AdaptPFL: Personalized Federated Learning with xxx Adaptation

February 7, 2025

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 $\mathbf{B}$.

...

The global objective can be rewritten as

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

where $\mathbf{Y} = \mathcal{A}((1-\alpha)\mathbf{W}_{\mathcal{I}}^{*\top}\hat{\mathbf{B}}^{*\top} + \alpha\hat{\mathbf{S}}_{\mathcal{I}}^{*\top}) \in \mathbb{R}^{rnm}$. Then we give the update rules of our algorithm:

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

$$\bar{\mathbf{S}}^{t+1} = \hat{\mathbf{B}}_{\perp}^{t}\hat{\mathbf{B}}_{\perp}^{t\top} \left(\widetilde{\mathbf{S}}^{t+1} \right), \tag{3}$$

$$\mathbf{S}^{t+1} = (1-\lambda_S)\mathbf{S}^t + \lambda_S \bar{\mathbf{S}}^{t+1}, \tag{4}$$

$$\bar{\mathbf{W}}^{t+1} = \arg\min_{\mathbf{W}\in\mathbb{R}^{k\times rn}} \frac{1}{2rnm} \left\| \mathcal{A}^t \left((1-\alpha) \left(\mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} - \mathbf{W}^{\top}\hat{\mathbf{B}}^{t\top} \right) + \alpha \left(\mathbf{S}^{*\top} - \bar{\mathbf{S}}^{t+1\top} \right) \right) \right\|_2^2, \tag{5}$$

$$\mathbf{W}^{t+1} = (1-\lambda_W)\mathbf{W}^t + \gamma\bar{\mathbf{W}}^{t+1}, \tag{6}$$

$$\bar{\mathbf{B}}^{t+1} = \hat{\mathbf{B}}^t - \frac{\eta}{rnm} \left((\mathcal{A}^t)^{\dagger}\mathcal{A}^t \left((1-\alpha) \left(\mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} \right) + \alpha \left(\mathbf{S}^{t+1\top} - \mathbf{S}^{*\top} \right) \right) \right)^{\top} \mathbf{W}_{\mathcal{I}^t}^{t+1\top}, \tag{7}$$

$$\hat{\mathbf{B}}^{t+1}, \mathbf{R}^{t+1} = \mathrm{QR}(\bar{\mathbf{B}}^{t+1}). \tag{8}$$

1.2 Auxiliary Lemmas

We first consider the update for $\mathbf{W}$.

Lemma 1 *Let $\bar{\Delta}_i^t := -\left(\mathbf{G}^i\right)^{-1}\mathbf{E}^i\Delta_i^t$, we have*

$$(1-\alpha)\bar{\mathbf{w}}_i^{t+1} = \hat{\mathbf{B}}^{t\top} \left((1-\alpha)\hat{\mathbf{B}}^*\mathbf{w}_i^* + \alpha\mathbf{s}_i^* \right) + \bar{\Delta}_i^t \tag{9}$$

Proof: According to the update rule of (5), $\mathbf{W}^{t+1}$ minimizes the function of $\widetilde{F}\left(\hat{\mathbf{B}}^t, \mathbf{W}, \bar{\mathbf{S}}^{t+1}\right) := \frac{1}{2rnm} \left\| \mathcal{A} \left((1-\alpha) \left(\mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} - \mathbf{W}^{\top}\hat{\mathbf{B}}^{t\top} \right) + \alpha \left(\mathbf{S}^{*\top} - \bar{\mathbf{S}}^{t+1\top} \right) \right) \right\|_2^2$.

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 $\mathbf{W}^{*\top}$, $\mathcal{S}_l^{t+1}$ denote the l-th column of $\bar{\mathbf{S}}^{t+1\top}$, $\mathcal{S}_l^*$ 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

$$\mathbf{0} = \nabla_{\mathcal{W}_p}\widetilde{F}\left(\hat{\mathbf{B}}^t, \mathbf{W}^{t+1}, \bar{\mathbf{S}}^{t+1}\right)$$

