PFL

December 16, 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**
      $\mathbf{w}_i^{t,\tau} \leftarrow \text{GRD}\left(f_i\left(\mathbf{w}_i^{t,\tau-1}, \mathbf{B}^{t-1}, \mathbf{s}_i^{t-1,\tau_s}\right), \mathbf{w}_i^{t,\tau-1}, \eta\right)$
    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**
      $\mathbf{s}_i^{t,\tau} \leftarrow \text{GRD}\left(f_i\left(\mathbf{w}_i^{t-1}, \mathbf{B}^{t-1}, \mathbf{s}_i^{t,\tau-1}\right), \mathbf{s}_i^{t,\tau-1}, \eta\right)$
    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**
      $\mathbf{B}_i^{t,\tau} \leftarrow \text{GRD}\left(f_i\left(\mathbf{w}_i^{t,\tau_w}, \mathbf{B}_i^{t,\tau-1}, \mathbf{s}_i^{t,\tau_s}\right), \mathbf{B}_i^{t,\tau-1}, \eta\right)$
    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 $\mathbf{B}$.

...

The global objective can be rewritten as

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

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:

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

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

$$\hat{\mathbf{S}}^{t+1} = \text{normalize}\left(\mathbf{S}^{t+1}\right) \tag{4}$$

$$\bar{\mathbf{B}} = \hat{\mathbf{B}}^t - \frac{\eta}{rnm}\left((\mathcal{A}^t)^{\dagger}\mathcal{A}^t(\mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} + \hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top})\right)^{\top}\mathbf{W}_{\mathcal{I}^t}^{t+1\top}, \tag{5}$$

$$\hat{\mathbf{B}}^{t+1}, \mathbf{R}^{t+1} = \text{QR}(\bar{\mathbf{B}}^t). \tag{6}$$

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

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 $\hat{\mathbf{S}}^{t\top}$, $\mathcal{S}_l^*$ denote the l-th column of $\hat{\mathbf{S}}^{*\top}$ 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

$$\begin{aligned}
\mathbf{0} &= \nabla_{\mathcal{W}_p}\widetilde{F}\left(\hat{\mathbf{B}}^t, \mathbf{W}^{t+1}, \hat{\mathbf{S}}^t\right) \\
&= \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle\mathbf{A}_{i,j}, \mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} + \hat{\mathbf{S}}^{t\top} - \hat{\mathbf{S}}^{*\top}\rangle\right)\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t \\
&= \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle\mathbf{A}_{i,j}, \mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top}\rangle + \langle\mathbf{A}_{i,j}, \hat{\mathbf{S}}^{t\top} - \hat{\mathbf{S}}^{*\top}\rangle\right)\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t \\
&= \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\sum_{q=1}^{k}\hat{\mathbf{b}}_q^{t\top}\mathbf{A}_{i,j}^{\top}\mathcal{W}_q^{t+1} - \sum_{q=1}^{k}\hat{\mathbf{b}}_q^{*\top}\mathbf{A}_{i,j}^{\top}\mathcal{W}_q^{*} + \sum_{l=1}^{d}\mathbf{e}_l^{\top}\mathbf{A}_{i,j}^{\top}\mathcal{S}_l^t - \sum_{l=1}^{d}\mathbf{e}_l^{\top}\mathbf{A}_{i,j}^{\top}\mathcal{S}_l^*\right)\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t,
\end{aligned} \tag{7}$$

which means

$$
\begin{aligned}
&\frac{1}{m}\sum_{q=1}^{k}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t\hat{\mathbf{b}}_q^{t\top}\mathbf{A}_{i,j}^\top\right)\mathcal{W}_q^{t+1}\\
=&\frac{1}{m}\sum_{q=1}^{k}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t\hat{\mathbf{b}}_q^{*\top}\mathbf{A}_{i,j}^\top\right)\mathcal{W}_q^{*}+\frac{1}{m}\sum_{l=1}^{d}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t\mathbf{e}_l^{\top}\mathbf{A}_{i,j}^\top\right)\left(\mathcal{S}_l^*-\mathcal{S}_l^t\right).
\end{aligned}
\tag{8}
$$

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,

$$
\mathbf{G} := \begin{bmatrix}\mathbf{G}_{11} & \cdots & \mathbf{G}_{1k}\\ \vdots & \ddots & \vdots\\ \mathbf{G}_{k1} & \cdots & \mathbf{G}_{kk}\end{bmatrix}, \mathbf{E} := \begin{bmatrix}\mathbf{E}_{11} & \cdots & \mathbf{E}_{1d}\\ \vdots & \ddots & \vdots\\ \mathbf{E}_{k1} & \cdots & \mathbf{E}_{kd}\end{bmatrix}. \tag{9}
$$

