)\left(\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}-\sum_{k=1}^K p_k\langle\boldsymbol{\theta}_k^{(u)}-\boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)})-g_k^*(\boldsymbol{\theta}_k^{(u)})\rangle\right)
\end{aligned}$$

by induction. When $t=1$, we have

$$\begin{aligned}
\left\|\bar{\boldsymbol{\theta}}^{(1)}-\boldsymbol{\theta}^*\right\|_2^2&\leq\left(8(E-1)^2+2\right)\beta_1^2G^2+2\beta_1\Bigg(\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}\\
&\quad-\sum_{k=1}^K p_k\langle\boldsymbol{\theta}_k^{(0)}-\boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(0)};\xi_k^{(0)})-g_k^*(\boldsymbol{\theta}_k^{(0)})\rangle\Bigg)
\end{aligned}$$

since $\left(1 - 2\eta^{(0)} \min_k \gamma_k \frac{1}{2}\right) < 0$. Assume the inequality holds for $t = l \geq 1$, then we have

$$
\begin{aligned}
&\left\|\bar{\boldsymbol{\theta}}^{(l+1)} - \boldsymbol{\theta}^*\right\|_2^2 \\
&\leq \left(1 - 2\eta^{(l)} \min_k \gamma_k \frac{1}{2}\right) \Bigg\{ \frac{2\beta_1^2 \left(8(E-1)^2 + 2\right) G^2}{l+1} \\
&\quad + \sum_{u=0}^{l-1} 2\eta^{(u+1)} \prod_{v=u+2}^{l} (1 - \eta^{(v)} \min_k \gamma_k) \Bigg(\max_k C_{3k}(\alpha_k, b_k) \frac{\log M_k}{M_k} - \sum_{k=1}^{K} p_k \langle \boldsymbol{\theta}_k^{(u)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(u)}; \xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)}) \rangle \Bigg) \Bigg\} \\
&\quad + \left(8(E-1)^2 + 2\right) \eta^{(l)2} G^2 + 2\eta^{(l)} \Bigg(\max_k C_{3k}(\alpha_k, b_k) \frac{\log M_k}{M_k} - \sum_{k=1}^{K} p_k \langle \boldsymbol{\theta}_k^{(l)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(l)}; \xi_k^{(l)}) - g_k^*(\boldsymbol{\theta}_k^{(l)}) \rangle \Bigg) \\
&\leq \frac{2\beta_1^2 \left(8(E-1)^2 + 2\right) G^2}{l+2} \\
&\quad + \sum_{u=0}^{l} 2\eta^{(u+1)} \prod_{v=u+2}^{l} (1 - \eta^{(v)} \min_k \gamma_k) \Bigg(\max_k C_{3k}(\alpha_k, b_k) \frac{\log M_k}{M_k} - \sum_{k=1}^{K} p_k \langle \boldsymbol{\theta}_k^{(u)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(u)}; \xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)}) \rangle \Bigg).
\end{aligned}
$$

To derive above inequality, we can first show that

$$
\left(1 - 2\eta^{(l)} \min_k \gamma_k \frac{1}{2}\right) \frac{2\beta_1^2 \left(8(E-1)^2 + 2\right) G^2}{l+1} + \left(8(E-1)^2 + 2\right) \eta^{(l)2} G^2 \leq \frac{2\beta_1^2 \left(8(E-1)^2 + 2\right) G^2}{l+2}
$$

as long as $\beta_1 \geq \frac{3l+1}{2l+2} \frac{1}{\min_k \gamma_k}$. This is true since the right-hand side is always less or equal to $\frac{3}{2 \min_k \gamma_k}$. The remaining part in the above inequality is apparent since $1 - 2\eta^{(l)} \min_k \gamma_k \frac{1}{2} \leq 1$. Thus, the proof of the induction step is complete. Using this fact, it can be shown that

$$
\begin{aligned}
&\left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^*\right\|_2^2 \\
&\le \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2} \\
&\quad + \sum_{u=0}^{t} 2\eta^{(u+1)} \prod_{v=u+2}^{t}(1-\eta^{(v)}\min_k \gamma_k)\Big(\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k} - \sum_{k=1}^{K} p_k\langle \boldsymbol{\theta}_k^{(u)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)})\rangle\Big) \\
&\le \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2} \\
&\quad + \sum_{u=0}^{t}\left(2\eta^{(u+1)} \prod_{v=u+2}^{t}(1-\eta^{(v)}\min_k \gamma_k)\right)\left(\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}\right) \\
&\quad - \sum_{u=0}^{t}\left(2\eta^{(u+1)} \prod_{v=u+2}^{t}(1-\eta^{(v)}\min_k \gamma_k)\right)\left(\sum_{k=1}^{K} p_k\langle \boldsymbol{\theta}_k^{(u)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)})\rangle\right) \\
&\le \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2} \\
&\quad + \sum_{u=0}^{t}\frac{2\beta_1}{t+1}\left(\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}\right) + \sum_{u=0}^{t}\frac{2\beta_1}{t+1}\left(\sum_{k=1}^{K} p_k\left\|\boldsymbol{\theta}_k^{(u)} - \boldsymbol{\theta}^*\right\|_2\left\|g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)})\right\|_2\right) \\
&\le \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2} \\
&\quad + \left(\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}\right) + \sum_{u=0}^{t}\frac{2\beta_1}{t+1}\left(\sum_{k=1}^{K} p_k\left\|\boldsymbol{\theta}_k^{(u)} - \boldsymbol{\theta}^*\right\|_2\left\|g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)})\right\|_2\right) \\
&\le \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2} \\
&\quad + \max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k} + \frac{2\beta_1}{t+1}\sum_{u=0}^{t}\left(\sum_{k=1}^{K}\sqrt{2}p_k(\theta_{max}-\theta_{min})\left\|g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)})\right\|_2\right) \\
&\le \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2} \\
&\quad + 2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k} + 2\beta_1\max_{0\le u\le t}\left(\sum_{k=1}^{K}\sqrt{2}p_k(\theta_{max}-\theta_{min})\left\|g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)})\right\|_2\right)
\end{aligned}
$$

In the third inequality, we use the Cauchy–Schwarz inequality and the fact that $2\eta^{(u+1)}\prod_{v=u+2}^{t}(1-\eta^{(v)}\min_k\gamma_k) \le 2\frac{\beta_1}{u+1}\prod_{v=u+2}^{t}(1-\frac{3}{2v}) \le \frac{2\beta_1}{t+1}$.