$$
\begin{aligned}
&= \frac{1-\alpha}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^{m} \left(\langle \mathbf{A}_{i,j}, (1-\alpha) \left(\mathbf{W}^{t+1\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} \right) + \alpha \left(\bar{\mathbf{S}}^{t+1\top} - \mathbf{S}^{*\top} \rangle \right) \right) \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \\
&= \frac{1-\alpha}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^{m} \left((1-\alpha) \langle \mathbf{A}_{i,j}, \mathbf{W}^{t+1\top} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top} \hat{\mathbf{B}}^{*\top} \rangle + \alpha \langle \mathbf{A}_{i,j}, \bar{\mathbf{S}}^{t+1\top} - \mathbf{S}^{*\top} \rangle \right) \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \\
&= \frac{1-\alpha}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^{m} \left((1-\alpha) \left(\sum_{q=1}^{k} \hat{\mathbf{b}}_q^{t\top} \mathbf{A}_{i,j}^\top \mathcal{W}_q^{t+1} - \sum_{q=1}^{k} \hat{\mathbf{b}}_q^{*\top} \mathbf{A}_{i,j}^\top \mathcal{W}_q^{*} \right) + \alpha \left(\sum_{l=1}^{d} \mathbf{e}_l^\top \mathbf{A}_{i,j}^\top \mathcal{S}_l^{t+1} - \sum_{l=1}^{d} \mathbf{e}_l^\top \mathbf{A}_{i,j}^\top \mathcal{S}_l^{*} \right) \right) \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t
\end{aligned}
\tag{10}
$$

which means

$$
\begin{aligned}
&\frac{1}{m} \sum_{q=1}^{k} \left(\sum_{i=1}^{rn} \sum_{j=1}^{m} \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \hat{\mathbf{b}}_q^{t\top} \mathbf{A}_{i,j}^\top \right) (1-\alpha) \mathcal{W}_q^{t+1} \\
=& \frac{1}{m} \sum_{q=1}^{k} \left(\sum_{i=1}^{rn} \sum_{j=1}^{m} \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \hat{\mathbf{b}}_q^{*\top} \mathbf{A}_{i,j}^\top \right) (1-\alpha) \mathcal{W}_q^{*} + \frac{1}{m} \sum_{l=1}^{d} \left(\sum_{i=1}^{rn} \sum_{j=1}^{m} \mathbf{A}_{i,j} \hat{\mathbf{b}}_p^t \mathbf{e}_l^\top \mathbf{A}_{i,j}^\top \right) \alpha \left(\mathcal{S}_l^* - \mathcal{S}_l^{t+1} \right).
\end{aligned}
\tag{11}
$$

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,

$$
\mathbf{G} := \begin{bmatrix} \mathbf{G}_{11} & \cdots & \mathbf{G}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{k1} & \cdots & \mathbf{G}_{kk} \end{bmatrix}, \mathbf{E} := \begin{bmatrix} \mathbf{E}_{11} & \cdots & \mathbf{E}_{1d} \\ \vdots & \ddots & \vdots \\ \mathbf{E}_{k1} & \cdots & \mathbf{E}_{kd} \end{bmatrix}. \tag{12}
$$

Then we define $\widetilde{\mathcal{W}}^{t+1} := \mathrm{vec}(\mathbf{W}^{t+1\top}) \in \mathbb{R}^{rnk}$, $\widetilde{\mathcal{W}}^{*} := \mathrm{vec}(\mathbf{W}^{*\top}) \in \mathbb{R}^{rnk}$, $\widetilde{\mathcal{S}}^{t+1} := \mathrm{vec}(\bar{\mathbf{S}}^{t+1\top}) \in \mathbb{R}^{rnd}$ and $\widetilde{\mathcal{S}}^{*} := \mathrm{vec}(\mathbf{S}^{*\top}) \in \mathbb{R}^{rnd}$. From (11) we reach,

$$
\begin{aligned}
(1-\alpha)\widetilde{\mathcal{W}}^{t+1} &= (1-\alpha)\mathbf{G}^{-1}\mathbf{C}\widetilde{\mathcal{W}}^{*} + \alpha \mathbf{G}^{-1}\mathbf{E} \left(\widetilde{\mathcal{S}}^{*} - \widetilde{\mathcal{S}}^{t+1} \right) \\
&= (1-\alpha)\mathbf{D}\widetilde{\mathcal{W}}^{*} - (1-\alpha)\mathbf{G}^{-1} \left(\mathbf{G}\mathbf{D} - \mathbf{C} \right) \widetilde{\mathcal{W}}^{*} + \alpha \mathbf{G}^{-1}\mathbf{E} \left(\widetilde{\mathcal{S}}^{*} - \widetilde{\mathcal{S}}^{t+1} \right),
\end{aligned}
\tag{13}
$$

