

PFL

November 27, 2024

1 Introduction

xxx (Collins et al., 2021)

Algorithm 1

Input: Participation rate r , step size η , number of local updates for the head τ_w , for the shortcut τ_s and for the representation τ_b , number of communication rounds T .

```
1: Initialize  $\mathbf{B}^0, \mathbf{w}_1^0, \dots, \mathbf{w}_n^0, \mathbf{s}_1^0, \dots, \mathbf{s}_n^0$ 
2: for  $t = 0, 1, 2, \dots, T - 1$  do
3:   Server receives a batch of clients  $\mathcal{I}^t$  of size  $rn$ 
4:   Server sends current representation  $\phi^t$  to clients in  $\mathcal{I}^t$ 
5:   for each client  $i$  in  $\mathcal{I}^t$  do
6:     Client  $i$  initializes  $\mathbf{w}_i^{t,0} \leftarrow \mathbf{w}_i^{t-1, \tau_h}$ 
7:     Client updates its head for  $\tau_h$  steps:
8:     for  $\tau = 1$  to  $\tau_w$  do
9:        $\mathbf{w}_i^{t,\tau} \leftarrow \text{GRD} \left( f_i \left( \mathbf{w}_i^{t,\tau-1}, \mathbf{B}^{t-1}, \mathbf{s}_i^{t-1, \tau_s} \right), \mathbf{w}_i^{t,\tau-1}, \eta \right)$ 
10:    end for
11:    Client  $i$  initializes  $\mathbf{s}_i^{t,0} \leftarrow \mathbf{s}_i^{t-1, \tau_s}$ 
12:    Client  $i$  updates its shortcut for  $\tau_s$  steps:
13:    for  $\tau = 1$  to  $\tau_s$  do
14:       $\mathbf{s}_i^{t,\tau} \leftarrow \text{GRD} \left( f_i \left( \mathbf{w}_i^{t-1}, \mathbf{B}^{t-1}, \mathbf{s}_i^{t,\tau-1} \right), \mathbf{s}_i^{t,\tau-1}, \eta \right)$ 
15:    end for
16:    Client  $i$  initializes  $\mathbf{B}_i^{t,0} \leftarrow \mathbf{B}^{t-1}$ 
17:    Client  $i$  updates its representation for  $\tau_b$  steps:
18:    for  $\tau = 1$  to  $\tau_b$  do
19:       $\mathbf{B}_i^{t,\tau} \leftarrow \text{GRD} \left( f_i \left( \mathbf{w}_i^{t,\tau_w}, \mathbf{B}_i^{t,\tau-1}, \mathbf{s}_i^{t,\tau_s} \right), \mathbf{B}_i^{t,\tau-1}, \eta \right)$ 
20:    end for
21:    Client  $i$  sends updated representation  $\mathbf{B}_i^{t,\tau_b}$  to server
22:  end for
23:  for each client  $j$  not in  $\mathcal{I}^t$  do
24:    Set  $\mathbf{w}_i^{t,\tau_w} \leftarrow \mathbf{w}_i^{t-1, \tau_w}$  and  $\mathbf{s}_i^{t,\tau_s} \leftarrow \mathbf{s}_i^{t-1, \tau_s}$ 
25:  end for
26:  Server computes new representation:  $\mathbf{B}^t = \frac{1}{rn} \sum_{i \in \mathcal{I}^t} \mathbf{B}_i^{t,\tau_b}$ 
27: end for
```

1.1 Preliminaries

First, we establish the notations that will be used throughout our proof. Let $\mathbf{S} := [\mathbf{s}_1, \dots, \mathbf{s}_{rn}] \in \mathbb{R}^{d \times rn}$ represent the personalized layers, and let $\mathbf{W} := [\mathbf{w}_1, \dots, \mathbf{w}_{rn}] \in \mathbb{R}^{k \times rn}$ denote the personalized heads, which follow the global representation \mathbf{B} . Since our algorithm updates \mathbf{w}_i and \mathbf{s}_i for each client i simultaneously, we define $\mathbf{h}_i^\top := [\mathbf{w}_i^\top, \mathbf{s}_i^\top]$ and $\mathbf{H} := [\mathbf{h}_1, \dots, \mathbf{h}_{rn}] \in \mathbb{R}^{(k+d) \times rn}$.