Let $\phi_k = (\log M_k)^{\epsilon_k - \frac{1}{2}}$. By Lemma 6 and using a union bound over u, with probability at least $1 - C_{\boldsymbol{\theta}}(T+1)\exp(-c_{\boldsymbol{\theta}}(\log M_k)^{2\epsilon_k})$, we have

$$\max_{0\le u\le t}\left\|g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)})\right\|_2 \le C_{\boldsymbol{\theta}}(\log M_k)^{\epsilon_k-\frac{1}{2}}.$$

Therefore,

$$
\begin{aligned}
&\left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*\right\|_2^2 \\
\leq\ & \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2}+2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+2\beta_1\max_{0\leq u\leq t}\sum_{k=1}^K\sqrt{2}p_k(\theta_{max}-\theta_{min})C_{\boldsymbol{\theta}}(\log M_k)^{\epsilon_k-\frac{1}{2}} \\
=\ & \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2}+2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+2\sqrt{2}\beta_1(\theta_{max}-\theta_{min})C_{\boldsymbol{\theta}}\sum_{k=1}^K p_k(\log M_k)^{\epsilon_k-\frac{1}{2}}.
\end{aligned}
$$

Using the same proof technique, we can also derive a same bound on $\left\|\bar{\theta}_2^{(t+1)}-\theta_2^*\right\|_2^2$. Let $\phi_k=M_k^{\epsilon_k-\frac{1}{2}}$. By Lemma 6 and using a union bound over u, with probability at least $1-C_\theta(t+1)(\log(M_k^{\epsilon_k-\frac{1}{2}}))^4\exp\{-c_\theta M_k^{2\epsilon_k}\}$,

$$
\max_{0\leq u\leq t}\left\|g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)})-g_k^*(\boldsymbol{\theta}_k^{(u)})\right\|_2\leq CM_k^{\epsilon_k-\frac{1}{2}}.
$$

Therefore,

$$
\begin{aligned}
&\left\|\bar{\theta}_2^{(t+1)}-\theta_2^*\right\|_2^2 \\
\leq\ & \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2}+2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+2\beta_1\max_{0\leq u\leq t}\sum_{k=1}^K\sqrt{2}p_k(\theta_{max}-\theta_{min})CM_k^{\epsilon_k-\frac{1}{2}} \\
=\ & \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+2}+2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+2\sqrt{2}\beta_1(\theta_{max}-\theta_{min})C\sum_{k=1}^K p_kM_k^{\epsilon_k-\frac{1}{2}}.
\end{aligned}
$$

4.3 Proof of Theorem 2

We slightly modify the definition of $\boldsymbol{\theta}_k^{(t+1)}$ such that $\boldsymbol{\theta}_k^{(t+1)}=\frac{1}{|\mathcal{S}_c|}\sum_{k\in\mathcal{S}_c}\boldsymbol{v}_k^{(t+1)}$ if $t+1\in\mathcal{I}_E$. Under the scenario of asynchronous update, it can be seen that $\bar{\boldsymbol{\theta}}^{(t+1)}\neq\bar{\boldsymbol{v}}^{(t+1)}$. Therefore, we want to establish a bound on the difference $\left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\bar{\boldsymbol{v}}^{(t+1)}\right\|_2^2$. We have

$$
\begin{aligned}
\left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*\right\|_2^2 &= \left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\bar{\boldsymbol{v}}^{(t+1)}+\bar{\boldsymbol{v}}^{(t+1)}-\boldsymbol{\theta}^*\right\|_2^2 \\
&= \underbrace{\left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\bar{\boldsymbol{v}}^{(t+1)}\right\|_2^2}_{\text{A}}+\underbrace{\left\|\bar{\boldsymbol{v}}^{(t+1)}-\boldsymbol{\theta}^*\right\|_2^2}_{\text{B}}+\underbrace{2\langle\bar{\boldsymbol{\theta}}^{(t+1)}-\bar{\boldsymbol{v}}^{(t+1)},\bar{\boldsymbol{v}}^{(t+1)}-\boldsymbol{\theta}^*\rangle}_{\text{C}}.
\end{aligned}
$$

We can show

$$
\mathbb{E}_{\mathcal{S}_c}\left\{\bar{\boldsymbol{\theta}}^{(t+1)}\right\}=\mathbb{E}_{\mathcal{S}_c}\left\{\frac{1}{|\mathcal{S}_c|}\sum_{k\in\mathcal{S}_c}\boldsymbol{v}_k^{(t+1)}\right\}=\frac{1}{|\mathcal{S}_c|}\sum_{k\in\mathcal{S}_c}\mathbb{E}_{\mathcal{S}_c}\left\{\boldsymbol{v}_k^{(t+1)}\right\}=\mathbb{E}_{\mathcal{S}_c}\left\{\boldsymbol{v}_1^{(t+1)}\right\}=\sum_{k=1}^K p_k\boldsymbol{v}_k^{(t+1)}=\bar{\boldsymbol{v}}^{(t+1)}
$$

since the sampling distribution is identical. Therefore, $\mathbb{E}_{\mathcal{S}_c}[\mathrm{C}] = 0$.

For part A, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_c}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{v}}^{(t+1)}\right\|_2^2\right\} \\
&= \mathbb{E}_{\mathcal{S}_c}\left\{\frac{1}{|\mathcal{S}_c|^2}\sum_{k\in\mathcal{S}_c}\left\|\boldsymbol{v}_k^{(t+1)} - \bar{\boldsymbol{v}}^{(t+1)}\right\|_2^2\right\} = \frac{1}{|\mathcal{S}_c|}\sum_{k=1}^{K} p_k\left\|\boldsymbol{v}_k^{(t+1)} - \bar{\boldsymbol{v}}^{(t+1)}\right\|_2^2.
\end{aligned}$$

The first equality uses the fact that $\boldsymbol{v}_k^{(t+1)}$ is independent of each other and is an unbiased estimator of $\bar{\boldsymbol{v}}^{(t+1)}$. Therefore, we have

$$\begin{aligned}
\sum_{k=1}^{K} p_k\left\|\boldsymbol{v}_k^{(t+1)} - \bar{\boldsymbol{v}}^{(t+1)}\right\|_2^2 &= \sum_{k=1}^{K} p_k\left\|\boldsymbol{v}_k^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t_0)} - (\bar{\boldsymbol{\theta}}^{(t_0)} - \bar{\boldsymbol{v}}^{(t+1)})\right\|_2^2 \\
&\le \sum_{k=1}^{K} p_k\left\|\boldsymbol{v}_k^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t_0)}\right\|_2^2 \\
&\le \sum_{k=1}^{K} p_k\left\|\boldsymbol{v}_k^{(t+1)} - \boldsymbol{\theta}_k^{(t_0)}\right\|_2^2 \\
&= \sum_{k=1}^{K} p_k\left\|\sum_{i=t_0}^{t}\eta^{(i)} g_k(\boldsymbol{\theta}_k^{(i)};\xi_k^{(i)})\right\|_2^2 \\
&\le \sum_{k=1}^{K} p_k\sum_{i=t_0}^{t} E\left\|\eta^{(i)} g_k(\boldsymbol{\theta}_k^{(i)};\xi_k^{(i)})\right\|_2^2 \\
&\le E^2\eta^{(t_0)2}G^2 \le 4E^2\eta^{(t)2}G^2
\end{aligned}$$

where $t_0 = t - E + 1$ is the iteration where communication happens. Therefore,

$$\mathbb{E}_{\mathcal{S}_c}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{v}}^{(t+1)}\right\|_2^2\right\} \le \frac{4E^2\eta^{(t)2}G^2}{|\mathcal{S}_c|}.$$

For part B, we can follow the exact proof in Theorem 1 to get an upper bound after taking expectation with respect to $\mathcal{S}_c$. In a nutshell, we can obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_c}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*\right\|_2^2\right\}\\
&\leq \frac{4E^2\eta^{(t)2}G^2}{|\mathcal{S}_c|}+\left(1-2\eta^{(t)}\min_k\gamma_k\frac{1}{2}\right)\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|_2^2+\left(8(E-1)^2+2\right)\eta^{(t)2}G^2+2\eta^{(t)}\Bigg(\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}\\
&\qquad-\sum_{k=1}^K p_k\langle\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*,g_k(\boldsymbol{\theta}_k^{(t)};\xi_k^{(t)})-g_k^*(\boldsymbol{\theta}_k^{(t)})\rangle\Bigg).
\end{aligned}$$