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

$$
\begin{aligned}
\mathbf{G}_{pq} &= \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^{m} \mathbf{A}_{i,j} \hat{\mathbf{b}}_p \hat{\mathbf{b}}_q^\top \mathbf{A}_{i,j}^\top \\
&= \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^{m} \mathbf{e}_i \left(\mathbf{x}_i^j \right)^\top \hat{\mathbf{b}}_p \hat{\mathbf{b}}_q^\top \mathbf{x}_i^j \mathbf{e}_i^\top,
\end{aligned}
\tag{14}
$$

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

$$(\mathbf{G}_{pq})_{ii} = \frac{1}{m}\sum_{j=1}^{m}\left(\mathbf{x}_i^j\right)^\top \hat{\mathbf{b}}_p\hat{\mathbf{b}}_q^\top \mathbf{x}_i^j = \hat{\mathbf{b}}_p^\top \left(\frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_i^j\left(\mathbf{x}_i^j\right)^\top\right)\hat{\mathbf{b}}_q. \tag{15}$$

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

$$\mathbf{G}^i := \left[\hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \hat{\mathbf{b}}_q\right]_{1\le p,q\le k+d} = \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}, \qquad \mathbf{C}^i := \left[\hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \hat{\mathbf{b}}_q^*\right]_{1\le p,q\le k+d} = \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}^*, \tag{16}$$

$$\mathbf{D}^i := \left[\langle \hat{\mathbf{b}}_p, \hat{\mathbf{b}}_q^*\rangle\right]_{1\le p,q\le k+d} = \hat{\mathbf{B}}^\top \hat{\mathbf{B}}^*, \qquad \mathbf{E}^i := \left[\hat{\mathbf{b}}_p^\top \mathbf{\Pi}^i \mathbf{e}_l\right]_{1\le p\le k, 1\le l\le d} = \hat{\mathbf{B}}^\top \mathbf{\Pi}^i, \tag{17}$$

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 (13) into i vectors, defined as

$$\begin{aligned}
\mathbf{f}_i :=&(1-\alpha)\left(\mathbf{G}^i\right)^{-1}\left(\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\right)\mathbf{w}_i^* &(18)\\
=&(1-\alpha)\left(\mathbf{G}^i\right)^{-1}\left(\hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}}\hat{\mathbf{B}}^\top\hat{\mathbf{B}}^* - \hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}}^*\right)\mathbf{w}_i^* &(19)\\
=&-(1-\alpha)\left(\mathbf{G}^i\right)^{-1}\hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}}_\perp\hat{\mathbf{B}}_\perp^\top\hat{\mathbf{B}}^*\mathbf{w}_i^* &(20)
\end{aligned}$$

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

$$\mathbf{h}_i := \alpha\left(\mathbf{G}^i\right)^{-1}\mathbf{E}^i\left(\mathbf{s}_i^* - \bar{\mathbf{s}}_i^{t+1}\right), \tag{21}$$

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

According to (2), $\widetilde{\mathbf{S}}^{t+1}$ minimizes

$$\mathbf{\Phi}\left(\hat{\mathbf{B}}^t, \bar{\mathbf{W}}^t, \widetilde{\mathbf{S}}\right) := \frac{1}{2rnm}\left\|\mathcal{A}\left((1-\alpha)\left(\mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} - \bar{\mathbf{W}}^{t\top}\hat{\mathbf{B}}^{t\top}\right) + \alpha\left(\mathbf{S}^{*\top} - \mathbf{S}^\top\right)\right)\right\|_2^2 + \frac{\beta}{2}\left\|\widetilde{\mathbf{S}}\right\|_{\mathrm{F}}^2. \tag{22}$$