...
The global objective can be rewritten as

$$\min_{\mathbf{B} \in \mathbb{R}^{d \times k}, \mathbf{W} \in \mathbb{R}^{k \times rn}, \mathbf{S} \in \mathbb{R}^{d \times rn}} \left\{ F(\hat{\mathbf{B}}, \mathbf{W}, \mathbf{S}) := \frac{1}{2rnm} \mathbb{E}_{\mathcal{A}, \mathcal{I}} \left\| \mathbf{Y} - \mathcal{A}(\mathbf{W}_{\mathcal{I}}^\top \hat{\mathbf{B}}^\top + \mathbf{S}_{\mathcal{I}}^\top) \right\|_2^2 \right\}, \quad (1)$$

where $\mathbf{Y} = \mathcal{A}(\mathbf{W}_{\mathcal{I}}^{*\top} \hat{\mathbf{B}}^{*\top} + \mathbf{S}_{\mathcal{I}}^{*\top}) \in \mathbb{R}^{rnm}$. Then we give the update rules of our algorithm:

$$\mathbf{W}^{t+1} = \arg \min_{\mathbf{W} \in \mathbb{R}^{k \times rn}} \frac{1}{2rnm} \left\| \mathcal{A}^t \left(\mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} - \mathbf{W}^\top \hat{\mathbf{B}}^{t\top} + \mathbf{S}^{*\top} - \mathbf{S}^{t\top} \right) \right\|_2^2, \quad (2)$$

$$\mathbf{S}^{t+1} = \arg \min_{\mathbf{S} \in \mathbb{R}^{d \times rn}} \frac{1}{2rnm} \left\| \mathcal{A}^t \left(\mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} - \mathbf{W}^{t\top} \hat{\mathbf{B}}^{t\top} + \mathbf{S}^{*\top} - \mathbf{S}^\top \right) \right\|_2^2, \quad (3)$$

$$\bar{\mathbf{B}} = \hat{\mathbf{B}}^t - \frac{\eta}{rnm} \left((\mathcal{A}^t)^\dagger \mathcal{A}^t (\mathbf{W}^{t+1\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} + \mathbf{S}^{t+1\top} - \mathbf{S}^{*\top}) \right)^\top \mathbf{W}_{\mathcal{I}^t}^{t+1\top}, \quad (4)$$

$$\hat{\mathbf{B}}^{t+1}, \mathbf{R}^{t+1} = \text{QR}(\bar{\mathbf{B}}^t). \quad (5)$$

1.2 Auxiliary Lemmas

We first consider the update for \mathbf{W} . According to the update rule of (2), \mathbf{W}^{t+1} minimizes the function of $\tilde{F}(\hat{\mathbf{B}}^t, \mathbf{W}, \mathbf{S}^t) := \frac{1}{2rnm} \left\| \mathcal{A} \left(\mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} - \mathbf{W}^\top \hat{\mathbf{B}}^{t\top} + \mathbf{S}^{*\top} - \mathbf{S}^{t\top} \right) \right\|_2^2$.

Let \mathcal{W}_p^{t+1} be the p -th column of $\mathbf{W}^{t+1\top}$, \mathcal{W}_p^* denote the p -th column of $\mathbf{W}^{*\top}$, \mathcal{S}_l^t denote the l -th column of $\mathbf{S}^{t\top}$, \mathcal{S}_l^* denote the l -th column of $\mathbf{S}^{*\top}$ and $\hat{\mathbf{b}}_p^t$ be the p -th column of $\hat{\mathbf{B}}^t$, then for any $p \in [k]$, $l \in [d]$, we have

