PFL

December 5, 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 $\mathbf{B}^0, \mathbf{w}_1^0, ..., \mathbf{w}_n^0, \mathbf{s}_1^0, ..., \mathbf{s}_n^0$
2: **for** $t = 0, 1, 2, ..., T-1$ **do**
3:     Server receives a batch of clients $\mathcal{I}^t$ of size rn
4:     Server sends current representation ϕ^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 τ_h steps:
8:         **for** $\tau = 1$ to τ_w **do**
9:             $\mathbf{w}_i^{t,\tau} \leftarrow \mathrm{GRD}\left(f_i\left(\mathbf{w}_i^{t,\tau-1}, \mathbf{B}^{t-1}, \mathbf{s}_i^{t-1,\tau_s}\right), \mathbf{w}_i^{t,\tau-1}, \eta\right)$
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 τ_s steps:
13:         **for** $\tau = 1$ to τ_s **do**
14:             $\mathbf{s}_i^{t,\tau} \leftarrow \mathrm{GRD}\left(f_i\left(\mathbf{w}_i^{t-1}, \mathbf{B}^{t-1}, \mathbf{s}_i^{t,\tau-1}\right), \mathbf{s}_i^{t,\tau-1}, \eta\right)$
15:         **end for**
16:         Client i initializes $\mathbf{B}_i^{t,0} \leftarrow \mathbf{B}^{t-1}$
17:         Client i updates its representation for τ_b steps:
18:         **for** $\tau = 1$ to τ_b **do**
19:             $\mathbf{B}_i^{t,\tau} \leftarrow \mathrm{GRD}\left(f_i\left(\mathbf{w}_i^{t,\tau_w}, \mathbf{B}_i^{t,\tau-1}, \mathbf{s}_i^{t,\tau_s}\right), \mathbf{B}_i^{t,\tau-1}, \eta\right)$
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 $\mathbf{S} := [\mathbf{s}_1, ..., \mathbf{s}_{rn}] \in \mathbb{R}^{d \times rn}$ 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:

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

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

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

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

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

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

which means

$$\frac{1}{m} \sum_{q=1}^{k} \left(\sum_{i=1}^{rn} \sum_{j=1}^{m} \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \hat{\mathbf{b}}_q^{t\top} \mathbf{A}_{i,j}^\top \right) \mathcal{W}_q^{t+1}$$

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

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,

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

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,

$$\widetilde{\mathcal{W}}^{t+1} = \mathbf{G}^{-1}\mathbf{C}\widetilde{\mathcal{W}}^* - \mathbf{G}^{-1}\mathbf{E} \left(\widetilde{\mathcal{S}}^* - \widetilde{\mathcal{S}}^t \right)$$

$$= \mathbf{D}\widetilde{\mathcal{W}}^* - \mathbf{G}^{-1} \left(\mathbf{GD} - \mathbf{C} \right) \widetilde{\mathcal{W}}^* - \mathbf{G}^{-1}\mathbf{E} \left(\widetilde{\mathcal{S}}^* - \widetilde{\mathcal{S}}^t \right), \qquad (9)$$

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

$$\mathbf{G}_{pq} = \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^{m} \mathbf{A}_{i,j} \hat{\mathbf{b}}_p \hat{\mathbf{b}}_q^\top \mathbf{A}_{i,j}^\top$$

$$= \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^{m} \mathbf{e}_i \left(\mathbf{x}_i^j \right)^\top \hat{\mathbf{b}}_p \hat{\mathbf{b}}_q^\top \mathbf{x}_i^j \mathbf{e}_i^\top, \qquad (10)$$

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

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

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 $(\mathbf{C}_{pq})_{ii} = \hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \hat{\mathbf{b}}_q^*$, 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

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

3

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

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

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

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 $\mathbf{G}^{-1}\mathbf{E}\left(\widetilde{\mathcal{S}}^* - \widetilde{\mathcal{S}}^t\right)$ into i vectors, defined as