Then we define $\widetilde{\mathcal{W}}^{t+1} := \mathrm{vec}(\mathbf{W}^{t+1\top})\in\mathbb{R}^{rnk}$, $\widetilde{\mathcal{W}}^{*} := \mathrm{vec}(\mathbf{W}^{*\top})\in\mathbb{R}^{rnk}$, $\widetilde{\mathcal{S}}^{t} := \mathrm{vec}(\hat{\mathbf{S}}^{t\top})\in\mathbb{R}^{rnd}$ and $\widetilde{\mathcal{S}}^{*} := \mathrm{vec}(\hat{\mathbf{S}}^{*\top})\in\mathbb{R}^{rnd}$. From (8) we reach,

$$
\begin{aligned}
\widetilde{\mathcal{W}}^{t+1} &= \mathbf{G}^{-1}\mathbf{C}\widetilde{\mathcal{W}}^{*} + \mathbf{G}^{-1}\mathbf{E}\left(\widetilde{\mathcal{S}}^{*}-\widetilde{\mathcal{S}}^{t}\right)\\
&= \mathbf{D}\widetilde{\mathcal{W}}^{*} - \mathbf{G}^{-1}\left(\mathbf{G}\mathbf{D}-\mathbf{C}\right)\widetilde{\mathcal{W}}^{*} + \mathbf{G}^{-1}\mathbf{E}\left(\widetilde{\mathcal{S}}^{*}-\widetilde{\mathcal{S}}^{t}\right),
\end{aligned}
\tag{10}
$$

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

$$
\begin{aligned}
\mathbf{G}_{pq} &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p\hat{\mathbf{b}}_q^\top\mathbf{A}_{i,j}^\top\\
&= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top\hat{\mathbf{b}}_p\hat{\mathbf{b}}_q^\top\mathbf{x}_i^j\mathbf{e}_i^\top,
\end{aligned}
\tag{11}
$$

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

$$
\left(\mathbf{G}_{pq}\right)_{ii} = \frac{1}{m}\sum_{j=1}^{m}\left(\mathbf{x}_i^j\right)^\top\hat{\mathbf{b}}_p\hat{\mathbf{b}}_q^\top\mathbf{x}_i^j = \hat{\mathbf{b}}_p^\top\left(\frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_i^j\left(\mathbf{x}_i^j\right)^\top\right)\hat{\mathbf{b}}_q. \tag{12}
$$

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 $\mathbf{D}_{pq} = \langle\hat{\mathbf{b}}_p,\hat{\mathbf{b}}_q^*\rangle\mathbf{I}_{rn}$ is also diagonal, then we define

$$
\mathbf{G}^i := \left[\hat{\mathbf{b}}_p^\top\mathbf{\Pi}^i\hat{\mathbf{b}}_q\right]_{1\le p,q\le k+d} = \hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}}, \qquad \mathbf{C}^i := \left[\hat{\mathbf{b}}_p^\top\mathbf{\Pi}^i\hat{\mathbf{b}}_q^*\right]_{1\le p,q\le k+d} = \hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}}^*, \tag{13}
$$

$$\mathbf{D}^i := \left[\langle \hat{\mathbf{b}}_p, \hat{\mathbf{b}}_q^* \rangle\right]_{1\leq p,q\leq k+d} = \hat{\mathbf{B}}^\top \hat{\mathbf{B}}^*, \qquad \mathbf{E}^i := \left[\hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \mathbf{e}_l\right]_{1\leq p\leq k, 1\leq l\leq d} = \hat{\mathbf{B}}^\top \mathbf{\Pi}^i, \tag{14}$$

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 $\mathbf{G}^{-1}\left(\mathbf{GD} - \mathbf{C}\right)\widetilde{\mathcal{W}}^*$ in (10) into i vectors, defined as

$$\mathbf{f}_i := \left(\mathbf{G}^i\right)^{-1}\left(\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\right)\mathbf{w}_i^*, \tag{15}$$

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

$$\mathbf{h}_i = \left(\mathbf{G}^i\right)^{-1}\mathbf{E}^i\left(\hat{\mathbf{s}}_i^* - \hat{\mathbf{s}}_i^t\right), \tag{16}$$