$$\begin{aligned} \mathbf{0} &= \nabla_{\mathcal{W}_p} \tilde{F}(\hat{\mathbf{B}}^t, \mathbf{W}^{t+1}, \mathbf{S}^t) \\ &= \frac{1}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^m \left(\left\langle \mathbf{A}_{i,j}, \mathbf{W}^{t+1\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} + \mathbf{S}^{t\top} - \mathbf{S}^{*\top} \right\rangle \right) \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \\ &= \frac{1}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^m \left(\left\langle \mathbf{A}_{i,j}, \mathbf{W}^{t+1\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} \right\rangle + \left\langle \mathbf{A}_{i,j}, \mathbf{S}^{t\top} - \mathbf{S}^{*\top} \right\rangle \right) \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \\ &= \frac{1}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^m \left(\sum_{q=1}^k \hat{\mathbf{b}}_q^{t\top} \mathbf{A}_{i,j}^\top \mathcal{W}_q^{t+1} - \sum_{q=1}^k \hat{\mathbf{b}}_q^{*\top} \mathbf{A}_{i,j}^\top \mathcal{W}_q^* + \sum_{l=1}^d \mathbf{e}_l^\top \mathbf{A}_{i,j}^\top \mathcal{S}_l^t - \sum_{l=1}^d \mathbf{e}_l^\top \mathbf{A}_{i,j}^\top \mathcal{S}_l^* \right) \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t, \end{aligned} \quad (6)$$

which means

$$\begin{aligned} & \frac{1}{m} \sum_{q=1}^k \left(\sum_{i=1}^{rn} \sum_{j=1}^m \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \hat{\mathbf{b}}_q^{t\top} \mathbf{A}_{i,j}^\top \right) \mathcal{W}_q^{t+1} \\ &= \frac{1}{m} \sum_{q=1}^k \left(\sum_{i=1}^{rn} \sum_{j=1}^m \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \hat{\mathbf{b}}_q^{*\top} \mathbf{A}_{i,j}^\top \right) \mathcal{W}_q^* + \frac{1}{m} \sum_{l=1}^d \left(\sum_{i=1}^{rn} \sum_{j=1}^m \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \mathbf{e}_l^\top \mathbf{A}_{i,j}^\top \right) (\mathcal{S}_l^t - \mathcal{S}_l^*). \end{aligned} \quad (7)$$

Then, define $\mathbf{G}_{pq} := \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^m \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \hat{\mathbf{b}}_q^{t\top} \mathbf{A}_{i,j}^\top$, $\mathbf{C}_{pq} := \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^m \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \hat{\mathbf{b}}_q^{*\top} \mathbf{A}_{i,j}^\top$ and $\mathbf{D}_{pq} := \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^m \langle \hat{\mathbf{b}}_p^t, \hat{\mathbf{b}}_q^* \rangle \mathbf{I}_{rn}$, for all $p, q \in [k]$, and define $\mathbf{E}_{pl} := \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^m \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \mathbf{e}_l^\top \mathbf{A}_{i,j}^\top$, for all $p \in [k], l \in [d]$. Further, we define block matrices $\mathbf{G}, \mathbf{C}, \mathbf{D} \in \mathbb{R}^{rnk \times rnk}$ and $\mathbf{E} \in \mathbb{R}^{rnk \times rnd}$, which are formed by $\mathbf{G}_{pq}, \mathbf{C}_{pq}, \mathbf{D}_{pq}$ and \mathbf{E}_{pl} , respectively. In detail, take \mathbf{G} and \mathbf{E} for example,

$$\mathbf{G} := \begin{bmatrix} \mathbf{G}_{11} & \cdots & \mathbf{G}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{k1} & \cdots & \mathbf{G}_{kk} \end{bmatrix}, \mathbf{E} := \begin{bmatrix} \mathbf{E}_{11} & \cdots & \mathbf{E}_{1d} \\ \vdots & \ddots & \vdots \\ \mathbf{E}_{k1} & \cdots & \mathbf{E}_{kd} \end{bmatrix}. \quad (8)$$