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

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

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

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

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

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 $\mathbf{S}$. Similarly, $\mathbf{S}^{t+1}$ minimizes $\Phi\left(\hat{\mathbf{B}}^t, \mathbf{W}^t, \mathbf{S}\right) := \frac{1}{2rnm}\left\|\mathcal{A}\left(\mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} - \mathbf{W}^{t\top}\hat{\mathbf{B}}^{t\top} + \mathbf{S}^{*\top} - \mathbf{S}^\top\right)\right\|_2^2$, therefore we have

$$\begin{aligned}
\mathbf{0} &= \nabla_{\mathbf{S}}\Phi\left(\hat{\mathbf{B}}^t, \mathbf{W}^t, \mathbf{S}^{t+1}\right) \\
&= \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{A}_{i,j}, \mathbf{W}^{t\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} + \mathbf{S}^{t+1\top} - \mathbf{S}^{*\top}\rangle\right)\mathbf{A}_{i,j} \qquad (18)
\end{aligned}$$

Then we reach

$$\mathbf{S}^{t+1} = \mathbf{S}^* + \hat{\mathbf{B}}^* \mathbf{W}^* - \hat{\mathbf{B}}^t \mathbf{W}^t. \qquad (19)$$

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*

$$\left\|\mathbf{G}^{-1}\right\|_2 \le \frac{1}{1 - \delta_k} \qquad (20)$$

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

4

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*

$$\left\|(\mathbf{GD} - \mathbf{C})\,\widetilde{\mathcal{W}}^*\right\|_2 \leq \delta_k \left\|\mathbf{W}^*\right\|_2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \tag{21}$$

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*

$$\left\|\mathbf{F}\right\|_{\mathrm{F}} \leq \frac{\delta_k}{1 - \delta_k} \left\|\mathbf{W}^*\right\|_2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \tag{22}$$

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

$$\mathbf{h}_i = \left(\mathbf{G}^i\right)^{-1} \mathbf{E}^i \left(\mathbf{s}_i^* - \mathbf{s}_i^t\right) = \hat{\mathbf{B}}^{t\top}\left(\mathbf{s}_i^* - \mathbf{s}_i^t\right) - \left(\mathbf{G}^i\right)^{-1}\left(\mathbf{G}^i\hat{\mathbf{B}}^{t\top} - \mathbf{E}^i\right)\left(\mathbf{s}_i^* - \mathbf{s}_i^t\right), \tag{23}$$

then we focus on the term of $\mathbf{G}^i\hat{\mathbf{B}}^{t\top} - \mathbf{E}^i$, for which we have

$$\mathbf{G}^i\hat{\mathbf{B}}^{t\top} - \mathbf{E}^i = \hat{\mathbf{B}}^{t\top}\left(\frac{1}{m}\mathbf{X}_i^\top\mathbf{X}_i\right)\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right). \tag{24}$$

Let $\mathbf{U} := \frac{1}{\sqrt{m}}\mathbf{X}_i\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)$ and $\mathbf{V} := \frac{1}{\sqrt{m}}\mathbf{X}_i\hat{\mathbf{B}}^t$, then we have the j-th row of $\mathbf{U}$ and $\mathbf{V}$ as the following, respectively:

$$\mathbf{u}_j = \frac{1}{\sqrt{m}}\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\mathbf{x}_i^j, \quad \mathbf{v}_j = \frac{1}{\sqrt{m}}\hat{\mathbf{B}}^{t\top}\mathbf{x}_i^j. \tag{25}$$

Note that $\mathbf{u}_j$ is $\frac{1}{\sqrt{m}}\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)$-sub-gaussian and $\mathbf{v}_j$ is $\frac{1}{\sqrt{m}}\hat{\mathbf{B}}^t$-sub-gaussian, therefore we can argue similarly as the derivatives for Theorem 4.4.5 in ([Vershynin, 2018](#)). First, let $\mathcal{S}^{d-1}$ be the d-dimension unit sphere and $\mathcal{S}^{k-1}$ be the k-dimension unit sphere, then let $\mathcal{N}_d$ be the $\frac{1}{4}$-th net on $\mathcal{S}^{d-1}$ and $\mathcal{N}_k$ be the $\frac{1}{4}$-th net on $\mathcal{S}^{k-1}$, such that $|\mathcal{N}_d| \leq 9^d$ and $|\mathcal{N}_k| \leq 9^k$, which exists according to Corollary 4.2.13 in ([Vershynin, 2018](#)). Next, by leveraging inequality 4.13 in ([Vershynin, 2018](#)), we have

