PFL

November 29, 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.

Initialize $\mathbf{B}^0, \mathbf{w}_1^0, ..., \mathbf{w}_n^0, \mathbf{s}_1^0, ..., \mathbf{s}_n^0$
for $t = 0, 1, 2, ..., T-1$ **do**
    Server receives a batch of clients $\mathcal{I}^t$ of size rn
    Server sends current representation ϕ^t to clients in $\mathcal{I}^t$
    for each client i in $\mathcal{I}^t$ **do**
        Client i initializes $\mathbf{w}_i^{t,0} \leftarrow \mathbf{w}_i^{t-1,\tau_h}$
        Client updates its head for τ_h steps:
        for $\tau = 1$ to τ_w **do**
            $\mathbf{w}_i^{t,\tau} \leftarrow \mathrm{GRD}\left(f_i\left(\mathbf{w}_i^{t,\tau-1}, \mathbf{B}^{t-1}, \mathbf{s}_i^{t-1,\tau_s}\right), \mathbf{w}_i^{t,\tau-1}, \eta\right)$
        end for
        Client i initializes $\mathbf{s}_i^{t,0} \leftarrow \mathbf{s}_i^{t-1,\tau_s}$
        Client i updates its shortcut for τ_s steps:
        for $\tau = 1$ to τ_s **do**
            $\mathbf{s}_i^{t,\tau} \leftarrow \mathrm{GRD}\left(f_i\left(\mathbf{w}_i^{t-1}, \mathbf{B}^{t-1}, \mathbf{s}_i^{t,\tau-1}\right), \mathbf{s}_i^{t,\tau-1}, \eta\right)$
        end for
        Client i initializes $\mathbf{B}_i^{t,0} \leftarrow \mathbf{B}^{t-1}$
        Client i updates its representation for τ_b steps:
        for $\tau = 1$ to τ_b **do**
            $\mathbf{B}_i^{t,\tau} \leftarrow \mathrm{GRD}\left(f_i\left(\mathbf{w}_i^{t,\tau_w}, \mathbf{B}_i^{t,\tau-1}, \mathbf{s}_i^{t,\tau_s}\right), \mathbf{B}_i^{t,\tau-1}, \eta\right)$
        end for
        Client i sends updated representation $\mathbf{B}_i^{t,\tau_b}$ to server
    end for
    for each client j not in $\mathcal{I}^t$ **do**
        Set $\mathbf{w}_i^{t,\tau_w} \leftarrow \mathbf{w}_i^{t-1,\tau_w}$ and $\mathbf{s}_i^{t,\tau_s} \leftarrow \mathbf{s}_i^{t-1,\tau_s}$
    end for
    Server computes new representation: $\mathbf{B}^t = \frac{1}{rn}\sum_{i\in\mathcal{I}^t}\mathbf{B}_i^{t,\tau_b}$
end for

1.1 Preliminaries

First, we establish the notations that will be used throughout our proof. Let $\mathbf{S} := [\mathbf{s}_1, ..., \mathbf{s}_{rn}] \in \mathbb{R}^{d\times rn}$ represent the personalized layers, and let $\mathbf{W} := [\mathbf{w}_1, ..., \mathbf{w}_{rn}] \in \mathbb{R}^{k\times rn}$ 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 $\mathbf{H} := [\mathbf{h}_1, ..., \mathbf{h}_{rn}] \in \mathbb{R}^{(k+d)\times rn}$.

...

The global objective can be rewritten as

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

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:

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

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

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

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

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 $\widetilde{F}\left(\hat{\mathbf{B}}^t, \mathbf{W}, \mathbf{S}^t\right) := \frac{1}{2rnm} \left\| \mathcal{A}\left(\mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} - \mathbf{W}^\top\hat{\mathbf{B}}^{t\top} + \mathbf{S}^{*\top} - \mathbf{S}^{t\top} \right) \right\|_2^2$.