Following the induction proof in Theorem 1, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_c}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*\right\|_2^2\right\}\\
&\leq \frac{2\beta_1^2\left(\frac{1}{|\mathcal{S}_c|}4E^2+8(E-1)^2+2\right)G^2}{t+2}+2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+2\sqrt{2}\beta_1(\theta_{max}-\theta_{min})C\sum_{k=1}^K p_k M_k^{\epsilon_k-\frac{1}{2}}.
\end{aligned}$$

4.4 Proof of Theorem 3

Convergence of Parameter Iterate

Define $C_{4k}:=(\theta_{max}-\theta_{min})^2\left(\frac{1}{2\theta_{min}^2}+\frac{C_{mat,k}(4b_k+3)}{2\theta_{min}^2(4b_k^2-6b_k-3)}\right)$. Following the same proof strategy in Theorem 1 and using Lemma 3 and 5, we can show that

$$\begin{aligned}
&\left\|\bar{\theta}_2^{(t+1)}-\theta_2^*\right\|_2^2\\
&\leq \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+1}+\sum_{u=0}^t 2\eta^{(u+1)}\prod_{v=u+2}^t(1-\eta^{(v)}\min_k\gamma_k)\Bigg(\max_k C_{4k}M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1}-\\
&\qquad\sum_{k=1}^K p_k\langle\theta_{2k}^{(u)}-\theta_2^*,[g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)})]_2-[g_k^*(\boldsymbol{\theta}_k^{(u)})]_2\rangle\Bigg)\\
&\leq \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+1}+2\beta_1\max_k C_{4k}M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1}\\
&\qquad+2\beta_1\max_{0\leq u\leq t}\left(\sum_{k=1}^K\sqrt{2}p_k(\theta_{max}-\theta_{min})\left\|[g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)})]_2-[g_k^*(\boldsymbol{\theta}_k^{(u)})]_2\right\|_2\right).
\end{aligned}$$

Let $\phi_k=M_k^{\epsilon_k-\frac{1}{2}}$. By Lemma 6, for any $0<\alpha_k<\frac{8b_k^2-12b_k-6}{4b_k+3},\epsilon_k<\frac{1}{2}$, with probability at least $1-C_{\boldsymbol{\theta}}(t+1)(\log(M_k^{\epsilon_k-\frac{1}{2}}))^4\exp\{-c_{\boldsymbol{\theta}}M_k^{2\epsilon_k}\}$, we have

$$\max_{0\leq u\leq t}\left\|[g_k(\boldsymbol{\theta}_k^{(u)};\xi_k^{(u)})]_2-[g_k^*(\boldsymbol{\theta}_k^{(u)})]_2\right\|_2\leq C_{\boldsymbol{\theta}}M_k^{\epsilon_k-\frac{1}{2}}.$$

Therefore,

$$
\begin{aligned}
&\left\|\bar{\theta}_2^{(t+1)} - \theta_2^*\right\|_2^2 \\
&\leq \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+1} + 2\beta_1 \max_k C_{4k} M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1} + 2\beta_1\Big(\sum_{k=1}^K \sqrt{2} p_k(\theta_{max}-\theta_{min}) C_{\boldsymbol{\theta}} M_k^{\epsilon_k-\frac{1}{2}}\Big) \\
&= \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+1} + \mathcal{O}\left(\max_k M_k^{-\frac{8b_k^2-12b_k-6-3\alpha_k-4\alpha_k b_k}{8b_k^2-4b_k}}\right) + \mathcal{O}\left(\sum_{k=1}^K p_k M_k^{\epsilon_k-\frac{1}{2}}\right).
\end{aligned}
$$

The partial device participation proof is similar to Theorem 2. Again, using Lemma 3 and 5, we can show that

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_c}\left\{\left\|\bar{\theta}_2^{(t+1)} - \theta_2^*\right\|_2^2\right\} \\
&\leq \frac{2\beta_1^2\left(\frac{4E^2}{|\mathcal{S}_c|}+8(E-1)^2+2\right)G^2}{t+1} + 2\beta_1 \max_k C_{4k} M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1} + 2\beta_1\Big(\sum_{k=1}^K \sqrt{2} p_k(\theta_{max}-\theta_{min}) C_{\boldsymbol{\theta}} M_k^{\epsilon_k-\frac{1}{2}}\Big) \\
&= \frac{2\beta_1^2\left(\frac{4E^2}{|\mathcal{S}_c|}+8(E-1)^2+2\right)G^2}{t+1} + \mathcal{O}\left(\max_k M_k^{-\frac{8b_k^2-12b_k-6-3\alpha_k-4\alpha_k b_k}{8b_k^2-4b_k}}\right) + \mathcal{O}\left(\sum_{k=1}^K p_k M_k^{\epsilon_k-\frac{1}{2}}\right).
\end{aligned}
$$

Convergence of Full Gradient

We follow the same proof strategy in Theorem 4. We defer this proof to the subsection after it.

4.5 Proof of Theorem 4

Proof. Our final goal is to bound the squared norm of full gradient $\left\|\nabla L(\bar{\boldsymbol{\theta}})\right\|_2^2 = \left\|\sum_{k=1}^K p_k \nabla L_k(\bar{\boldsymbol{\theta}}; D_k)\right\|_2^2$. We define a conditional expectation of $\nabla L_k(\bar{\boldsymbol{\theta}}^{(t)}; D_k)$ as $\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)}) := \mathbb{E}\left(\nabla L_k(\bar{\boldsymbol{\theta}}^{(t)}; D_k) | \boldsymbol{X}_k\right)$. By the definition of $\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})$, for $i \in \{1, 2\}$, we have

$$
\begin{aligned}
&\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_i \\
&= \frac{1}{2N_k} \operatorname{Tr}\left[\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})^{-1}\left(\boldsymbol{I}_{N_k} - \boldsymbol{K}_{N_k}(\boldsymbol{\theta}_k^*)\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})^{-1}\right)\frac{\partial \boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})}{\partial \bar{\theta}_i^{(t)}}\right] \\
&= \frac{1}{2N_k} \operatorname{Tr}\left[\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})^{-1}\left(\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)}) - \boldsymbol{K}_{N_k}(\boldsymbol{\theta}_k^*)\right)\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})^{-1}\frac{\partial \boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})}{\partial \bar{\theta}_i^{(t)}}\right].
\end{aligned}
$$

where $\boldsymbol{\theta}_k^* := (\theta_{1k}^*, \theta_{2k}^*)$ is the set of optimal model parameters for device k. By definition, $\boldsymbol{K}_{N_k}(\boldsymbol{\theta}_k) = \theta_{1k}\boldsymbol{K}_{f,N_k} + \theta_{2k}\boldsymbol{I}_{N_k}$, where θ_{1k}, θ_{2k} are device-specific model parameters. Therefore, we obtain

$$\begin{aligned}&\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_i \\ &= \frac{1}{2N_k} \operatorname{Tr}\left[\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})^{-1}\left((\bar{\theta}_1^{(t)} - \theta_{1k}^*)\boldsymbol{K}_{f,N_k} + (\bar{\theta}_2^{(t)} - \theta_{2k}^*)\boldsymbol{I}_{N_k}\right)\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})^{-1}\frac{\partial \boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})}{\partial \bar{\theta}_i^{(t)}}\right],\end{aligned}$$

where $\bar{\theta}_i^{(t)} = \sum_{k=1}^K p_k \theta_{ik}^{(t)}, i = 1, 2$.