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 (10) we have

$$\begin{aligned}\mathbf{w}_i^{t+1} &= \mathbf{D}^i\mathbf{w}_i^* - \left(\mathbf{G}^i\right)^{-1}\left(\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\right)\mathbf{w}_i^* + \left(\mathbf{G}^i\right)^{-1}\mathbf{E}^i\left(\hat{\mathbf{s}}_i^* - \hat{\mathbf{s}}_i^t\right) \\ &= \hat{\mathbf{B}}^\top\hat{\mathbf{B}}^*\mathbf{w}_i^* - \mathbf{f}_i + \mathbf{h}_i.\end{aligned} \tag{17}$$

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

$$\mathbf{W}^{t+1} = \hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F} + \mathbf{H}, \tag{18}$$

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

$$\begin{aligned}\mathbf{0} &= \nabla_{\mathbf{S}}\mathbf{\Phi}\left(\hat{\mathbf{B}}^t, \mathbf{W}^t, \mathbf{S}^{t+1}\right) \\ &= \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{A}_{i,j}, \mathbf{W}^{t\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} + \mathbf{S}^{t+1\top} - \hat{\mathbf{S}}^{*\top}\rangle\right)\mathbf{A}_{i,j}\end{aligned} \tag{19}$$

Then we reach

$$\mathbf{S}^{t+1} = \hat{\mathbf{S}}^* + \hat{\mathbf{B}}^*\mathbf{W}^* - \hat{\mathbf{B}}^t\mathbf{W}^t. \tag{20}$$

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*

$$\left\|\mathbf{G}^{-1}\right\|_2 \leq \frac{1}{1-\delta_k} \tag{21}$$

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*

$$\left\|\left(\mathbf{GD}-\mathbf{C}\right)\widetilde{\mathcal{W}}^*\right\|_2 \leq \delta_k \left\|\mathbf{W}^*\right\|_2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \tag{22}$$

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*

$$\left\|\mathbf{F}\right\|_{\mathrm{F}} \leq \frac{\delta_k}{1-\delta_k}\left\|\mathbf{W}^*\right\|_2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \tag{23}$$

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

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

Lemma 4 *Let $\delta_k = c\frac{k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}, \delta_d = c_3\frac{\sqrt{d\log(rn)}}{\sqrt{m}}, \delta = \frac{\delta_d}{1-\delta_k}$ for some absolute constant c, c_2, then*

$$\frac{1}{\sqrt{rn}}\left\|\mathbf{H}\right\|_2 \leq (1+\delta)\left\|\hat{\mathbf{S}}^* - \hat{\mathbf{S}}^t\right\|_2 \tag{24}$$

with probability at least $1-e^{-120k^3\log(rn)}$.

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

$$\mathbf{h}_i = \left(\mathbf{G}^i\right)^{-1}\mathbf{E}^i\left(\hat{\mathbf{s}}_i^* - \hat{\mathbf{s}}_i^t\right) = \hat{\mathbf{B}}^{t\top}\left(\hat{\mathbf{s}}_i^* - \hat{\mathbf{s}}_i^t\right) - \left(\mathbf{G}^i\right)^{-1}\left(\mathbf{G}^i\hat{\mathbf{B}}^{t\top} - \mathbf{E}^i\right)\left(\hat{\mathbf{s}}_i^* - \hat{\mathbf{s}}_i^t\right), \tag{25}$$

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

$$\mathbf{G}^i\hat{\mathbf{B}}^{t\top} - \mathbf{E}^i = \hat{\mathbf{B}}^{t\top}\left(\frac{d}{m}\mathbf{X}_i^\top\mathbf{X}_i\right)\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right). \tag{26}$$

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:

$$\mathbf{u}_j = \frac{1}{\sqrt{m}}\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\mathbf{x}_i^j, \;\; \mathbf{v}_j = \frac{1}{\sqrt{m}}\hat{\mathbf{B}}^{t\top}\mathbf{x}_i^j. \tag{27}$$

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

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

$$
\begin{aligned}
&= 2 \max_{\mathbf{z}\in\mathcal{N}_d,\mathbf{y}\in\mathcal{N}_k} \mathbf{z}^\top \left(\sum_{j=1}^m \mathbf{u}_j\mathbf{v}_j^\top\right)\mathbf{y} \\
&= 2 \max_{\mathbf{z}\in\mathcal{N}_d,\mathbf{y}\in\mathcal{N}_k} \sum_{j=1}^m \langle \mathbf{z},\mathbf{u}_j\rangle\langle \mathbf{v}_j,\mathbf{y}\rangle. \qquad (28)
\end{aligned}
$$

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

$$
\mathbb{E}[\langle \mathbf{z},\mathbf{u}_j\rangle\langle \mathbf{v}_j,\mathbf{y}\rangle] = \mathbb{E}[\mathbf{z}^\top\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top}-\mathbf{I}_d\right)\frac{d}{m}\mathbf{X}_i^\top\mathbf{X}_i\hat{\mathbf{B}}^t] = 0. \qquad (29)
$$