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

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

which means

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

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 $\mathbf{D}_{pq} := \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left\langle\hat{\mathbf{b}}_p^t,\hat{\mathbf{b}}_q^*\right\rangle\mathbf{I}_{rn}$, 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,

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

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

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

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

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

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

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

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

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

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

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

$$\mathbf{f}_i := \left(\mathbf{G}^i\right)^{-1}\left(\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\right)\mathbf{w}_i^*, \tag{14}$$

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

$$\mathbf{h}_i = \left(\mathbf{G}^i\right)^{-1}\mathbf{E}^i\left(\mathbf{s}_i^t - \mathbf{s}_i^*\right), \tag{15}$$

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

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

$$\begin{aligned}\mathbf{w}_i^{t+1} &= \mathbf{D}^i\mathbf{w}_i^* - \left(\mathbf{G}^i\right)^{-1}\left(\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\right)\mathbf{w}_i^* + \left(\mathbf{G}^i\right)^{-1}\mathbf{E}^i\left(\mathbf{s}_i^t - \mathbf{s}_i^*\right)\\ &= \hat{\mathbf{B}}^\top\hat{\mathbf{B}}^*\mathbf{w}_i^* - \mathbf{f}_i + \mathbf{h}_i.\end{aligned} \tag{16}$$

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

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

where $\mathbf{F} := [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_{rn}]$ and $\mathbf{H} := [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{rn}]$. Then, we consider the update for $\mathbf{S}$, for which we apply the least squares methods, leading to

$$\mathbf{S}^{t+1} = \mathbf{S}^* + \hat{\mathbf{B}}^*\mathbf{W}^* - \hat{\mathbf{B}}^t\mathbf{W}^t. \tag{18}$$

Next, we recall three lemmas from (Collins et al., 2021) to bound $\mathbf{F}$.

Lemma 1 (Collins et al., 2021) *Let $\delta_k = c\frac{k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}$ for some absolute constant c, then*

$$\left\|\mathbf{G}^{-1}\right\|_2 \leq \frac{1}{1-\delta_k} \tag{19}$$

with probability at least $1-e^{-111k^3\log(rn)}$.

Lemma 2 (Collins et al., 2021) *Let $\delta_k = c\frac{k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}$ for some absolute constant c, then*

$$\left\|\left(\mathbf{GD}-\mathbf{C}\right)\widetilde{\mathcal{W}}^*\right\|_2 \leq \delta_k \left\|\mathbf{W}^*\right\|_2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \tag{20}$$

with probability at least $1-e^{-111k^2\log(rn)}$

Lemma 3 (Collins et al., 2021) *Let* $\delta_k = c\frac{k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}$ *for some absolute constant* c*, then*

$$\|\mathbf{F}\|_{\mathrm{F}} \leq \frac{\delta_k}{1-\delta_k} \|\mathbf{W}^*\|_2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \tag{21}$$

with probability at least $1 - e^{-110k^2 \log(rn)}$

Next, we focus on bounding $\|\mathbf{H}\|_2$.

Lemma 4 *...*

Proof: Recall that $\mathbf{H} := [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{rn}]$ and

$$\mathbf{h}_i = \left(\mathbf{G}^i\right)^{-1} \mathbf{E}^i \left(\mathbf{s}_i^t - \mathbf{s}_i^*\right), \tag{22}$$

then we have

$$\|\mathbf{h}_i\|_2 \leq \left\|\left(\mathbf{G}^i\right)^{-1}\right\|_2 \left\|\mathbf{E}^i\right\|_2 \left\|\mathbf{s}_i^t - \mathbf{s}_i^*\right\|_2 \tag{23}$$

□

Lemma 5 *Let* $\delta_k' = ...$ *for some absolute constant* c*. Then for any* t*,*

$$\frac{1}{rn}\left\|\left(\frac{1}{m}\mathcal{A}^{\dagger}\mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top}\right)^{\top} \mathbf{W}^{t+1\top}\right\|_2 \leq \tag{24}$$