By Eigendecomposition, we can write $\boldsymbol{K}_{f,N_k} = \boldsymbol{Q}_{N_k}\boldsymbol{\Lambda}_{N_k}\boldsymbol{Q}_{N_k}^{-1}$ where $\boldsymbol{Q}_{N_k}$ contains eigenvectors of $\boldsymbol{K}_{f,N_k}$, $\boldsymbol{\Lambda}_{N_k} := \text{diag}(\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1N_k})$ is a diagonal matrix with eigenvalues of $\boldsymbol{K}_{f,N_k}$ and λ_{1j} is the j^{th} largest eigenvalue of $\boldsymbol{K}_{f,N_k}$. Here note that the values of $\lambda_{\cdot\cdot}$ are different for each device k. For simplicity, we drop the notation k in the eigenvalues unless there is an ambiguity. When $i = 1$, we can simplify $\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_i$ as

$$\begin{aligned}&\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_1 \\ &= \frac{1}{2N_k} \operatorname{Tr}\left[\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})^{-1}\left((\bar{\theta}_1^{(t)} - \theta_{1k}^*)\boldsymbol{K}_{f,N_k} + (\bar{\theta}_1^{(t)} - \theta_{2k}^*)\boldsymbol{I}_{N_k}\right)\boldsymbol{K}_{N_k}(\bar{\boldsymbol{\theta}}^{(t)})^{-1}\boldsymbol{K}_{f,N_k}\right] \\ &= \frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_{1k}^*)\sum_{j=1}^{N_k}\frac{\lambda_{1j}^2}{\left(\bar{\theta}_1^{(t)}\lambda_{1j} + \bar{\theta}_2^{(t)}\lambda_{2j}\right)^2} + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_{2k}^*)\sum_{j=1}^{N_k}\frac{\lambda_{2j}\lambda_{1j}}{\left(\bar{\theta}_1^{(t)}\lambda_{1j} + \bar{\theta}_2^{(t)}\lambda_{2j}\right)^2}.\end{aligned}$$

where $\lambda_{2j} = 1$ is the j^{th} largest eigenvalue of $\boldsymbol{I}_{N_k}$. Similarly, it can be shown that, when $i = 2$,

$$\begin{aligned}&\left[\nabla L^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_2 \\ &= \frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_{1k}^*)\sum_{j=1}^{N_k}\frac{\lambda_{1j}\lambda_{2j}}{\left(\bar{\theta}_1^{(t)}\lambda_{1j} + \bar{\theta}_2^{(t)}\lambda_{2j}\right)^2} + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_{2k}^*)\sum_{j=1}^{N_k}\frac{\lambda_{2j}^2}{\left(\bar{\theta}_1^{(t)}\lambda_{1j} + \bar{\theta}_2^{(t)}\lambda_{2j}\right)^2}.\end{aligned}$$

Our first goal is to bound eigenvalues of $\boldsymbol{K}_{f,N_k}$ using Lemma 1 and 2.

Part I: Bounding eigenvalues By Lemma 1 and 2, for any $0 < \epsilon_k, \alpha_k < 1$, $C_{1k}(\alpha, b) > 0$ and

$N_k > C_{2k}(\epsilon_k, b_k)$, with probability at least $1 - \frac{3}{N_k^{\alpha_k}}$,

$$\begin{aligned}
\frac{\epsilon_k \log N_k}{8 b_k \theta_{max}^2} &\leq \sum_{j=1}^{N_k} \frac{\lambda_{1j}^2}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} \leq \frac{4 + 2\alpha_k}{b_k \theta_{min}^2} \log N_k \\
\frac{N_k - C_{1k}(\alpha_k, b_k) \log N_k}{4 \theta_{max}^2} &\leq \sum_{j=1}^{N_k} \frac{\lambda_{2j}^2}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} \leq \frac{N_k}{\theta_{min}^2} \\
0 &< \sum_{j=1}^{N_k} \frac{\lambda_{1j} \lambda_{2j}}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} \leq \frac{5 + 2\alpha_k}{7 b_k \theta_{min}^2} \log N_k.
\end{aligned}$$

Therefore, we can show that, with probability at least $1 - \frac{3}{N_k^{\alpha_k}}$,

$$\begin{aligned}
\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_1 &= \frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_{1k}^*) \sum_{j=1}^{N_k} \frac{\lambda_{1j}^2}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_{2k}^*) \sum_{j=1}^{N_k} \frac{\lambda_{2j} \lambda_{1j}}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} \\
&\leq \frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_{1k}^*) \frac{4 + 2\alpha_k}{b_k \theta_{min}^2} \log N_k + \frac{1}{2N_k} \frac{5 + 2\alpha_k}{7 b_k \theta_{min}^2} \log N_k \\
&\leq \frac{(\theta_{max} - \theta_{min})(33 + 16\alpha_k)}{14 N_k b_k \theta_{min}^2} \log N_k = \frac{(\theta_{max} - \theta_{min})(33 + 16\alpha_k)}{14 b_k \theta_{min}^2} \frac{\log N_k}{N_k},
\end{aligned}$$

and

$$\begin{aligned}
\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_1 &= \frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_{1k}^*) \sum_{j=1}^{N_k} \frac{\lambda_{1j}^2}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_{2k}^*) \sum_{j=1}^{N} \frac{\lambda_{2j} \lambda_{1j}}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} \\
&\geq \frac{1}{2N_k} \frac{\epsilon_k \log N_k}{8 b_k \theta_{max}^2} = \frac{\epsilon_k}{16 b_k \theta_{max}^2} \frac{\log N_k}{N_k} > 0.
\end{aligned}$$

Similarly, it can be shown that

$$\begin{aligned}
&\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_2 \\
&= \frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_{1k}^*) \sum_{j=1}^{N_k} \frac{\lambda_{1j} \lambda_{2j}}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_{2k}^*) \sum_{j=1}^{N_k} \frac{\lambda_{2j}^2}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} \\
&\leq \frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_{1k}^*) \frac{5 + 2\alpha_k}{7 b_k \theta_{min}^2} \log N_k + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_{2k}^*) \frac{N_k}{\theta_{min}^2} \\
&\leq \frac{(\theta_{max} - \theta_{min})(5 + 2\alpha_k)}{14 b_k \theta_{min}^2} \frac{\log N_k}{N_k} + (\bar{\theta}_2^{(t)} - \theta_{2k}^*) \frac{1}{2\theta_{min}^2},
\end{aligned}$$

and

$$
\begin{aligned}
&\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_2 \\
&= \frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_{1k}^*)\sum_{j=1}^{N_k} \frac{\lambda_{1j}\lambda_{2j}}{\left(\bar{\theta}_1^{(t)}\lambda_{1j} + \bar{\theta}_2^{(t)}\lambda_{2j}\right)^2} + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_{2k}^*)\sum_{j=1}^{N_k} \frac{\lambda_{2j}^2}{\left(\bar{\theta}_1^{(t)}\lambda_{1j} + \bar{\theta}_2^{(t)}\lambda_{2j}\right)^2} \\
&\geq \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_{2k}^*)\frac{N_k - C_{1k}(\alpha_k, b_k)\log N_k}{4\theta_{max}^2} \\
&\geq (\bar{\theta}_2^{(t)} - \theta_{2k}^*)\frac{1}{8\theta_{max}^2} - \frac{(\theta_{max} - \theta_{min})C_{1k}(\alpha_k, b_k)}{8\theta_{max}^2}\frac{\log N_k}{N_k}.
\end{aligned}
$$

By combining above inequalities, we obtain

$$
\begin{aligned}
&\left\|\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2^2 \\
&= \left(\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_1^2 + \left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_2^2\right) \\
&\leq \left\{\left(\frac{(\theta_{max} - \theta_{min})(33 + 16\alpha_k)}{14b_k\theta_{min}^2}\frac{\log N_k}{N_k}\right)^2 + \left(\frac{(\theta_{max} - \theta_{min})(5 + 2\alpha_k)}{14b_k\theta_{min}^2}\frac{\log N_k}{N_k} + (\bar{\theta}_2^{(t)} - \theta_{2k}^*)\frac{1}{2\theta_{min}^2}\right)^2\right\}.
\end{aligned}
$$

Our next goal is therefore to study the behavior of $\bar{\theta}_2^{(t)} - \theta_{2k}^*$ during iteration and provide bound on this parameter iterate.