Then we define $\widetilde{\mathcal{W}}^{t+1} := \text{vec}(\mathbf{W}^{t+1\top}) \in \mathbb{R}^{rnk}$, $\widetilde{\mathcal{W}}^* := \text{vec}(\mathbf{W}^{*\top}) \in \mathbb{R}^{rnk}$, $\widetilde{\mathcal{S}}^t := \text{vec}(\mathbf{S}^{t\top}) \in \mathbb{R}^{rnd}$ and $\widetilde{\mathcal{S}}^* := \text{vec}(\mathbf{S}^{*\top}) \in \mathbb{R}^{rnd}$. From (7) we reach,

$$\begin{aligned} \widetilde{\mathcal{W}}^{t+1} &= \mathbf{G}^{-1} \mathbf{C} \widetilde{\mathcal{W}}^* + \mathbf{G}^{-1} \mathbf{E} (\widetilde{\mathcal{S}}^t - \widetilde{\mathcal{S}}^*) \\ &= \mathbf{D} \widetilde{\mathcal{W}}^* - \mathbf{G}^{-1} (\mathbf{G} \mathbf{D} - \mathbf{C}) \widetilde{\mathcal{W}}^* + \mathbf{G}^{-1} \mathbf{E} (\widetilde{\mathcal{S}}^t - \widetilde{\mathcal{S}}^*), \end{aligned} \quad (9)$$

where \mathbf{G} is invertible will be proved in the following lemma. Here, we consider \mathbf{G}_{pq} ,

$$\begin{aligned} \mathbf{G}_{pq} &= \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^m \mathbf{A}_{i,j} \hat{\mathbf{b}}_p \hat{\mathbf{b}}_q^\top \mathbf{A}_{i,j}^\top \\ &= \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^m \mathbf{e}_i (\mathbf{x}_i^j)^\top \hat{\mathbf{b}}_p \hat{\mathbf{b}}_q^\top \mathbf{x}_i^j \mathbf{e}_i^\top, \end{aligned} \quad (10)$$

meaning that \mathbf{G}_{pq} is diagonal with diagonal entries

$$(\mathbf{G}_{pq})_{ii} = \frac{1}{m} \sum_{j=1}^m (\mathbf{x}_i^j)^\top \hat{\mathbf{b}}_p \hat{\mathbf{b}}_q^\top \mathbf{x}_i^j = \hat{\mathbf{b}}_p^\top \left(\frac{1}{m} \sum_{j=1}^m \mathbf{x}_i^j (\mathbf{x}_i^j)^\top \right) \hat{\mathbf{b}}_q. \quad (11)$$

Define $\mathbf{\Pi}^i := \frac{1}{m} \sum_{j=1}^m \mathbf{x}_i^j (\mathbf{x}_i^j)^\top$ for all $i \in [rn]$, then \mathbf{C}_{pq} is diagonal with entries $(\mathbf{C}_{pq})_{ii} = \hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \hat{\mathbf{b}}_q^*$, and \mathbf{E}_{pl} is diagonal with entries $(\mathbf{E}_{pl})_{ii} = \hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \mathbf{e}_l$. Note that $\mathbf{D}_{pq} = \langle \hat{\mathbf{b}}_p, \hat{\mathbf{b}}_q^* \rangle \mathbf{I}_{rn}$ is also diagonal, then we define

$$\mathbf{G}^i := \left[\hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \hat{\mathbf{b}}_q \right]_{1 \leq p, q \leq k+d} = \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}, \quad \mathbf{C}^i := \left[\hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \hat{\mathbf{b}}_q^* \right]_{1 \leq p, q \leq k+d} = \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}^*, \quad (12)$$