$$\left\|\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\left(\frac{1}{m}\mathbf{X}_i^\top\mathbf{X}_i\right)\hat{\mathbf{B}}^t\right\|_2 = \left\|\mathbf{U}^\top\mathbf{V}\right\|_2 \leq 2\max_{\mathbf{z}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k}\mathbf{z}^\top\mathbf{U}^\top\mathbf{V}\mathbf{y}$$

$$= 2\max_{\mathbf{z}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k}\mathbf{z}^\top\left(\sum_{j=1}^m \mathbf{u}_j\mathbf{v}_j^\top\right)\mathbf{y}$$

$$= 2\max_{\mathbf{z}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k}\sum_{j=1}^m \langle\mathbf{z}, \mathbf{u}_j\rangle\langle\mathbf{v}_j, \mathbf{y}\rangle. \tag{26}$$

By the definition of sub-gaussianity, $\langle \mathbf{z}, \mathbf{u}_j \rangle$ is sub-gaussian with norm at most $\frac{1}{\sqrt{m}} \left\| \hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} - \mathbf{I}_d \right\|_2 \leq \frac{2}{\sqrt{m}}$ and $\langle \mathbf{v}_j, \mathbf{y} \rangle$ is sub-gaussian with norm at most $\frac{1}{\sqrt{m}} \left\| \hat{\mathbf{B}}^{t\top} \right\|_2 = \frac{1}{\sqrt{m}}$. Therefore, $\langle \mathbf{z}, \mathbf{u}_j \rangle \langle \mathbf{v}_j, \mathbf{y} \rangle$ is sub-exponential with norm at most $\frac{c}{m}$ for some absolute constant c, for all $j \in [m]$. Also, for any $j \in [m]$ and any $\mathbf{z} \in \mathcal{N}_d, \mathbf{y} \in \mathcal{N}_k$, we have

$$\mathbb{E}[\langle \mathbf{z}, \mathbf{u}_j \rangle \langle \mathbf{v}_j, \mathbf{y} \rangle] = \mathbb{E}[\mathbf{z}^\top \left(\hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} - \mathbf{I}_d \right) \frac{1}{m} \mathbf{X}_i^\top \mathbf{X}_i \hat{\mathbf{B}}^t] = 0. \tag{27}$$

Thus, we obtain a sum of m mean-zero, independent sub-exponential random variables, for which we apply Bernstein's inequality, for any $\mathbf{z} \in \mathcal{N}_d, \mathbf{y} \in \mathcal{N}_k$,

$$\mathbb{P} \left(\sum_{j=1}^{m} \langle \mathbf{z}, \mathbf{u}_j \rangle \langle \mathbf{v}_j, \mathbf{y} \rangle \geq s \right) \leq e^{-c'm \min\left(s^2, s\right)}. \tag{28}$$

Union bounding over all $\mathbf{z} \in \mathcal{N}_d, \mathbf{y} \in \mathcal{N}_k$, we obtain

$$\mathbb{P} \left(\left\| \left(\hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} - \mathbf{I}_d \right) \left(\frac{1}{m} \mathbf{X}_i^\top \mathbf{X}_i \right) \hat{\mathbf{B}}^t \right\|_2 \geq 2s \right) \leq 9^{d+k} e^{-c'm \min\left(s^2, s\right)}. \tag{29}$$

Here, let $s = \max\left(\varepsilon, \varepsilon^2\right)$ for some $\varepsilon > 0$, then we have $\min\left(s^2, s\right) = \varepsilon^2$. Then we reach

$$\mathbb{P} \left(\left\| \left(\hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} - \mathbf{I}_d \right) \left(\frac{1}{m} \mathbf{X}_i^\top \mathbf{X}_i \right) \hat{\mathbf{B}}^t \right\|_2 \geq 2 \max\left(\varepsilon, \varepsilon^2\right) \right) \leq 9^{d+k} e^{-c'm\varepsilon^2}. \tag{30}$$