Part II: Bounding parameter iterates We consider the full device participation scenario and the partial device participation scenario separately.

Under the full device participation scenario, following the same procedure in the proof of Theorem 1, we can show that

$$
\begin{aligned}
\left\|\bar{\theta}_2^{(t+1)} - \bar{\theta}_{2k}^*\right\|_2^2 \leq &\frac{2\beta_1^2\left(8(E-1)^2 + 2\right)G^2}{t+2} \\
&+ 2\beta_1 \max_k C_{3k}(\alpha_k, b_k)\frac{\log M_k}{M_k} + 2\beta_1 \max_{0\leq u\leq t}\left(\sum_{k=1}^K \sqrt{2}p_k(\theta_{max} - \theta_{min})\left\|g_k(\boldsymbol{\theta}_k^{(u)}; \xi_k^{(u)}) - g_k^*(\boldsymbol{\theta}_k^{(u)})\right\|_2\right) \\
\leq &\frac{2\beta_1^2\left(8(E-1)^2 + 2\right)G^2}{t+2} + 2\beta_1 \max_k C_{3k}(\alpha_k, b_k)\frac{\log M_k}{M_k} + 2\sqrt{2}\beta_1(\theta_{max} - \theta_{min})CM_k^{\epsilon_k - \frac{1}{2}},
\end{aligned}
$$

with probability at least $1 - C_{\theta,k}(t+1)(\log(M_k^{\epsilon - \frac{1}{2}}))^4 \exp\{-c_{\theta,k}M_k^{2\epsilon_k}\}$.

Under the partial device participation scenario, following the same procedure in the proof of Theorem

2, we can show

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_c}\left\{\left\|\bar{\theta}^{(t+1)}-\bar{\theta}^*_{2k}\right\|_2^2\right\}\\
&\le \frac{2\beta_1^2\left(\frac{1}{|\mathcal{S}_c|}4E^2+8(E-1)^2+2\right)G^2}{t+2}+2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+2\sqrt{2}\beta_1(\theta_{max}-\theta_{min})C_kM_k^{\epsilon_k-\frac{1}{2}}.
\end{aligned}
$$

Part III: Bounding $\left[\nabla L(\bar{\boldsymbol{\theta}}^{(t)})\right]_i$ for $i=1,2$ and Proving convergence Finally, equipped with all aforementioned results, we are going to prove our convergence result.

From Part I, we know

$$
\begin{aligned}
&\left\|\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2^2\\
&\le \left\{\left(\frac{(\theta_{max}-\theta_{min})(33+16\alpha_k)}{14b_k\theta_{min}^2}\frac{\log N_k}{N_k}\right)^2+\left(\frac{(\theta_{max}-\theta_{min})(5+2\alpha_k)}{14b_k\theta_{min}^2}\frac{\log N_k}{N_k}+(\bar{\theta}_2^{(t)}-\theta^*_{2k})\frac{1}{2\theta_{min}^2}\right)^2\right\}\\
&\le w_{1k}^2\left(\frac{\log N_k}{N_k}\right)^2+w_{2k}^2\left(\frac{\log N_k}{N_k}\right)^2+2w_{2k}\frac{\log N_k}{N_k}(\bar{\theta}_2^{(t)}-\theta^*_{2k})\frac{1}{2\theta_{min}^2}+\left\|\bar{\theta}_2^{(t)}-\bar{\theta}^*_{2k}\right\|_2^2\frac{1}{4\theta_{min}^4}\\
&\le (w_{1k}^2+w_{2k}^2)\left(\frac{\log N_k}{N_k}\right)^2+2w_{2k}\frac{\log N_k}{N_k}(\theta_{max}-\theta_{min})\frac{1}{2\theta_{min}^2}\\
&+\left(\frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+1}+2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+2\sqrt{2}\beta_1(\theta_{max}-\theta_{min})CM_k^{\epsilon_k-\frac{1}{2}}\right)\frac{1}{4\theta_{min}^4},
\end{aligned}
$$

where

$$
\begin{aligned}
w_{1k}&=\frac{(\theta_{max}-\theta_{min})(33+16\alpha_k)}{14b_k\theta_{min}^2},\\
w_{2k}&=\frac{(\theta_{max}-\theta_{min})(5+2\alpha_k)}{14b_k\theta_{min}^2}.
\end{aligned}
$$

By Lemma 6, with probability at least $1-C_{\theta,k}(t+1)(\log(M_k^{\epsilon-\frac{1}{2}}))^4\exp\{-c_{\theta,k}M_k^{2\epsilon_k}\}$,

$$
\begin{aligned}
&\left\|\nabla L_k(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2^2\le\left(C_{\theta,k}M_k^{\epsilon_k-\frac{1}{2}}\right)^2+\left\|\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2^2+2\left\|\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2\left(C_{\theta,k}M_k^{\epsilon_k-\frac{1}{2}}\right)\\
&\le C_{\theta,k}^2M_k^{2\epsilon_k-1}+(w_{1k}^2+w_{2k}^2)\left(\frac{\log N_k}{N_k}\right)^2+2w_{2k}\frac{\log N_k}{N_k}(\theta_{max}-\theta_{min})\frac{1}{2\theta_{min}^2}\\
&+\left(\frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+1}+2\beta_1\max_k C_{3k}(\alpha_k,b_k)\frac{\log M_k}{M_k}+2\sqrt{2}\beta_1(\theta_{max}-\theta_{min})CM_k^{\epsilon_k-\frac{1}{2}}\right)\frac{1}{4\theta_{min}^4}\\
&+2\left\|\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2\left(C_{\theta,k}M_k^{\epsilon_k-\frac{1}{2}}\right).
\end{aligned}
$$

Therefore,

$$\left\|\nabla L(\bar{\boldsymbol{\theta}})\right\|_2^2 = \left\|\sum_{k=1}^{K} p_k \nabla L_k(\bar{\boldsymbol{\theta}}; D_k)\right\|_2^2 \leq \sum_{k=1}^{K} p_k \left\|\nabla L_k(\bar{\boldsymbol{\theta}}; D_k)\right\|_2^2 \leq \max_k \left\|\nabla L_k(\bar{\boldsymbol{\theta}}; D_k)\right\|_2^2$$
$$\leq \max_k \left(\frac{2\beta_1^2 \left(8(E-1)^2+2\right) G^2}{4\theta_{min}^4 (t+1)} + \mathcal{O}\left(\frac{\log M_k}{M_k} + M_k^{\epsilon_k - \frac{1}{2}} + \frac{\log N_k}{N_k} \right) \right).$$

Under the partial device participation scenario, we have

$$\left\|\nabla L(\bar{\boldsymbol{\theta}})\right\|_2^2 \leq \max_k \left(\frac{2\beta_1^2 \left(\frac{1}{|\mathcal{S}_c|} 4E^2 + 8(E-1)^2+2\right) G^2}{4\theta_{min}^4 (t+1)} + \mathcal{O}\left(\frac{\log M_k}{M_k} + M_k^{\epsilon_k - \frac{1}{2}} + \frac{\log N_k}{N_k} \right) \right).$$