$$\mathbf{D}^i := \left[\langle \hat{\mathbf{b}}_p, \hat{\mathbf{b}}_q^* \rangle \right]_{1 \leq p, q \leq k+d} = \hat{\mathbf{B}}^\top \hat{\mathbf{B}}^*, \quad \mathbf{E}^i := \left[\hat{\mathbf{b}}_p^\top \boldsymbol{\Pi}^i \mathbf{e}_l \right]_{1 \leq p \leq k, 1 \leq l \leq d} = \hat{\mathbf{B}}^\top \boldsymbol{\Pi}^i, \quad (13)$$

where \mathbf{G}^i , \mathbf{C}^i and \mathbf{D}^i are the $k \times k$ matrices that formed by taking the i -th diagonal entry of each block \mathbf{G}_{pq} , \mathbf{C}_{pq} and \mathbf{D}_{pq} , respectively. Similarly, \mathbf{E}^i is the $k \times d$ matrix that formed by taking the i -th diagonal entry of each block \mathbf{E}_{pl} . Then we can decouple the term of $\mathbf{G}^{-1}(\mathbf{GD} - \mathbf{C})\widetilde{\mathcal{W}}^*$ in (9) into i vectors, defined as

$$\mathbf{f}_i := (\mathbf{G}^i)^{-1} (\mathbf{G}^i \mathbf{D}^i - \mathbf{C}^i) \widetilde{\mathcal{W}}_i^*, \quad (14)$$

where $\widetilde{\mathcal{W}}_i^* \in \mathbb{R}^k$ is the vector formed by taking the $((p-1)rn+i)$ -th elements of $\widetilde{\mathcal{W}}^*$ for $p = 1, \dots, k$. Similarly, we can decouple $\mathbf{G}^{-1}\mathbf{E}(\tilde{\mathcal{S}}^t - \tilde{\mathcal{S}}^*)$ into i vectors, defined as

$$\mathbf{h}_i = (\mathbf{G}^i)^{-1} \mathbf{E}^i (\tilde{\mathcal{S}}_i^t - \tilde{\mathcal{S}}_i^*), \quad (15)$$

where $\tilde{\mathcal{S}}_i^t \in \mathbb{R}^d$ and $\tilde{\mathcal{S}}_i^* \in \mathbb{R}^d$ are vectors formed by taking the $((l-1)rn+i)$ -th elements of $\tilde{\mathcal{S}}^t$ and $\tilde{\mathcal{S}}^*$, respectively.

Next, we consider the vector $\widetilde{\mathcal{W}}_i^{t+1}$ formed by taking the $((p-1)rn+i)$ -th elements of $\widetilde{\mathcal{W}}^{t+1}$ for $p = 1, \dots, k$, from (9) we have

$$\begin{aligned} \widetilde{\mathcal{W}}_i^{t+1} &= \mathbf{D}^i \widetilde{\mathcal{W}}_i^* - (\mathbf{G}^i)^{-1} (\mathbf{G}^i \mathbf{D}^i - \mathbf{C}^i) \widetilde{\mathcal{W}}_i^* + (\mathbf{G}^i)^{-1} \mathbf{E}^i (\tilde{\mathcal{S}}_i^t - \tilde{\mathcal{S}}_i^*) \\ &= \hat{\mathbf{B}}^\top \hat{\mathbf{B}}^* \widetilde{\mathcal{W}}_i^* - \mathbf{f}_i + \mathbf{h}_i. \end{aligned} \quad (16)$$

Finally, we reach to the update of \mathbf{W}^{t+1} as

$$\mathbf{W}^{t+1} = \hat{\mathbf{B}}^\top \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} + \mathbf{H}, \quad (17)$$

where $\mathbf{F} := [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{rn}]$ and $\mathbf{H} := [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{rn}]$.