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

$$
\mathbb{P}\left(\sum_{j=1}^m \langle \mathbf{z},\mathbf{u}_j\rangle\langle \mathbf{v}_j,\mathbf{y}\rangle \geq s\right) \leq e^{-c'm\min\left(s^2,s\right)}. \qquad (30)
$$

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

$$
\mathbb{P}\left(\left\|\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top}-\mathbf{I}_d\right)\left(\frac{1}{m}\mathbf{X}_i^\top\mathbf{X}_i\right)\hat{\mathbf{B}}^t\right\|_2 \geq 2s\right) \leq 9^{d+k}e^{-c'm\min\left(s^2,s\right)}. \qquad (31)
$$

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

$$
\mathbb{P}\left(\left\|\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top}-\mathbf{I}_d\right)\left(\frac{d}{m}\mathbf{X}_i^\top\mathbf{X}_i\right)\hat{\mathbf{B}}^t\right\|_2 \geq 2\max\left(\varepsilon,\varepsilon^2\right)\right) \leq 9^{d+k}e^{-c'm\varepsilon^2}. \qquad (32)
$$

Further, let $\varepsilon=\sqrt{\frac{c_2 d\log(rn)}{m}}$ for some constant c_2. Then conditioned on $\varepsilon\leq 1$, we have

$$
\mathbb{P}\left(\left\|\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top}-\mathbf{I}_d\right)\left(\frac{1}{m}\mathbf{X}_i^\top\mathbf{X}_i\right)\hat{\mathbf{B}}^t\right\|_2 \geq c_3\sqrt{\frac{d\log(rn)}{m}}\right) \leq 9^{d+k}e^{-c_4 d\log(rn)} \leq e^{-110d\log(rn)}, \qquad (33)
$$

for a large enough constant c_1. According to (25),

$$
\begin{aligned}
\|\mathbf{h}_i\|_2 &\leq \left\|\hat{\mathbf{B}}^{t\top}\right\|_2\left\|\hat{\mathbf{s}}_i^*-\hat{\mathbf{s}}_i^t\right\|_2 + \left\|\left(\mathbf{G}^i\right)^{-1}\right\|_2\left\|\mathbf{G}^i\hat{\mathbf{B}}^{t\top}-\mathbf{E}^i\right\|_2\left\|\hat{\mathbf{s}}_i^*-\hat{\mathbf{s}}_i^t\right\|_2 \qquad (34)\\
&= \left(1+\left\|\left(\mathbf{G}^i\right)^{-1}\right\|_2\left\|\mathbf{G}^i\hat{\mathbf{B}}^{t\top}-\mathbf{E}^i\right\|_2\right)\left\|\hat{\mathbf{s}}_i^*-\hat{\mathbf{s}}_i^t\right\|_2.
\end{aligned}
$$

From (33), we know that

$$
\mathbb{P}\left(\left\|\mathbf{G}^i\hat{\mathbf{B}}^{t\top}-\mathbf{E}^i\right\|_2 \geq \delta_d\right) \leq e^{-110d\log(rn)}, \qquad (35)
$$

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

$$\mathbb{P}\left(\left\|\left(\mathbf{G}^{i}\right)^{-1}\right\|_{2} \geq \frac{1}{1-\delta_{k}}\right) \leq e^{-121 k^{3} \log (rn)} \tag{36}$$

Therefore, we obtain

$$\|\mathbf{h}_{i}\|_{2} \leq (1+\delta)\left\|\hat{\mathbf{s}}_{i}^{*}-\hat{\mathbf{s}}_{i}^{t}\right\|_{2} \tag{37}$$

with probability at least $1-e^{-110 d \log (rn)}-e^{-121 k^{3} \log (rn)}$. Finally we take a union bound over $i \in [rn]$, leading to