□

4.6 Missing Proof in Theorem 3

Following the same strategy in Theorem 4, we can show that

$$\begin{aligned}
&\left\|\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2^2 \\
&= \left(\left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_1^2 + \left[\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right]_2^2 \right) \\
&= \left(\frac{1}{2N_k} (\bar{\theta}_1^{(t)} - \theta_1^*) \sum_{j=1}^{N_k} \frac{\lambda_{1j}^2}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} + \frac{1}{2N_k} (\bar{\theta}_2^{(t)} - \theta_2^*) \sum_{j=1}^{N_k} \frac{\lambda_{2j}\lambda_{1j}}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} \right)^2 \\
&\quad + \left(\frac{1}{2N_k} (\bar{\theta}_1^{(t)} - \theta_1^*) \sum_{j=1}^{N_k} \frac{\lambda_{1j}\lambda_{2j}}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} + \frac{1}{2N_k} (\bar{\theta}_2^{(t)} - \theta_2^*) \sum_{j=1}^{N_k} \frac{\lambda_{2j}^2}{\left(\bar{\theta}_1^{(t)} \lambda_{1j} + \bar{\theta}_2^{(t)} \lambda_{2j}\right)^2} \right)^2.
\end{aligned}$$

By Lemma 3, we have

$$
\begin{aligned}
&\left\|\nabla L_k^*(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2^2 \\
&\leq \Bigg(\frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_1^*)N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}^2(4b_k+3)}{\theta_{min}^2(8b_k^2-8b_k-3)}\right) \\
&\quad + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_2^*)N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2-6b_k-3)}\right)\Bigg)^2 \\
&\quad + \left(\frac{1}{2N_k}(\bar{\theta}_1^{(t)} - \theta_1^*)N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2-6b_k-3)}\right) + \frac{1}{2N_k}(\bar{\theta}_2^{(t)} - \theta_2^*)\frac{N_k}{\theta_{min}^2}\right)^2 \\
&\leq \Bigg(\frac{1}{2}(\theta_{max} - \theta_{min})N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1}\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}^2(4b_k+3)}{\theta_{min}^2(8b_k^2-8b_k-3)}\right) \\
&\quad + \frac{1}{2}(\theta_{max} - \theta_{min})N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1}\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2-6b_k-3)}\right)\Bigg)^2 \\
&\quad + \left(\frac{1}{2}(\theta_{max} - \theta_{min})N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1}\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2-6b_k-3)}\right) + \frac{(\bar{\theta}_2^{(t)} - \theta_2^*)}{2\theta_{min}^2}\right)^2 \\
&\leq a_{mat,1}N_k^{\frac{2(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-2} + \left(a_{mat,2}N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1} + \frac{(\bar{\theta}_2^{(t)} - \theta_2^*)}{2\theta_{min}^2}\right)^2 \\
&\leq a_{mat,1}N_k^{\frac{2(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-2} + a_{mat,2}^2N_k^{\frac{2(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-2} + 2a_{mat,2}N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1}\frac{(\bar{\theta}_2^{(t)} - \theta_2^*)}{2\theta_{min}^2} + \frac{(\bar{\theta}_2^{(t)} - \theta_2^*)^2}{4\theta_{min}^4} \\
&\leq a_{mat,1}N_k^{\frac{2(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-2} + a_{mat,2}^2N_k^{\frac{2(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-2} + 2a_{mat,2}N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1}\frac{(\bar{\theta}_2^{(t)} - \theta_2^*)}{2\theta_{min}^2} \\
&\quad + \frac{1}{4\theta_{min}^4}\Bigg(\frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{t+1} + 2\beta_1\max_k C_{4k}M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1} + 2\beta_1\Bigg(\sum_{k=1}^{K}\sqrt{2}p_k(\theta_{max} - \theta_{min})C_{\boldsymbol{\theta}}M_k^{\epsilon_k-\frac{1}{2}}\Bigg)\Bigg)
\end{aligned}
$$

where $a_{mat,1} = \left(\frac{1}{2}(\theta_{max} - \theta_{min})\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}^2(4b_k+3)}{\theta_{min}^2(8b_k^2-8b_k-3)}\right) + \frac{1}{2}(\theta_{max} - \theta_{min})\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2-6b_k-3)}\right)\right)^2$ and $a_{mat,2} = \frac{1}{2}(\theta_{max} - \theta_{min})\left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2-6b_k-3)}\right)$

By Lemma 6 and Lemma 5, with probability at least $1-\max_k\{C_{\boldsymbol{\theta}}(t+1)(\log(M_k^{\epsilon_k-\frac{1}{2}}))^4\exp\{-c_{\boldsymbol{\theta}}M_k^{2\epsilon_k}\}\}$

$$
\begin{aligned}
&\left\|\nabla L_k(\bar{\boldsymbol{\theta}}^{(t)})\right\|_2^2 \\
&\leq \frac{2\beta_1^2\left(8(E-1)^2+2\right)G^2}{4\theta_{min}^4(t+1)} + \mathcal{O}\left(M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1} + \sum_{k=1}^{K}p_kM_k^{\epsilon_k-\frac{1}{2}} + N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}-1}\right).
\end{aligned}
$$

For the partial device participation scenario, the proof is similar.

5 Proof of Lemmas

5.1 Proof of Lemma 3

Remember that the eigenvalues in each device k are different. For the sake of neatness, we omit the subscript k in the eigenvalues. Let $r_k = j^{\frac{4b_k}{4b_k+3}}$ and $\delta_k = \frac{1}{N_k^{\alpha_k+1}}$, where $0 < \alpha_k < \frac{8b_k^2-12b_k-6}{4b_k+3}$, then, by Lemma 1, with probability at least $1-\delta_k$, we have

$$\begin{aligned} C(r_k, N_k) &< r_k\sqrt{\frac{r_k(r_k+1)}{N_k\delta_k\lambda^*_{r_k}}} = j^{\frac{4b_k}{4b_k+3}}\sqrt{\frac{j^{\frac{4b_k}{4b_k+3}}(j^{\frac{4b_k}{4b_k+3}}+1)}{C_k j^{\frac{-8b_k^2}{4b_k+3}}}N_k^{\frac{\alpha}{2}}} \\ &= N_k^{\frac{\alpha}{2}} j^{\frac{4b_k^2+6b_k}{4b_k+3}}\sqrt{\frac{j^{\frac{4b_k}{4b_k+3}}(j^{-\frac{4b_k}{4b_k+3}}+1)}{C_k}} \le N_k^{\frac{\alpha}{2}} j^{\frac{4b_k^2+8b_k}{4b_k+3}}\sqrt{\frac{2}{C_k}} \end{aligned}$$

and

$$\begin{aligned} H(r_k, N_k) &< \frac{C_k}{2b_k-1}r^{-(2b_k-1)} + \sqrt{\frac{2C_k}{2b_k-1}}r^{-(b_k-1/2)}N_k^{\alpha/2} \\ &\le \left(\frac{C_k}{2b_k-1} + \sqrt{\frac{2C_k}{2b_k-1}}\right) j^{-\frac{2b_k(2b_k-1)}{4b_k+3}} N_k^{\alpha/2}. \end{aligned}$$