Next, we consider the update for \mathbf{S} . According to the update rule of (3), \mathbf{S}^{t+1} minimizes the function of $\tilde{F}(\hat{\mathbf{B}}^t, \mathbf{W}^t, \mathbf{S}) := \frac{1}{2rnm} \left\| \mathbf{A}(\mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} - \mathbf{W}^{t\top} \hat{\mathbf{B}}^{t\top} + \mathbf{S}^{*\top} - \mathbf{S}^\top) \right\|_2^2$. Similarly, we have

$$\begin{aligned} \mathbf{0} &= \nabla_{\mathbf{S}_l} \tilde{F}(\hat{\mathbf{B}}^t, \mathbf{W}^t, \mathbf{S}^{t+1}) \\ &= \frac{1}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^m \left(\left\langle \mathbf{A}_{i,j}, \mathbf{W}^{t\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} + \mathbf{S}^{t+1\top} - \mathbf{S}^{*\top} \right\rangle \right) \mathbf{A}_{i,j} \mathbf{e}_l \\ &= \frac{1}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^m \left(\left\langle \mathbf{A}_{i,j}, \mathbf{W}^{t\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} \right\rangle + \left\langle \mathbf{A}_{i,j}, \mathbf{S}^{t+1\top} - \mathbf{S}^{*\top} \right\rangle \right) \mathbf{A}_{i,j} \mathbf{e}_l \\ &= \frac{1}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^m \left(\sum_{q=1}^k \hat{\mathbf{b}}_q^{t\top} \mathbf{A}_{i,j}^\top \mathcal{W}_q^t - \sum_{q=1}^k \hat{\mathbf{b}}_q^{*\top} \mathbf{A}_{i,j}^\top \mathcal{W}_q^* + \sum_{u=1}^d \mathbf{e}_u^\top \mathbf{A}_{i,j}^\top \mathcal{S}_u^{t+1} - \sum_{u=1}^d \mathbf{e}_u^\top \mathbf{A}_{i,j}^\top \mathcal{S}_u^* \right) \mathbf{A}_{i,j} \mathbf{e}_l, \end{aligned} \quad (18)$$

Lemma 1 Bounding $\|\mathbf{G}^{-1}\|_2$

In order to give bounding on $\|\mathbf{G}^{-1}\|_2$, we need to lower bound $\sigma_{\min}(\mathbf{G})$. For some vector $\mathbf{z} \in \mathbb{R}^{rn(k+d)}$, let $\mathbf{z}^i \in \mathbb{R}^{k+d}$ be the vector formed by taking the $((p-1)rn+i)$ -th elements of \mathbf{z} for $p = 1, \dots, k+d$, then we have

$$\begin{aligned}\sigma_{\min}(\mathbf{G}) &= \min_{\mathbf{z}: \|\mathbf{z}\|_2=1} \mathbf{z}^\top \mathbf{G} \mathbf{z} \\ &= \min_{\mathbf{z}: \|\mathbf{z}\|_2=1} \sum_{i=1}^{rn} (\mathbf{z}^i)^\top \mathbf{G}^i \mathbf{z}^i \\ &\geq \min_{i \in [rn]} \sigma_{\min}(\mathbf{G}^i),\end{aligned}$$

Lemma 2 ...

Let $\mathbf{Q}^{t\top} = \mathbf{W}^{t+1\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} + \mathbf{S}^{t+1\top} - \mathbf{S}^{*\top}$. To bound $\frac{1}{rn} \left\| \left(\frac{1}{m} \mathcal{A}^\top \mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \right\|_2$, we first consider the bound of the columns of \mathbf{Q} . Let $\mathbf{q}_i \in \mathbb{R}^d$ be the i -th column of \mathbf{Q} , for all $i \in [rn]$ we have

$$\begin{aligned}\mathbf{q}_i &= \tilde{\mathbf{B}}^t \tilde{\mathbf{B}}^{t\top} \tilde{\mathbf{B}}^* \mathbf{h}_i^* - \tilde{\mathbf{B}}^t \mathbf{f}_i - \tilde{\mathbf{B}}^* \mathbf{h}_i^* \\ &= \left((1-\alpha)^2 \hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} + \alpha^2 \mathbf{I}_d \right) \tilde{\mathbf{B}}^* \mathbf{h}_i^* - \tilde{\mathbf{B}}^t \mathbf{f}_i - \tilde{\mathbf{B}}^* \mathbf{h}_i^* \\ &= \left((1-\alpha)^2 \hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} + \alpha^2 \mathbf{I}_d \right) \left((1-\alpha) \hat{\mathbf{B}}^* \mathbf{w}_i^* + \alpha \mathbf{s}_i^* \right) - (1-\alpha) \hat{\mathbf{B}}^t \mathbf{f}_{i1} - \alpha \mathbf{f}_{i2} - (1-\alpha) \hat{\mathbf{B}}^* \mathbf{w}_i^* - \alpha \mathbf{s}_i^*\end{aligned}$$