Further, let $\varepsilon = \sqrt{\frac{ck^2 \log(rn)}{4m}}$ for some constant c, and $m \geq ck^2 \log(rn)$, we have

$$\mathbb{P} \left(\left\| \left(\hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} - \mathbf{I}_d \right) \left(\frac{1}{m} \mathbf{X}_i^\top \mathbf{X}_i \right) \hat{\mathbf{B}}^t \right\|_2 \geq \sqrt{\frac{ck^2 \log(rn)}{m}} \right) \leq 9^{d+k} e^{-c_1 k^2 \log(rn)} \leq e^{-111 k^2 \log(rn)}, \tag{31}$$

for a large enough constant c_1. According to (23),

$$\|\mathbf{h}_i\|_2 \leq \left\| \hat{\mathbf{B}}^{t\top} \right\|_2 \left\| \mathbf{s}_i^* - \mathbf{s}_i^t \right\|_2 + \left\| \left(\mathbf{G}^i \right)^{-1} \right\|_2 \left\| \mathbf{G}^i \hat{\mathbf{B}}^{t\top} - \mathbf{E}^i \right\|_2 \left\| \mathbf{s}_i^* - \mathbf{s}_i^t \right\|_2 \tag{32}$$
$$= \left(1 + \left\| \left(\mathbf{G}^i \right)^{-1} \right\|_2 \left\| \mathbf{G}^i \hat{\mathbf{B}}^{t\top} - \mathbf{E}^i \right\|_2 \right) \left\| \mathbf{s}_i^* - \mathbf{s}_i^t \right\|_2.$$

From (31), we know that

$$\mathbb{P} \left(\left\| \mathbf{G}^i \hat{\mathbf{B}}^{t\top} - \mathbf{E}^i \right\|_2 \geq \delta_k \right) \leq e^{-111 k^2 \log(rn)}, \tag{33}$$

and from equation (43) in (Collins et al., 2021) we have

$$\mathbb{P} \left(\left\| \left(\mathbf{G}^i \right)^{-1} \right\|_2 \geq \frac{1}{1 - \delta_k} \right) \leq e^{-121 k^3 \log(rn)} \tag{34}$$

Therefore, we obtain

$$\|\mathbf{h}_i\|_2 \leq \left(1 + \frac{\delta_k}{1 - \delta_k} \right) \left\| \mathbf{s}_i^* - \mathbf{s}_i^t \right\|_2 = \frac{1}{1 - \delta_k} \left\| \mathbf{s}_i^* - \mathbf{s}_i^t \right\|_2 \tag{35}$$

6

with probability at least $1 - e^{-111k^2 \log(rn)}$. Finally we take a union bound over $i \in [rn]$, leading to

$$\mathbb{P}\left(\frac{1}{rn}\|\mathbf{H}\|_2^2 \geq \frac{1}{(1-\delta_k)^2}\|\mathbf{S}^* - \mathbf{S}^t\|_{\mathrm{F}}^2\right) \leq \mathbb{P}\left(\frac{1}{rn}\sum_{i=1}^{rn}\|\mathbf{h}_i\|_2^2 \geq \frac{1}{(1-\delta_k)^2}\sum_{i=1}^{rn}\|\mathbf{s}_i^* - \mathbf{s}_i^t\|_2^2\right)$$

$$\leq rn\mathbb{P}\left(\|\mathbf{h}_i\|_2^2 \geq \frac{1}{(1-\delta_k)^2}\|\mathbf{s}_i^* - \mathbf{s}_i^t\|_2^2\right)$$

$$\leq e^{-110k^2 \log(rn)} \tag{36}$$

and thus completing the proof. $\qquad\square$

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

$$\frac{1}{rn}\left\|\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}\right\|_2 \leq \tag{37}$$

Proof: 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

$$\mathbf{q}_i = \hat{\mathbf{B}}^t\mathbf{w}_i^{t+1} - \hat{\mathbf{B}}^*\mathbf{w}_i^* + \mathbf{s}_i^{t+1} - \mathbf{s}_i^*$$
$$= \hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{w}_i^* - \hat{\mathbf{B}}^t\mathbf{f}_i - \hat{\mathbf{B}}^t\mathbf{h}_i - \hat{\mathbf{B}}^*\mathbf{w}_i^* + \hat{\mathbf{B}}^*\left(\mathbf{w}_i^* - \mathbf{w}_i^t\right)$$
$$= \left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\mathbf{w}_i^* - \hat{\mathbf{B}}^t\mathbf{f}_i - \hat{\mathbf{B}}^t\mathbf{h}_i + \hat{\mathbf{B}}^*\left(\mathbf{w}_i^* - \mathbf{w}_i^t\right) \tag{38}$$