Therefore, by Lemma 1, we obtain

$$\begin{aligned} \frac{\lambda_j(\boldsymbol{K}_{f,N_k})}{N_k} &\le \lambda^*_j + \lambda^*_j N_k^{\frac{\alpha}{2}} j^{\frac{4b_k^2+8b_k}{4b_k+3}}\sqrt{\frac{2}{C_k}} + \left(\frac{C_k}{2b_k-1} + \sqrt{\frac{2C_k}{2b_k-1}}\right) j^{-\frac{2b_k(2b_k-1)}{4b_k+3}} N_k^{\alpha/2} \\ &= C_k j^{-2b_k} + C_k j^{-2b_k} N_k^{\frac{\alpha}{2}} j^{\frac{4b_k^2+8b_k}{4b_k+3}}\sqrt{\frac{2}{C_k}} + \left(\frac{C_k}{2b_k-1} + \sqrt{\frac{2C_k}{2b_k-1}}\right) j^{-\frac{2b_k(2b_k-1)}{4b_k+3}} N_k^{\alpha/2}. \end{aligned}$$

This implies

$$\begin{aligned} &\lambda_j(\boldsymbol{K}_{f,N_k}) \le C_k j^{-2b_k}\left(N_k + N_k^{1+\frac{\alpha}{2}} j^{\frac{4b_k^2+8b_k}{4b_k+3}}\sqrt{\frac{2}{C_k}}\right) + \left(\frac{C_k}{2b_k-1} + \sqrt{\frac{2C_k}{2b_k-1}}\right) j^{-\frac{2b_k(2b_k-1)}{4b_k+3}} N_k^{1+\alpha/2} \\ &\le \left(2\sqrt{2C_k} + \frac{C_k}{2b_k-1} + \sqrt{\frac{2C_k}{2b_k-1}}\right) j^{-\frac{2b_k(2b_k-1)}{4b_k+3}} N_k^{1+\alpha/2}, \end{aligned}$$

where probability at least $1 - \frac{1}{N_k^{\alpha_k+1}}$. Let $C_{mat,k} = \left(2\sqrt{2C_k} + \frac{C_k}{2b_k-1} + \sqrt{\frac{2C_k}{2b_k-1}}\right)$. Therefore, we have

$$\sum_{j=1}^{N_k} \frac{\lambda_{1j}^2}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} \leq \frac{L_{mat,k}}{\theta_{min}^2} + \frac{C_{mat,k}^2}{\theta_{min}^2} \sum_{j=L_{mat,k}}^{\infty} j^{-\frac{4b_k(2b_k-1)}{4b_k+3}} N_k^{2+\alpha}$$

for any $0 < L_{mat,k} \leq N_k$. Let $L_{mat,k} = N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}$, then we obtain

$$\sum_{j=1}^{N_k} \frac{\lambda_{1j}^2}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} \leq N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}} \left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}^2(4b_k+3)}{\theta_{min}^2(8b_k^2 - 8b_k - 3)}\right).$$

Similarly, we have

$$\sum_{j=1}^{N_k} \frac{\lambda_{1j}\lambda_{2j}}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} \leq N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}} \left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2 - 6b_k - 3)}\right).$$

Additionally, we can show that

$$\sum_{j=1}^{N_k} \frac{\lambda_{2j}^2}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} \geq \frac{|\{j : \theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j} \leq 2\theta_{max}\}|}{4\theta_{max}^2}.$$

The fact that $j : \theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j} \leq 2\theta_{max}$ implies

$$\begin{aligned} &C_{mat,k}\theta_{max} j^{-\frac{2b_k(2b_k-1)}{4b_k+3}} N_k^{1+\alpha/2} \leq \theta_{max} \\ &\Rightarrow j \geq C_{mat,k}^{\frac{4b_k+3}{2b_k(2b_k-1)}} N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}} \geq C_{mat,k} N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}} \end{aligned}$$

since $b_k \geq \frac{\sqrt{21}+3}{4}$. Therefore,

$$\frac{N_k - C_{mat,k} N_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}}{4\theta_{max}} \leq \sum_{j=1}^{N_k} \frac{\lambda_{2j}^2}{\left(\theta_{1k}^{(t)}\lambda_{1j} + \theta_{2k}^{(t)}\lambda_{2j}\right)^2} \leq \frac{N_k}{\theta_{min}^2}$$

where the upper bound is trivially true.

5.2 Proof of Lemma 4

Proof. For device k, denote by $\boldsymbol{\theta}_k^{(t)} = (\theta_{1k}^{(t)}, \theta_{2k}^{(t)})$ the model parameter at iteration t. Let $\lambda_{1jk}^{(t)}$ be the j^{th} largest eigenvalue of $\boldsymbol{K}_{f,\xi_k^{(t)}}$ and $\lambda_{2jk}^{(t)} = 1$ be the j^{th} largest eigenvalue of $\boldsymbol{I}_M$. By definition,

$$
\begin{aligned}
&\left[g_k^*(\boldsymbol{\theta}_k^{(t)})\right]_1 \\
&= \frac{1}{2s_1(M)} \operatorname{Tr}\left[\boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})^{-1}\left(\boldsymbol{I}_M - \boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^*)\boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})^{-1}\right)\frac{\partial \boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})}{\partial \theta_{1k}^{(t)}}\right] \\
&= \frac{1}{2s_1(M)}(\theta_{1k}^{(t)} - \theta_{1k}^*)\sum_{j=1}^{M}\frac{\lambda_{1jk}^{(t)2}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} + \frac{1}{2s_1(M)}(\theta_{2k}^{(t)} - \theta_{2k}^*)\sum_{j=1}^{M}\frac{\lambda_{2jk}^{(t)}\lambda_{1jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2}
\end{aligned}
$$

and

$$
\begin{aligned}
&\left[g_k^*(\boldsymbol{\theta}_k^{(t)})\right]_2 \\
&= \frac{1}{2M} \operatorname{Tr}\left[\boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})^{-1}\left(\boldsymbol{I}_M - \boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^*)\boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})^{-1}\right)\frac{\partial \boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})}{\partial \theta_{2k}^{(t)}}\right] \\
&= \frac{1}{2M}(\theta_{1k}^{(t)} - \theta_{1k}^*)\sum_{j=1}^{M}\frac{\lambda_{1jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} + \frac{1}{2s_1(M)}(\theta_{2k}^{(t)} - \theta_{2k}^*)\sum_{j=1}^{M}\frac{\lambda_{2jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2}.
\end{aligned}
$$

Based on those two expressions, we can obtain

$$
\begin{aligned}
&\langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k^*(\boldsymbol{\theta}_k^{(t)})\rangle \\
&= (\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*)^\intercal \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} (\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*)
\end{aligned}
$$

where $A_{11}, A_{12}, A_{21}, A_{22}$ will be clarified shortly. Let $\epsilon_k = \frac{1}{2}$, by Lemma 2, with probability at least $1 - \frac{2}{M^{\alpha_k}}$,