Let $\mathbf{Q}^t = \hat{\mathbf{B}}^t\mathbf{W}^{t+1} - \hat{\mathbf{B}}^*\mathbf{W}^* + \mathbf{S}^{t+1} - \mathbf{S}^*$. 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

$$\begin{aligned}
\mathbf{q}_i &= \hat{\mathbf{B}}^t\mathbf{w}_i^{t+1} - \hat{\mathbf{B}}^*\mathbf{w}_i^* + \mathbf{s}_i^{t+1} - \mathbf{s}_i^* \\
&= \hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{w}_i^* - \hat{\mathbf{B}}^t\mathbf{f}_i + \hat{\mathbf{B}}^t\mathbf{h}_i - \hat{\mathbf{B}}^*\mathbf{w}_i^* + \hat{\mathbf{B}}^*\left(\mathbf{w}_i^* - \mathbf{w}_i^t\right) \\
&= \left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\mathbf{w}_i^* - \hat{\mathbf{B}}^t\mathbf{f}_i + \hat{\mathbf{B}}^t\mathbf{h}_i + \hat{\mathbf{B}}^*\left(\mathbf{w}_i^* - \mathbf{w}_i^t\right)
\end{aligned} \tag{25}$$

Thus,

$$\begin{aligned}
\|\mathbf{q}_i\|_2 &= \left\|\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\mathbf{w}_i^* - \hat{\mathbf{B}}^t\mathbf{f}_i + \hat{\mathbf{B}}^t\mathbf{h}_i + \hat{\mathbf{B}}^*\left(\mathbf{w}_i^* - \mathbf{w}_i^t\right)\right\|_2 \\
&\leq \left\|\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\right\|_2 \|\mathbf{w}_i^*\|_2 + \|\mathbf{f}_i\|_2 + \|\mathbf{h}_i\|_2 + \left\|\mathbf{w}_i^* - \mathbf{w}_i^t\right\|_2
\end{aligned} \tag{26}$$

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

$$
\begin{aligned}
\bar{\mathbf{B}}^{t+1} &= \hat{\mathbf{B}}^t - \frac{\eta}{rnm}\left((\mathcal{A}^t)^\top \mathcal{A}^t(\mathbf{Q}^{t\top})\right)^\top \mathbf{W}^{t+1\top} \\
&= \hat{\mathbf{B}}^t - \frac{\eta}{rn}\mathbf{Q}^t\mathbf{W}^{t+1\top} - \frac{\eta}{rn}\left(\frac{1}{m}(\mathcal{A}^t)^\top \mathcal{A}^t(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}. \qquad (27)
\end{aligned}
$$

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

$$
\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^{t+1} = \left(\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\mathbf{Q}^t\mathbf{W}^{t+1\top} - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}(\mathcal{A}^t)^\top \mathcal{A}^t(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}\right)(\mathbf{R}^{t+1})^{-1}. \qquad (28)
$$

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

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

plugging this into (28) then we reach to

$$
\begin{aligned}
\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^{t+1} = \Bigg(&\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t\left(\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right) + \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\mathbf{S}^* - \mathbf{S}^{t+1}\right)\mathbf{W}^{t+1\top} \\
&- \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}(\mathcal{A}^t)^\top \mathcal{A}^t(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}\Bigg)(\mathbf{R}^{t+1})^{-1}. \qquad (29)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\operatorname{dist}(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*) &= \left\|\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^{t+1}\right\|_2 \\
&\le \left\|\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t\left(\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right)\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \\
&\quad + \frac{\eta}{rn}\left\|\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}(\mathcal{A}^t)^\top \mathcal{A}^t(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \\
&\quad + \frac{\eta}{rn}\left\|\hat{\mathbf{B}}_\perp^{*\top}\left(\mathbf{S}^* - \mathbf{S}^{t+1}\right)\mathbf{W}^{t+1\top}\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2. \qquad (30)
\end{aligned}
$$

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