Thus,

1.3 Main Result

Recall that $\mathbf{Q}^{t\top} = \mathbf{W}^{t+1\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} + \mathbf{S}^{t+1\top} - \mathbf{S}^{*\top}$, plugging this into (4), and without losing generality, we drop the subscripts of \mathcal{I}^t and obtain

$$\begin{aligned}\bar{\mathbf{B}}^{t+1} &= \hat{\mathbf{B}}^t - \frac{\eta}{rnm} \left((\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) \right)^\top \mathbf{W}^{t+1\top} \\ &= \hat{\mathbf{B}}^t - \frac{\eta}{rn} \mathbf{Q}^t \mathbf{W}^{t+1\top} - \frac{\eta}{rn} \left(\frac{1}{m} (\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top}.\end{aligned}\tag{19}$$

Since $\bar{\mathbf{B}}^{t+1} = \hat{\mathbf{B}}^{t+1} \mathbf{R}^{t+1}$, we right multiply $(\mathbf{R}^{t+1})^{-1}$ and left multiply $\hat{\mathbf{B}}_\perp^{*\top}$ on both sides to get

$$\hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^{t+1} = \left(\hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t - \frac{\eta}{rn} \hat{\mathbf{B}}_\perp^{*\top} \mathbf{Q}^t \mathbf{W}^{t+1\top} - \frac{\eta}{rn} \hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m} (\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \right) (\mathbf{R}^{t+1})^{-1}.\tag{20}$$

Then we consider the term of $\hat{\mathbf{B}}_\perp^{*\top} \mathbf{Q}^t \mathbf{W}^{t+1\top}$:

$$\hat{\mathbf{B}}_\perp^{*\top} \mathbf{Q}^t \mathbf{W}^{t+1\top} = \hat{\mathbf{B}}_\perp^{*\top} \left(\mathbf{W}^{t+1\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} + \mathbf{S}^{t+1\top} - \mathbf{S}^{*\top} \right) \mathbf{W}^{t+1\top}$$

$$= \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} - \hat{\mathbf{B}}_{\perp}^{*\top} (\mathbf{S}^* - \mathbf{S}^{t+1}) \mathbf{W}^{t+1\top},$$

plugging this into (20) then we reach to

$$\begin{aligned} \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^{t+1} &= \left(\hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t \left(\mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right) + \frac{\eta}{rn} \hat{\mathbf{B}}_{\perp}^{*\top} (\mathbf{S}^* - \mathbf{S}^{t+1}) \mathbf{W}^{t+1\top} \right. \\ &\quad \left. - \frac{\eta}{rn} \hat{\mathbf{B}}_{\perp}^{*\top} \left(\frac{1}{m} (\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \right) (\mathbf{R}^{t+1})^{-1}. \end{aligned} \quad (21)$$

Therefore,

$$\begin{aligned} \text{dist}(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*) &= \left\| \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^{t+1} \right\|_2 \\ &\leq \left\| \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t \left(\mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right) \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 \\ &\quad + \frac{\eta}{rn} \left\| \hat{\mathbf{B}}_{\perp}^{*\top} \left(\frac{1}{m} (\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 \\ &\quad + \frac{\eta}{rn} \left\| \hat{\mathbf{B}}_{\perp}^{*\top} (\mathbf{S}^* - \mathbf{S}^{t+1}) \mathbf{W}^{t+1\top} \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2. \end{aligned} \quad (22)$$

References

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

A Proofs