$$\begin{aligned}
A_{11} &:= \frac{1}{2\tau \log M} \sum_{j=1}^{M} \frac{\lambda_{1jk}^{(t)2}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} \geq \frac{1}{2\tau \log M} \frac{\epsilon_k \log M}{8 b_k \theta_{max}^2} = \frac{1}{32 \tau b_k \theta_{max}^2}, \\
A_{12} &:= \frac{1}{2\tau \log M} \sum_{j=1}^{M} \frac{\lambda_{1jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} \leq \frac{1}{2\tau \log M} \frac{5 + 2\alpha_k}{7 b_k \theta_{min}^2} \log M \leq \frac{1}{2\tau b_k \theta_{min}^2}, \\
A_{21} &:= \frac{1}{2M} \sum_{j=1}^{M} \frac{\lambda_{1jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} \leq \frac{1}{2M} \frac{5 + 2\alpha_k}{7 b_k \theta_{min}^2} \log M = \frac{5 + 2\alpha_k}{14 b_k \theta_{min}^2} \frac{\log M}{M} \leq \frac{1}{2 b_k \theta_{min}^2} \frac{\log M}{M}, \\
A_{12} &:= \frac{1}{2M} \sum_{j=1}^{M} \frac{1}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} \geq \frac{1}{2M} \frac{M - C_{1k}(\alpha_k, b_k) \log M}{4\theta_{max}^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k^*(\boldsymbol{\theta}_k^{(t)}) \rangle \\
&\geq \left(\frac{1}{64 \tau b_k \theta_{max}^2} - \frac{\log M}{64 \theta_{max}^2 b_k M} \right) (\theta_{1k}^{(t)} - \theta_1^*)^2 \\
&\quad + \left(\frac{1}{8\theta_{max}^2} - \frac{4\theta_{max}^2}{\tau b_k \theta_{min}^4} - \frac{C_{1k}(\alpha_k, b_k) \log M}{8 \theta_{max}^2 M} + \frac{4\theta_{max}^2 \log M}{b \theta_{min}^4 M} \right) (\theta_{2k}^{(t)} - \theta_1^*)^2 \\
&= \frac{1}{64 \tau b_k \theta_{max}^2} (\theta_{1k}^{(t)} - \theta_1^*)^2 + \left(\frac{1}{8\theta_{max}^2} - \frac{4\theta_{max}^2}{\tau b_k \theta_{min}^4} \right) (\theta_{2k}^{(t)} - \theta_1^*)^2 \\
&\quad - \frac{\log M}{64 \theta_{max}^2 b_k M} \theta_{max}^2 - \frac{C_{1k}(\alpha_k, b_k) \log M}{8\theta_{max}^2 M} \theta_{max}^2 + \frac{4\theta_{max}^2 \log M}{b_k \theta_{min}^4 M} \theta_{min}^2 \\
&\geq \frac{\gamma_k}{2} \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}_k^* \right\|_2^2 - C_{3k}(\alpha_k, b_k) \frac{\log M}{M},
\end{aligned}$$

where $\gamma_k = \min \left\{ \frac{1}{32 \tau b_k \theta_{max}^2}, \frac{1}{4\theta_{max}^2} - \frac{8\theta_{max}^2}{\tau b_k \theta_{min}^4} \right\} > 0$ and $C_{3k}(\alpha_k, b_k) = \frac{1}{64 b_k} + \frac{C_{1k}(\alpha_k, b_k)}{8} - \frac{4\theta_{max}^2}{b\theta_{min}^2}$. □

5.3 Proof of Lemma 5

By definition, we can show that

$$\begin{aligned}
&\left[g_k^*(\boldsymbol{\theta}_k^{(t)}) \right]_2 \\
&= \frac{1}{2M} \operatorname{Tr} \left[\boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})^{-1} \left(\boldsymbol{I}_M - \boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^*) \boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})^{-1} \right) \frac{\partial \boldsymbol{K}_{\xi_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})}{\partial \theta_{2k}^{(t)}} \right] \\
&= \frac{1}{2M} (\theta_{1k}^{(t)} - \theta_{1k}^*) \sum_{j=1}^{M} \frac{\lambda_{1jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} + \frac{1}{2M} (\theta_{2k}^{(t)} - \theta_{2k}^*) \sum_{j=1}^{M} \frac{\lambda_{2jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\left[g_k^*(\boldsymbol{\theta}_k^{(t)})\right]_2 (\theta_{2k}^{(t)} - \theta_{2k}^*) \\
&= \frac{1}{2M}(\theta_{1k}^{(t)} - \theta_{1k}^*)(\theta_{2k}^{(t)} - \theta_{2k}^*) \sum_{j=1}^{M} \frac{\lambda_{1jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} + \frac{1}{2M}(\theta_{2k}^{(t)} - \theta_{2k}^*)^2 \sum_{j=1}^{M} \frac{\lambda_{2jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} \\
&\geq \frac{1}{2M}(\theta_{2k}^{(t)} - \theta_{2k}^*)^2 \sum_{j=1}^{M} \frac{1}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} - \frac{1}{2M}(\theta_{max} - \theta_{min})^2 \sum_{j=1}^{M} \frac{\lambda_{1jk}^{(t)}}{(\theta_{1k}^{(t)}\lambda_{1jk}^{(t)} + \theta_{2k}^{(t)}\lambda_{2jk}^{(t)})^2} \\
&\geq \frac{1}{2M_k}(\theta_{2k}^{(t)} - \theta_{2k}^*)^2 \frac{M_k - C_{mat,k} M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}}{4\theta_{max}} - \frac{1}{2M_k}(\theta_{max} - \theta_{min})^2 M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}} \left(\frac{1}{\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{\theta_{min}^2(4b_k^2 - 6b_k - 3)}\right)
\end{aligned}
$$

with probability at least $1 - \frac{1}{M_k^{1+\alpha_k}}$. Therefore,

$$
\left[g_k^*(\boldsymbol{\theta}_k^{(t)})\right]_2 (\theta_{2k}^{(t)} - \theta_{2k}^*) \geq \frac{\gamma_k}{2}(\theta_{2k}^{(t)} - \theta_{2k}^*)^2 - (\theta_{max} - \theta_{min})^2 M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)} - 1} \left(\frac{1}{2\theta_{min}^2} + \frac{C_{mat,k}(4b_k+3)}{2\theta_{min}^2(4b_k^2 - 6b_k - 3)}\right),
$$

where we slightly abuse the notation and define $\gamma_k := \frac{1}{2M_k} \frac{M_k - C_{mat,k} M_k^{\frac{(2+\alpha_k)(4b_k+3)}{4b_k(2b_k-1)}}}{4\theta_{max}}$. Here note that this γ_k is different from the γ_k in the Lemmas/Theorems involved with RBF kernels.

References

Braun, M. L. (2006). Accurate error bounds for the eigenvalues of the kernel matrix. *The Journal of Machine Learning Research*, 7:2303–2328.

Chen, H., Zheng, L., Al Kontar, R., and Raskutti, G. (2020). Stochastic gradient descent in correlated settings: A study on gaussian processes. *Advances in Neural Information Processing Systems*, 33.

Cox, D. D., Park, J.-S., and Singer, C. E. (2001). A statistical method for tuning a computer code to a data base. *Computational statistics & data analysis*, 37(1):77–92.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963.

Cutajar, K., Pullin, M., Damianou, A., Lawrence, N., and González, J. (2019). Deep gaussian processes for multi-fidelity modeling. *arXiv preprint arXiv:1903.07320*.

Gramacy, R. B. and Lian, H. (2012). Gaussian process single-index models as emulators for computer experiments. *Technometrics*, 54(1):30–41.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

Liu, K., Gebraeel, N. Z., and Shi, J. (2013). A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis. *IEEE Transactions on Automation Science and Engineering*, 10(3):652–664.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Moon, H., Dean, A. M., and Santner, T. J. (2012). Two-stage sensitivity-based group screening in computer experiments. *Technometrics*, 54(4):376–387.

Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N. D., and Karniadakis, G. E. (2017). Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751.

Song, C. and Liu, K. (2018). Statistical degradation modeling and prognostics of multiple sensor signals via data fusion: A composite health index approach. *IISE Transactions*, 50(10):853–867.

Xiong, S., Qian, P. Z., and Wu, C. J. (2013). Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics*, 55(1):37–46.

Yan, H., Liu, K., Zhang, X., and Shi, J. (2016). Multiple sensor data fusion for degradation modeling and prognostics under multiple operational conditions. *IEEE Transactions on Reliability*, 65(3):1416–1426.