Thus,

$$\|\mathbf{q}_i\|_2 = \left\|\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\mathbf{w}_i^* - \hat{\mathbf{B}}^t\mathbf{f}_i - \hat{\mathbf{B}}^t\mathbf{h}_i + \hat{\mathbf{B}}^*\left(\mathbf{w}_i^* - \mathbf{w}_i^t\right)\right\|_2$$
$$\leq \left\|\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\right\|_2\|\mathbf{w}_i^*\|_2 + \|\mathbf{f}_i\|_2 + \|\mathbf{h}_i\|_2 + \left\|\mathbf{w}_i^* - \mathbf{w}_i^t\right\|_2$$
$$\leq 2\sqrt{k}\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + 2\left\|\mathbf{s}_i^* - \mathbf{s}_i^t\right\|_2 + \left\|\mathbf{w}_i^* - \mathbf{w}_i^t\right\|_2 \tag{39}$$
$$\leq 2\sqrt{k}\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + 2\left\|\mathbf{S}^* - \mathbf{S}^t\right\|_{\mathrm{F}} + \left\|\mathbf{W}^* - \mathbf{W}^t\right\|_{\mathrm{F}}, \tag{40}$$

where (39) holds with probability at least $1 - e^{-110k^2 \log(rn)}$, by combining equation (44) in (Collins et al., 2021) and (35), conditioned on $\delta_k \leq \frac{1}{2}$. Similarly, combining equation (45) and (35), conditioned on $\delta_k \leq \frac{1}{2}$, we have

$$\left\|\mathbf{w}_i^{t+1}\right\|_2 \leq \left\|\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{w}_i^*\right\|_2 + \|\mathbf{f}_i\|_2 + \|\mathbf{h}_i\|_2$$
$$\leq 2\sqrt{k} + 2\left\|\mathbf{s}_i^* - \mathbf{s}_i^t\right\|_2 \tag{41}$$
$$\leq 2\sqrt{k} + 2\left\|\mathbf{S}^* - \mathbf{S}^t\right\|_{\mathrm{F}}, \tag{42}$$

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

7

Next, just for simple notation, let $\Delta \mathbf{S}^t$ denote $\mathbf{S}^* - \mathbf{S}^t$ and $\Delta \mathbf{W}^t$ denote $\mathbf{W}^* - \mathbf{W}^t$. and in the following proof, we condition on the event

$$\mathcal{E} := \bigcap_{i=1}^{rn} \left\{ \|\mathbf{q}_i\|_2 \leq 2\sqrt{k} \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + 2\left\|\Delta \mathbf{S}^t\right\|_{\mathrm{F}} + \left\|\Delta \mathbf{W}^t\right\|_{\mathrm{F}} \cap \left\|\mathbf{w}_i^{t+1}\right\|_2 \leq 2\sqrt{k} + 2\left\|\Delta \mathbf{S}^t\right\|_{\mathrm{F}} \right\}, \tag{43}$$

which holds with probability at least $1 - e^{-109k^2 \log(rn)}$. Next, we consider the following matrix:

$$\frac{1}{m}\mathcal{A}^\dagger \mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top} = \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left\langle \mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top, \mathbf{Q}^{t\top}\right\rangle \mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top - \mathbf{Q}^{t\top}$$

$$= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top - \mathbf{Q}^{t\top}, \tag{44}$$

further, we have

$$\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger \mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top} = \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{x}_i^j \mathbf{w}_i^\top - \mathbf{q}_i \mathbf{w}_i^\top\right). \tag{45}$$

Next, we establish similar arguments as the derivatives for Theorem 4.4.5 in (Vershynin, 2018) to bound $\left\|\frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{x}_i^j \mathbf{w}_i^\top - \mathbf{q}_i \mathbf{w}_i^\top\right)\right\|_2$. let $\mathcal{S}^{d-1}$ be the d-dimension unit sphere and $\mathcal{S}^{k-1}$ be the k-dimension unit sphere, then let $\mathcal{N}_d$ be the $\frac{1}{4}$-th net on $\mathcal{S}^{d-1}$ and $\mathcal{N}_k$ be the $\frac{1}{4}$-th net on $\mathcal{S}^{k-1}$, such that $|\mathcal{N}_d| \leq 9^d$ and $|\mathcal{N}_k| \leq 9^k$, which exists according to Corollary 4.2.13 in (Vershynin, 2018). Using equation 4.13 in (Vershynin, 2018), we have