Then via a similar process from (10) to (32), we can obtain

$$\alpha\widetilde{\mathbf{s}}_i^{t+1} = \left(\mathbf{\Pi}^i + \beta\mathbf{I}_d\right)^{-1}\mathbf{\Pi}^i\left(\alpha\mathbf{s}_i^* + (1-\alpha)\hat{\mathbf{B}}^*\mathbf{w}_i^* - (1-\alpha)\hat{\mathbf{B}}^t\bar{\mathbf{w}}_i^t\right), \tag{23}$$

further, let $\mathbf{s}_i^{t+1\prime} := \dfrac{1}{\alpha}\left(\alpha\mathbf{s}_i^* + (1-\alpha)\hat{\mathbf{B}}^*\mathbf{w}_i^* - (1-\alpha)\hat{\mathbf{B}}^t\bar{\mathbf{w}}_i^t\right)$, we have

$$\alpha\bar{\mathbf{s}}_i^{t+1} = \Delta_i^t + \hat{\mathbf{B}}_\perp^t\hat{\mathbf{B}}_\perp^{t\top}\alpha\mathbf{s}_i^{t+1\prime}, \tag{24}$$

where

$$\Delta_i^t := \hat{\mathbf{B}}_\perp^t \hat{\mathbf{B}}_\perp^{t\top} \left(\left(\mathbf{\Pi}^i + \beta \mathbf{I}_d \right)^{-1} \mathbf{\Pi}^i - \mathbf{I}_d \right) \alpha \mathbf{s}_i^{t+1\prime}. \tag{25}$$

By

$$\begin{aligned}
\alpha(\mathbf{s}_i^* - \bar{\mathbf{s}}_i^{t+1}) &= \alpha(\mathbf{s}_i^* - \hat{\mathbf{B}}_\perp^t \hat{\mathbf{B}}_\perp^{t\top} \widetilde{\mathbf{s}}_i^{t+1}) - \Delta_i^t && (26)\\
&= \alpha \mathbf{s}_i^* - \hat{\mathbf{B}}_\perp^t \hat{\mathbf{B}}_\perp^{t\top} (\alpha \mathbf{s}_i^* + (1-\alpha)\mathbf{B}^* \mathbf{w}_i^* - (1-\alpha)\mathbf{B}^t \mathbf{w}_i^t) - \Delta_i^t && (27)\\
&= \alpha \hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} \mathbf{s}_i^* - (1-\alpha) \hat{\mathbf{B}}_\perp^t \hat{\mathbf{B}}_\perp^{t\top} \mathbf{B}^* \mathbf{w}_i^* - \Delta_i^t && (28)
\end{aligned}$$

we have

$$\begin{aligned}
\mathbf{h}_i &= \left(\mathbf{G}^i\right)^{-1} \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \left(\alpha \hat{\mathbf{B}}^t \hat{\mathbf{B}}^{t\top} \mathbf{s}_i^* - (1-\alpha) \hat{\mathbf{B}}_\perp^t \hat{\mathbf{B}}_\perp^{t\top} \mathbf{B}^* \mathbf{w}_i^* - \Delta_i^t \right) && (29)\\
&= \alpha \hat{\mathbf{B}}^{t\top} \mathbf{s}_i^* + \mathbf{f}_i - \left(\mathbf{G}^i\right)^{-1} \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \Delta_i^t && (30)
\end{aligned}$$

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 (13) we have

$$\begin{aligned}
(1-\alpha)\bar{\mathbf{w}}_i^{t+1} &= (1-\alpha)\mathbf{D}^i \mathbf{w}_i^* - \mathbf{f}_i + \mathbf{h}_i && (31)\\
&= \hat{\mathbf{B}}^\top \left((1-\alpha) \hat{\mathbf{B}}^* \mathbf{w}_i^* + \alpha \mathbf{s}_i^* \right) - \left(\mathbf{G}^i\right)^{-1} \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \Delta_i^t && (32)
\end{aligned}$$

Let $\bar{\Delta}_i^t := -\left(\mathbf{G}^i\right)^{-1} \mathbf{E}^i \Delta_i^t$, and we can rewrite (32) as

$$(1-\alpha)\bar{\mathbf{w}}_i^{t+1} = \hat{\mathbf{B}}^{t\top} \left((1-\alpha) \hat{\mathbf{B}}^* \mathbf{w}_i^* + \alpha \mathbf{s}_i^* \right) + \bar{\Delta}_i^t \tag{33}$$