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{rn}\|\mathbf{H}\|_{2}^{2} \geq (1+\delta)^{2}\left\|\mathbf{S}^{*}-\mathbf{S}^{t}\right\|_{2}^{2}\right) & \leq \mathbb{P}\left(\frac{1}{rn} \sum_{i=1}^{rn}\|\mathbf{h}_{i}\|_{2}^{2} \geq (1+\delta)^{2}\left\|\mathbf{S}^{*}-\mathbf{S}^{t}\right\|_{2}^{2}\right) \\
& \leq rn \mathbb{P}\left(\|\mathbf{h}_{1}\|_{2}^{2} \geq (1+\delta)^{2}\left\|\mathbf{S}^{*}-\mathbf{S}^{t}\right\|_{2}^{2}\right) \\
& \leq rn \mathbb{P}\left(\|\mathbf{h}_{1}\|_{2}^{2} \geq (1+\delta)^{2}\left\|\mathbf{s}_{1}^{*}-\mathbf{s}_{1}^{t}\right\|_{2}^{2}\right) && (38) \\
& \leq e^{-120 k^{3} \log (rn)} && (39)
\end{aligned}$$

and thus completing the proof. □

Lemma 5 *Let $\delta' = c\sqrt{\frac{d+k}{rnm}}$ for some absolute constant c. Then for any t,*

$$\frac{1}{rn}\left\|\left(\frac{1}{m} \mathcal{A}^{\dagger} \mathcal{A}\left(\mathbf{Q}^{t \top}\right)-\mathbf{Q}^{t \top}\right)^{\top} \mathbf{W}^{t+1 \top}\right\|_{2} \leq \delta' \Delta^{t} \tag{40}$$

with probability at least ..., where Δ^{t} will be given in the following proof.

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

$$\begin{aligned}
\mathbf{q}_{i} & =\hat{\mathbf{B}}^{t} \mathbf{w}_{i}^{t+1}-\hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}+\hat{\mathbf{s}}_{i}^{t+1}-\hat{\mathbf{s}}_{i}^{*} \\
& =\hat{\mathbf{B}}^{t} \hat{\mathbf{B}}^{t \top} \hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}-\hat{\mathbf{B}}^{t} \mathbf{f}_{i}-\hat{\mathbf{B}}^{t} \mathbf{h}_{i}-\hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}+\hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}-\hat{\mathbf{B}}^{t} \mathbf{w}_{i}^{t} \\
& =\left(\hat{\mathbf{B}}^{t} \hat{\mathbf{B}}^{t \top}-\mathbf{I}_{d}\right) \hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}-\hat{\mathbf{B}}^{t} \mathbf{f}_{i}-\hat{\mathbf{B}}^{t} \mathbf{h}_{i}+\hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}-\hat{\mathbf{B}}^{t} \mathbf{w}_{i}^{t}
\end{aligned} \tag{41}$$

Thus,

$$\begin{aligned}
\|\mathbf{q}_{i}\|_{2} & =\left\|\left(\hat{\mathbf{B}}^{t} \hat{\mathbf{B}}^{t \top}-\mathbf{I}_{d}\right) \hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}-\hat{\mathbf{B}}^{t} \mathbf{f}_{i}-\hat{\mathbf{B}}^{t} \mathbf{h}_{i}+\hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}-\hat{\mathbf{B}}^{t} \mathbf{w}_{i}^{t}\right\|_{2} \\
& \leq\left\|\left(\hat{\mathbf{B}}^{t} \hat{\mathbf{B}}^{t \top}-\mathbf{I}_{d}\right) \hat{\mathbf{B}}^{*}\right\|_{2}\|\mathbf{w}_{i}^{*}\|_{2}+\|\mathbf{f}_{i}\|_{2}+\|\mathbf{h}_{i}\|_{2}+\left\|\hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}-\hat{\mathbf{B}}^{t} \mathbf{w}_{i}^{t}\right\|_{2} \\
& \leq 2 \sqrt{k} \operatorname{dist}\left(\hat{\mathbf{B}}^{t}, \hat{\mathbf{B}}^{*}\right)+(1+\delta)\left\|\mathbf{s}_{i}^{*}-\mathbf{s}_{i}^{t}\right\|_{2}+\left\|\hat{\mathbf{B}}^{*} \mathbf{w}_{i}^{*}-\hat{\mathbf{B}}^{t} \mathbf{w}_{i}^{t}\right\|_{2}
\end{aligned} \tag{42}$$

$$\leq 2\sqrt{k}\,\mathrm{dist}\left(\hat{\mathbf{B}}^t,\hat{\mathbf{B}}^*\right)+(1+\delta)\left\|\mathbf{S}^*-\mathbf{S}^t\right\|_{\mathrm{F}}+\left\|\hat{\mathbf{B}}^*\mathbf{W}^*-\hat{\mathbf{B}}^t\mathbf{W}^t\right\|_{\mathrm{F}}, \tag{43}$$

where (42) holds with probability at least $1-e^{-110k^2\log(rn)}$, by combining equation (44) in (Collins et al., 2021) and (37), conditioned on $\delta_k \leq \frac{1}{2}$. Similarly, combining equation (45) and (37), conditioned on $\delta_k \leq \frac{1}{2}$, we have

$$\begin{aligned}\left\|\mathbf{w}_i^{t+1}\right\|_2 &\leq \left\|\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{w}_i^*\right\|_2+\|\mathbf{f}_i\|_2+\|\mathbf{h}_i\|_2\\ &\leq 2\sqrt{k}+(1+\delta)\left\|\hat{\mathbf{s}}_i^*-\hat{\mathbf{s}}_i^t\right\|_2 \qquad (44)\\ &\leq (4+2\delta)\sqrt{k}, \qquad (45)\end{aligned}$$