$$\left\|\frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{x}_i^j \mathbf{w}_i^\top - \mathbf{q}_i \mathbf{w}_i^\top\right)\right\|_2$$

$$\leq 2\max_{\mathbf{z}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k} \mathbf{z}^\top \left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{x}_i^j \mathbf{w}_i^\top - \frac{1}{rnm}\mathbf{q}_i \mathbf{w}_i^\top\right)\right)\mathbf{y}$$

$$= 2\max_{\mathbf{z}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\langle \mathbf{z}, \mathbf{x}_i^j\rangle\langle \mathbf{w}_i, \mathbf{y}\rangle - \frac{1}{rnm}\langle \mathbf{z}, \mathbf{q}_i\rangle\langle \mathbf{w}_i, \mathbf{y}\rangle\right) \tag{46}$$

Since $\mathbf{x}_i^j$ is $\mathbf{I}_d$-sub-gaussian, $\langle \mathbf{z}, \mathbf{x}_i^j\rangle$ is sub-gaussian with norm at most $c\|\mathbf{z}\|_2 = c$ for some absolute constant c and any $\mathbf{z}\in\mathcal{N}_d$. Also $\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle$ is sub-gaussian with norm at most $\|\mathbf{q}_i\|_2$. Therefore, $\langle \mathbf{z}, \mathbf{x}_i^j\rangle\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle$ is sub-exponential with norm at most $c\|\mathbf{q}_i\|_2$, which indicates $\frac{1}{rnm}\langle \mathbf{z}, \mathbf{x}_i^j\rangle\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\langle \mathbf{w}_i, \mathbf{y}\rangle$ is sub-exponential with norm at most

$$\frac{c}{rnm}\|\mathbf{q}_i\|_2\langle \mathbf{w}_i, \mathbf{y}\rangle \leq \frac{c}{rnm}\|\mathbf{q}_i\|_2\|\mathbf{w}_i\|_2$$

$$\leq \frac{c}{rnm}\left(2\sqrt{k}\operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + 2\left\|\Delta \mathbf{S}^t\right\|_{\mathrm{F}} + \left\|\Delta \mathbf{W}^t\right\|_{\mathrm{F}}\right)\left(2\sqrt{k} + 2\left\|\Delta \mathbf{S}^t\right\|_{\mathrm{F}}\right) \tag{47}$$

$$\leq \frac{c_2}{rnm}\left(2k\operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + 2\sqrt{k}\left\|\Delta \mathbf{S}^t\right\|_{\mathrm{F}} + \sqrt{k}\left\|\Delta \mathbf{W}^t\right\|_{\mathrm{F}}\right)$$

8

$$+ \frac{c_2}{rnm} \left(2\sqrt{k} \operatorname{dist} \left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) \left\| \Delta \mathbf{S}^t \right\|_{\mathrm{F}} + 2 \left\| \Delta \mathbf{S}^t \right\|_{\mathrm{F}}^2 + \left\| \Delta \mathbf{S}^t \right\|_{\mathrm{F}} \left\| \Delta \mathbf{W}^t \right\|_{\mathrm{F}} \right) \quad (48)$$

Since $\mathbb{E}[\frac{1}{rnm} \langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \langle \mathbf{z}, \mathbf{x}_i^j \rangle \langle \mathbf{w}_i, \mathbf{y} \rangle - \frac{1}{rnm} \langle \mathbf{z}, \mathbf{q}_i \rangle \langle \mathbf{w}_i, \mathbf{y} \rangle] = 0$, we have a sum of rnm independent, mean zero, sub-exponential random variables, for which we can apply Bernstein's inequality and obtain

$$\mathbb{P} \left(\sum_{i=1}^{rn} \sum_{j=1}^{m} \left(\frac{1}{rnm} \langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \langle \mathbf{z}, \mathbf{x}_i^j \rangle \langle \mathbf{w}_i, \mathbf{y} \rangle - \frac{1}{rnm} \langle \mathbf{z}, \mathbf{q}_i \rangle \langle \mathbf{w}_i, \mathbf{y} \rangle \right) \geq s \right)$$
$$\leq \exp \left(\right) \quad (49)$$