□

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*

$$\left\| \mathbf{G}^{-1} \right\|_2 \le \frac{1}{1-\delta_k} \tag{34}$$

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

Lemma 3 *Let* $\delta_d = c\frac{\sqrt{d}+k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}$ *for some absolute constant* c*, then*

$$\left\| \mathbf{E}^i \right\|_2 \le 1 + \delta_d \tag{35}$$

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

Lemma 4

$$\|\bar{\Delta}_i^t\| \le \tag{36}$$

Proof: By

$$\widetilde{\mathbf{s}}_i^{t+1} = \arg\min_{\mathbf{s}_i \in \mathbb{R}^d} \frac{1}{2rnm} \left\| \left((1-\alpha) \left(\mathbf{w}_i^{*\top} \hat{\mathbf{B}}^{*\top} - \mathbf{w}_i^{t\top} \hat{\mathbf{B}}^{t\top} \right) + \alpha \left(\mathbf{s}_i^{*\top} - \mathbf{s}_i^\top \right) \right) \mathbf{X}_i^\top \right\|_2^2 + \frac{\beta}{2} \|\mathbf{s}_i\|_2^2 \quad (37)$$

$$= \arg\min_{\mathbf{s}_i \in \mathbb{R}^d} \frac{1}{2rnm} \left\| \alpha \left(\mathbf{s}_i^{t+1\prime\top} - \mathbf{s}_i^\top \right) \mathbf{X}_i^\top \right\|_2^2 + \frac{\beta}{2} \|\mathbf{s}_i\|_2^2 \quad (38)$$

we have

$$\frac{1}{2rnm} \left\| \alpha \left(\mathbf{s}_i^{t+1\prime\top} - \mathbf{s}_i^{t+1\prime\top} \right) \mathbf{X}_i^\top \right\|_2^2 + \frac{\beta}{2} \left\| \mathbf{s}_i^{t+1\prime\top} \right\|_2^2 \geq \min_{\mathbf{s}_i \in \mathbb{R}^d} \frac{1}{2rnm} \left\| \alpha \left(\mathbf{s}_i^{t+1\prime\top} - \mathbf{s}_i^\top \right) \mathbf{X}_i^\top \right\|_2^2 + \frac{\beta}{2} \|\mathbf{s}_i\|_2^2 \quad (39)$$

$$= \frac{1}{2rnm} \left\| \alpha \left(\mathbf{s}_i^{t+1\prime\top} - \widetilde{\mathbf{s}}_i^{t+1} \right) \mathbf{X}_i^\top \right\|_2^2 + \frac{\beta}{2} \left\| \widetilde{\mathbf{s}}_i^{t+1} \right\|_2^2 \quad (40)$$

so we can get

$$\frac{\beta}{2} \left\| \mathbf{s}_i^{t+1\prime\top} \right\|_2^2 \geq \frac{1}{2rnm} \left\| \alpha \left(\mathbf{s}_i^{t+1\prime\top} - \widetilde{\mathbf{s}}_i^{t+1} \right) \mathbf{X}_i^\top \right\|_2^2 \quad (41)$$

$$\geq \frac{1}{2rnm} \left\| \alpha \left(\mathbf{s}_i^{t\prime\top} - \widetilde{\mathbf{s}}_i^{t+1} \right) \right\|_2^2 \sigma_{\min}^2(\mathbf{X}_i) \quad (42)$$

Then,

$$\|\Delta_i^t\|^2 \leq \left\| \alpha \left(\mathbf{s}_i^{t+1\prime\top} - \widetilde{\mathbf{s}}_i^{t+1} \right) \right\|_2^2 \quad (43)$$

$$\leq \frac{2rnm\beta}{\sigma_{\min}^2(\mathbf{X}_i)} \quad (44)$$

Thus,

$$\|\bar{\Delta}_i^t\| = \| \left(\mathbf{G}^i \right)^{-1} \mathbf{E}^i \Delta_i^t \| \quad (45)$$

$$\leq \frac{(1+\delta_d)\sqrt{2rnm\beta}}{(1-\delta_k)\sigma_{\min}(\mathbf{X}_i)} \quad (46)$$

with high probability. □

Lemma 5 *Let $\delta_k' = c_4 k \frac{\sqrt{d}}{\sqrt{rnm}}$ for some absolute constant c_4. Then for any t,*

$$\frac{1}{rn} \left\| \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A} \left(\mathbf{Q}^{t\top} \right) - \mathbf{Q}^{t\top} \right)^\top (1-\alpha) \mathbf{W}^{t+1\top} \right\|_2 \leq \delta_k' \operatorname{dist} \left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) \quad (47)$$