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

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

$$\mathcal{E}:=\bigcap_{i=1}^{rn}\left\{\|\mathbf{q}_i\|_2\leq 2\sqrt{k}\,\mathrm{dist}\left(\hat{\mathbf{B}}^t,\hat{\mathbf{B}}^*\right)+(1+\delta)\left\|\Delta_{\mathbf{S}}^t\right\|_{\mathrm{F}}+\left\|\Delta_{\mathbf{BW}}^t\right\|_{\mathrm{F}}\cap\left\|\mathbf{w}_i^{t+1}\right\|_2\leq 2\sqrt{k}+(1+\delta)\left\|\Delta_{\mathbf{S}}^t\right\|_{\mathrm{F}}\right\}, \tag{46}$$

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

$$\begin{aligned}\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right)-\mathbf{Q}^{t\top}&=\frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left\langle\mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top,\mathbf{Q}^{t\top}\right\rangle\mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top-\mathbf{Q}^{t\top}\\ &=\frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\langle\mathbf{x}_i^j,\mathbf{q}_i\rangle\mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top-\mathbf{Q}^{t\top}, \qquad (47)\end{aligned}$$

further, we have

$$\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right)-\mathbf{Q}^{t\top}\right)^\top\mathbf{W}^{t+1\top}=\frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle\mathbf{x}_i^j,\mathbf{q}_i\rangle\mathbf{x}_i^j\mathbf{w}_i^\top-\mathbf{q}_i\mathbf{w}_i^\top\right). \tag{48}$$

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

$$\begin{aligned}&\left\|\frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle\mathbf{x}_i^j,\mathbf{q}_i\rangle\mathbf{x}_i^j\mathbf{w}_i^\top-\mathbf{q}_i\mathbf{w}_i^\top\right)\right\|_2\\ &\leq 2\max_{\mathbf{z}\in\mathcal{N}_d,\mathbf{y}\in\mathcal{N}_k}\mathbf{z}^\top\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle\mathbf{x}_i^j,\mathbf{q}_i\rangle\mathbf{x}_i^j\mathbf{w}_i^\top-\frac{1}{rnm}\mathbf{q}_i\mathbf{w}_i^\top\right)\right)\mathbf{y}\end{aligned}$$

$$= 2 \max_{\mathbf{z}\in\mathcal{N}_d,\mathbf{y}\in\mathcal{N}_k} \sum_{i=1}^{rn}\sum_{j=1}^{m} \left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\langle \mathbf{z}, \mathbf{x}_i^j\rangle\langle \mathbf{w}_i, \mathbf{y}\rangle - \frac{1}{rnm}\langle \mathbf{z}, \mathbf{q}_i\rangle\langle \mathbf{w}_i, \mathbf{y}\rangle \right) \tag{49}$$

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 $c\left\|\mathbf{z}\right\|_2 = c$ 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 $\left\|\mathbf{q}_i\right\|_2$. 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 $c\left\|\mathbf{q}_i\right\|_2$, 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

$$\begin{aligned}
\frac{c}{rnm}\left\|\mathbf{q}_i\right\|_2\langle \mathbf{w}_i, \mathbf{y}\rangle &\le \frac{c}{rnm}\left\|\mathbf{q}_i\right\|_2\left\|\mathbf{w}_i\right\|_2 \\
&\le \frac{c}{rnm}\left(2\sqrt{k}\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + (1+\delta)\left\|\Delta_{\mathbf{S}}^t\right\|_{\mathrm{F}} + \left\|\Delta_{\mathbf{BW}}^t\right\|_{\mathrm{F}}\right)\left(2\sqrt{k} + (1+\delta)\left\|\Delta_{\mathbf{S}}^t\right\|_{\mathrm{F}}\right) \quad (50) \\
&\le \frac{c}{rnm}\left(4k\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + 2\sqrt{k}(1+\delta)\left\|\Delta_{\mathbf{S}}^t\right\|_{\mathrm{F}} + 2\sqrt{k}\left\|\Delta_{\mathbf{BW}}^t\right\|_{\mathrm{F}}\right) \\
&\quad + \frac{c}{rnm}\left(2\sqrt{k}(1+\delta)\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right)\left\|\Delta_{\mathbf{S}}^t\right\|_{\mathrm{F}} + (1+\delta)^2\left\|\Delta_{\mathbf{S}}^t\right\|_{\mathrm{F}}^2 + (1+\delta)\left\|\Delta_{\mathbf{S}}^t\right\|_{\mathrm{F}}\left\|\Delta_{\mathbf{BW}}^t\right\|_{\mathrm{F}}\right) \quad (51) \\
&:= \frac{c}{rnm}\Delta^t. \quad (52)
\end{aligned}$$