$\square$

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

$$\bar{\mathbf{B}}^{t+1} = \hat{\mathbf{B}}^t - \frac{\eta}{rnm} \left((\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) \right)^\top \mathbf{W}^{t+1\top}$$
$$= \hat{\mathbf{B}}^t - \frac{\eta}{rn} \mathbf{Q}^t \mathbf{W}^{t+1\top} - \frac{\eta}{rn} \left(\frac{1}{m} (\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top}. \quad (50)$$

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

$$\hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^{t+1} = \left(\hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t - \frac{\eta}{rn} \hat{\mathbf{B}}_{\perp}^{*\top} \mathbf{Q}^t \mathbf{W}^{t+1\top} - \frac{\eta}{rn} \hat{\mathbf{B}}_{\perp}^{*\top} \left(\frac{1}{m} (\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \right) (\mathbf{R}^{t+1})^{-1}. \quad (51)$$

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

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

plugging this into (51) then we reach

$$\hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^{t+1} = \left(\hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t \left(\mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right) + \frac{\eta}{rn} \hat{\mathbf{B}}_{\perp}^{*\top} \left(\mathbf{S}^* - \mathbf{S}^{t+1} \right) \mathbf{W}^{t+1\top} \right.$$
$$\left. - \frac{\eta}{rn} \hat{\mathbf{B}}_{\perp}^{*\top} \left(\frac{1}{m} (\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \right) (\mathbf{R}^{t+1})^{-1}. \quad (52)$$

Therefore,

$$\operatorname{dist}(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*) = \left\| \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^{t+1} \right\|_2$$

9

$$\leq \left\| \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t \left(\mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right) \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2$$

$$+ \frac{\eta}{rn} \left\| \hat{\mathbf{B}}_{\perp}^{*\top} \left(\frac{1}{m} (\mathcal{A}^t)^\top \mathcal{A}^t (\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2$$

$$+ \frac{\eta}{rn} \left\| \hat{\mathbf{B}}_{\perp}^{*\top} \left(\mathbf{S}^* - \mathbf{S}^{t+1} \right) \mathbf{W}^{t+1\top} \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2. \tag{53}$$

Next, we focus on the term of $\left\| \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t \left(\mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right) \right\|_2$, for which we have

$$\left\| \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t \left(\mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right) \right\|_2 \leq \left\| \hat{\mathbf{B}}_{\perp}^{*\top} \hat{\mathbf{B}}^t \right\|_2 \left\| \mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right\|_2$$

$$\leq \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) \left\| \mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right\|_2. \tag{54}$$

To bound the term of $\left\| \mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right\|_2$, first we consider

$$\left\| \mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1} \mathbf{W}^{t+1\top} \right\|_2$$

$$= \left\| \mathbf{I}_k - \frac{\eta}{rn} \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} - \mathbf{H} \right) \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} - \mathbf{H} \right)^\top \right\|_2$$

$$= \left\| \mathbf{I}_k - \frac{\eta}{rn} \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right) \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right)^\top \right\|_2$$

$$+ \frac{\eta}{rn} \left\| \mathbf{H} \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right)^\top - \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right) \mathbf{H}^\top + \mathbf{H} \mathbf{H}^\top \right\|_2 \tag{55}$$

As for the first term, combining (66) and (67) in (Collins et al., 2021), we have

$$\left\| \mathbf{I}_k - \frac{\eta}{rn} \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right) \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right)^\top \right\|_2 \leq 1 - \eta \bar{\sigma}_{\min,*}^2 E_0 + 2\eta \frac{\delta_k}{1 - \delta_k} \bar{\sigma}_{\max,*}^2. \tag{56}$$

As for the second term,

$$\frac{\eta}{rn} \left\| \mathbf{H} \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right)^\top - \left(\hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right) \mathbf{H}^\top + \mathbf{H} \mathbf{H}^\top \right\|_2$$

$$\leq \frac{2\eta}{rn} \left\| \hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* - \mathbf{F} \right\|_2 \left\| \mathbf{H} \right\|_2 + \frac{\eta}{rn} \left\| \mathbf{H} \right\|_2^2 \tag{57}$$