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

$$\mathbb{P}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\langle \mathbf{z}, \mathbf{x}_i^j\rangle\langle \mathbf{w}_i, \mathbf{y}\rangle - \frac{1}{rnm}\langle \mathbf{z}, \mathbf{q}_i\rangle\langle \mathbf{w}_i, \mathbf{y}\rangle\right) \ge s\right) \le \exp\left(-c_2 rnm \min\left(\frac{s^2}{(\Delta^t)^2}, \frac{s}{\Delta^t}\right)\right). \tag{53}$$

Take union bound over all $\mathbf{z}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k$,

$$\mathbb{P}\left(\left\|\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top}\right)\mathbf{W}^{t+1\top}\right\|_2 \ge 2s \,\middle|\, \mathcal{E}\right) \le 9^{d+k}\exp\left(-c_2 rnm \min\left(\frac{s^2}{(\Delta^t)^2}, \frac{s}{\Delta^t}\right)\right). \tag{54}$$

Let $\frac{s}{\Delta^t} = \max(\varepsilon, \varepsilon^2)$ for some $\varepsilon > 0$, then $\varepsilon^2 = \min\left(\frac{s^2}{(\Delta^t)^2}, \frac{s}{\Delta^t}\right)$. Further, let $\varepsilon = \sqrt{\frac{113(d+k)}{c_2 rnm}}$, and conditioned on $\varepsilon \le 1$, we obtain

$$\mathbb{P}\left(\left\|\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top}\right)\mathbf{W}^{t+1\top}\right\|_2 \ge c_3\Delta^t\sqrt{\frac{d+k}{rnm}} \,\middle|\, \mathcal{E}\right) \le e^{-110(d+k)}. \tag{55}$$

$\square$

1.3 Main Result

Recall that $\mathbf{Q}^{t\top} = \mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} + \hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top}$, plugging this into (5), and without losing generality, we drop the subscripts of $\mathcal{I}^t$ and obtain

$$
\begin{aligned}
\bar{\mathbf{B}}^{t+1} &= \hat{\mathbf{B}}^t - \frac{\eta}{rnm}\left(\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top})\right)^\top \mathbf{W}^{t+1\top} \\
&= \hat{\mathbf{B}}^t - \frac{\eta}{rn}\mathbf{Q}^t\mathbf{W}^{t+1\top} - \frac{\eta}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}. \qquad (56)
\end{aligned}
$$

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

$$
\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^{t+1} = \left(\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\mathbf{Q}^t\mathbf{W}^{t+1\top} - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}\right)(\mathbf{R}^{t+1})^{-1}. \qquad (57)
$$

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

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

plugging this into (57) then we reach

$$
\begin{aligned}
\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^{t+1} = \Bigg(&\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t\left(\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right) + \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\hat{\mathbf{S}}^* - \hat{\mathbf{S}}^{t+1}\right)\mathbf{W}^{t+1\top} \\
&-\frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}\Bigg)(\mathbf{R}^{t+1})^{-1}. \qquad (58)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{dist}(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*) &= \left\|\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^{t+1}\right\|_2 \\
&\le \left\|\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t\left(\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right)\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \\
&\quad + \frac{\eta}{rn}\left\|\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}(\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)^\top \mathbf{W}^{t+1\top}\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \\
&\quad + \frac{\eta}{rn}\left\|\hat{\mathbf{B}}_\perp^{*\top}\left(\hat{\mathbf{S}}^* - \hat{\mathbf{S}}^{t+1}\right)\mathbf{W}^{t+1\top}\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2. \qquad (59)
\end{aligned}
$$

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

$$
\begin{aligned}
\left\|\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t\left(\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right)\right\|_2 &\le \left\|\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t\right\|_2 \left\|\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right\|_2 \\
&\le \mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right)\left\|\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right\|_2. \qquad (60)
\end{aligned}
$$

To bound the term of $\left\|\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right\|_2$, we assume that $\frac{1}{\sqrt{rn}}\mathbf{W}^{t+1}$ has non-zero minimum singular value, defined as σ_{min}^{t+1}. Then as long as $\eta \le (\sigma_{min}^{t+1})^2$, we have

$$\left\|\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right\|_2 = 1 - \eta(\sigma_{min}^{t+1})^2.$$

Then we consider the term of

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

$$\begin{aligned}
\mathbf{R}^{t+1\top}\mathbf{R}^{t+1} &= \bar{\mathbf{B}}^{t+1\top}\bar{\mathbf{B}}^{t+1} \\
&= \hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^{t} - \frac{\eta}{rn}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\right) + \frac{\eta^2}{(rn)^2}\mathbf{W}^{t+1}\mathbf{U}^{t}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} \\
&= \mathbf{I}_k - \frac{\eta}{rn}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\right) + \frac{\eta^2}{(rn)^2}\mathbf{W}^{t+1}\mathbf{U}^{t}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}. \qquad (61)
\end{aligned}$$

Using Weyl's Inequality, we reach

$$\begin{aligned}
\sigma^2_{\min}\left(\mathbf{R}^{t+1}\right) &\ge 1 - \frac{\eta}{rn}\lambda_{\max}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\right) + \frac{\eta^2}{(rn)^2}\lambda_{\min}\left(\mathbf{W}^{t+1}\mathbf{U}^{t}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}\right) \\
&\ge 1 - \frac{\eta}{rn}\lambda_{\max}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\right) \qquad (62)
\end{aligned}$$

where (62) holds since $\mathbf{W}^{t+1}\mathbf{U}^{t}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}$ is positive semi-definite. Further,

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

When considering the first term, we have

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^{t}\mathbf{z} \le \frac{2\eta}{rn}\left\|\mathbf{W}^{t+1}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)\right\|_2\left\|\hat{\mathbf{B}}^{t}\right\|_2 \le 2\eta\sqrt{\frac{d+k}{rnm}}\Delta^t \qquad (64)$$

Then we consider the second term in (63),

$$\begin{aligned}
\max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{Q}^{t\top}\hat{\mathbf{B}}^{t}\mathbf{z} \le &\max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^{*}\mathbf{W}^{*} - \mathbf{F}\right)\left(\mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top}\right)\hat{\mathbf{B}}^{t}\mathbf{z} \\
&+ \max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\frac{2\eta}{rn}\mathbf{z}^\top\left(\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^{*}\mathbf{W}^{*} - \mathbf{F}\right)\left(\hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top}\right) + \mathbf{H}\mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^{t}\mathbf{z}
\end{aligned} \qquad (65)$$

As for the first term in (65), from equation (81) in (Collins et al., 2021) we have

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{2\eta}{rn}\mathbf{z}^\top \left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F}\right)\left(\mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top}\right)\hat{\mathbf{B}}^t\mathbf{z} \leq 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}^2_{\max,*}. \tag{66}$$

As for the second term in (65),

$$\begin{aligned}
&\frac{2\eta}{rn}\left\|\left(\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F}\right)\left(\hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top}\right) + \mathbf{H}\mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^t\right\|_2 \\
&\leq \frac{2\eta}{rn}\left\|\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{W}^* - \mathbf{F}\right\|_2\left\|\hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top}\right\|_2 + \frac{2\eta}{rn}\left\|\mathbf{H}\mathbf{Q}^{t\top}\right\|_2 \\
&\leq \frac{4\eta}{rn}\|\mathbf{W}^*\|_2\left\|\Delta^t_{\mathbf{BW}}\right\|_2 + 2\eta\frac{1}{\sqrt{rn}}\|\mathbf{H}\|_2\frac{1}{\sqrt{rn}}\left\|\mathbf{Q}^t\right\|_2 \qquad (67)\\
&\leq \frac{4\eta}{\sqrt{rn}}\bar{\sigma}_{\max,*}\left\|\Delta^t_{\mathbf{BW}}\right\|_2 + 4\eta(1+\delta)\sqrt{k}\,\mathrm{dist}\left(\hat{\mathbf{B}}^t,\mathbf{B}^*\right)\left\|\Delta^t_{\mathbf{S}}\right\|_{\mathrm{F}} + 4\eta(1+\delta)\left\|\Delta^t_{\mathbf{S}}\right\|^2_{\mathrm{F}} + 2\eta(1+\delta)\left\|\Delta^t_{\mathbf{BW}}\right\|_{\mathrm{F}}\left\|\Delta^t_{\mathbf{S}}\right\|_{\mathrm{F}} \qquad (68)
\end{aligned}$$

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.

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

A Proofs