$$\leq 2\eta \left(\frac{1}{\sqrt{rn}} \left\| \hat{\mathbf{B}}^{t\top} \hat{\mathbf{B}}^* \mathbf{W}^* \right\|_2 + \frac{1}{\sqrt{rn}} \left\| \mathbf{F} \right\|_2 \right) \frac{1}{\sqrt{rn}} \left\| \mathbf{H} \right\|_2 + \frac{\eta}{rn} \left\| \mathbf{H} \right\|_2^2 \tag{58}$$

$$\leq 2\eta \left(\frac{1}{\sqrt{rn}} \left\| \mathbf{W}^* \right\|_2 + \frac{1}{\sqrt{rn}} \frac{\delta_k}{1 - \delta_k} \left\| \mathbf{W}^* \right\|_2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) \right) \frac{1}{1 - \delta_k} \left\| \Delta \mathbf{S}^t \right\|_{\mathrm{F}} + \frac{\eta}{(1 - \delta_k)^2} \left\| \Delta \mathbf{S}^t \right\|_{\mathrm{F}}^2 \tag{59}$$

Then, we focus on bounding $\left\| (\mathbf{R}^{t+1})^{-1} \right\|_2$. Just for simple notation, let $\mathbf{U}^t := \frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^{t\top})$, then we have

$$\mathbf{R}^{t+1\top} \mathbf{R}^{t+1} = \bar{\mathbf{B}}^{t+1\top} \bar{\mathbf{B}}^{t+1}$$

$$= \hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^t - \frac{\eta}{rn}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\right) + \frac{\eta^2}{(rn)^2}\mathbf{W}^{t+1}\mathbf{U}^t\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}$$

$$= \mathbf{I}_k - \frac{\eta}{rn}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\right) + \frac{\eta^2}{(rn)^2}\mathbf{W}^{t+1}\mathbf{U}^t\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}. \qquad (60)$$

Using Weyl's Inequality, we reach

$$\sigma_{\min}^2\left(\mathbf{R}^{t+1}\right) \geq 1 - \frac{\eta}{rn}\lambda_{\max}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\right) + \frac{\eta^2}{(rn)^2}\lambda_{\min}\left(\mathbf{W}^{t+1}\mathbf{U}^t\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}\right)$$

$$\geq 1 - \frac{\eta}{rn}\lambda_{\max}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\right) \qquad (61)$$

where (61) holds since $\mathbf{W}^{t+1}\mathbf{U}^t\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}$ is positive semi-definite. Further,

$$\frac{\eta}{rn}\lambda_{\max}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\right)$$

$$= \max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{\eta}{rn}\left(\mathbf{z}^\top\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}\mathbf{z} + \mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\mathbf{z}\right)$$

$$= \max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\mathbf{z}$$

$$= \max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\left(\frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^t\mathbf{z} + \frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{Q}^{t\top}\hat{\mathbf{B}}^t\mathbf{z}\right) \qquad (62)$$

When considering the first term, we have

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^t\mathbf{z} \leq \frac{2\eta}{rn}\left\|\mathbf{W}^{t+1}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)\right\|_2\left\|\hat{\mathbf{B}}^t\right\|_2 \qquad (63)$$

Then we consider the second term in (62),

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{Q}^{t\top}\hat{\mathbf{B}}^t\mathbf{z} \leq \max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F}\right)\left(\mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top}\right)\hat{\mathbf{B}}^t\mathbf{z}$$

$$+ \max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\left(\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F}\right)\left(\mathbf{S}^{t+1\top} - \mathbf{S}^{*\top}\right) + \mathbf{H}\mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^t\mathbf{z} \qquad (64)$$

As for the first term in (64), from equation (81) in (Collins et al., 2021) we have

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F}\right)\left(\mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top}\right)\hat{\mathbf{B}}^t\mathbf{z} \leq 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}_{\max,*}^2. \qquad (65)$$

As for the second term in (64),

$$\frac{2\eta}{rn}\left\|\left(\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F}\right)\left(\mathbf{S}^{t+1\top} - \mathbf{S}^{*\top}\right) + \mathbf{H}\mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^t\right\|_2$$

$$\leq \frac{2\eta}{rn}\left\|\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F}\right\|_2\left\|\mathbf{S}^{t+1\top} - \mathbf{S}^{*\top}\right\|_2 + \frac{2\eta}{rn}\left\|\mathbf{H}\mathbf{Q}^{t\top}\right\|_2 \qquad (66)$$

11

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.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

A Proofs