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

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

$$\mathbf{q}_i = (1-\alpha) \left(\hat{\mathbf{B}}^t \mathbf{w}_i^{t+1} - \hat{\mathbf{B}}^* \mathbf{w}_i^* \right) + \alpha \left(\hat{\mathbf{s}}_i^t - \hat{\mathbf{s}}_i^* \right)$$

$$= (1-\alpha)\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{w}_i^* - (1-\alpha)\hat{\mathbf{B}}^t\mathbf{f}_i - \alpha\hat{\mathbf{B}}^t\mathbf{h}_i - (1-\alpha)\hat{\mathbf{B}}^*\mathbf{w}_i^* + \alpha\hat{\mathbf{s}}_i^t - \alpha\hat{\mathbf{s}}_i^*$$
$$= (1-\alpha)\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\mathbf{w}_i^* - \hat{\mathbf{B}}^t\mathbf{k}_i + \alpha\hat{\mathbf{s}}_i^t - \alpha\hat{\mathbf{s}}_i^* \tag{48}$$

Thus,

$$\begin{aligned}
\|\mathbf{q}_i\|_2 &= \left\|(1-\alpha)\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\mathbf{w}_i^* - \hat{\mathbf{B}}^t\mathbf{k}_i + \alpha\hat{\mathbf{s}}_i^t - \alpha\hat{\mathbf{s}}_i^*\right\|_2 \\
&\le \left\|(1-\alpha)\left(\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t\top} - \mathbf{I}_d\right)\hat{\mathbf{B}}^*\right\|_2 \|\mathbf{w}_i^*\|_2 + \|\mathbf{k}_i\|_2 + \alpha\left\|\hat{\mathbf{s}}_i^t - \hat{\mathbf{s}}_i^*\right\|_2 \\
&\le (1-\alpha)\sqrt{k}\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + \alpha C_s\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + \left(\alpha C_s + (1-\alpha)\sqrt{k}\right)\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) &(49)\\
&\le 2\left((1-\alpha)\sqrt{k} + \alpha C_s\right)\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) &(50)\\
&\le 2\sqrt{k}\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) &(51)
\end{aligned}$$

where (49) holds with probability at least $1 - e^{-110k^2\log(rn)}$, by combining equation (44) in (Collins et al., 2021) and (**??**), conditioned on $\delta_k \le \frac{1}{2}$ and $\delta_d \le \frac{1}{2}$. Similarly, combining equation (45) and (**??**), conditioned on $\delta_k \le \frac{1}{2}$, we have

$$\begin{aligned}
\left\|(1-\alpha)\mathbf{w}_i^{t+1}\right\|_2 &\le \left\|(1-\alpha)\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^*\mathbf{w}_i^*\right\|_2 + \|\mathbf{k}_i\|_2 \\
&\le (1-\alpha)\sqrt{k} + \alpha C_s &(52)\\
&\le 2\sqrt{k} &(53)
\end{aligned}$$

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

$$\mathcal{E} := \bigcap_{i=1}^{rn}\left\{\|\mathbf{q}_i\|_2 \le 2\left((1-\alpha)\sqrt{k} + \alpha C_s\right)\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \cap \left\|(1-\alpha)\mathbf{w}_i^{t+1}\right\|_2 \le (1-\alpha)\sqrt{k} + \alpha C_s\right\}, \tag{54}$$

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

$$\begin{aligned}
\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top} &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left\langle \mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top, \mathbf{Q}^{t\top}\right\rangle \mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top - \mathbf{Q}^{t\top} \\
&= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{e}_i\left(\mathbf{x}_i^j\right)^\top - \mathbf{Q}^{t\top}, &(55)
\end{aligned}$$

further, we have

$$\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top}\right)^\top (1-\alpha)\mathbf{W}^{t+1\top} = \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{x}_i^j(1-\alpha)\mathbf{w}_i^{t+1\top} - \mathbf{q}_i(1-\alpha)\mathbf{w}_i^{t+1\top}\right). \tag{56}$$

Next, we establish similar arguments as the derivatives for Theorem 4.4.5 in (**?**) to bound $\left\|\frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\right.\right.$ 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 (**?**). Using equation 4.13 in (**?**), we have

$$
\begin{aligned}
&\left\|\frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{x}_i^j(1-\alpha)\mathbf{w}_i^{t+1\top} - \mathbf{q}_i(1-\alpha)\mathbf{w}_i^{t+1\top}\right)\right\|_2 \\
&\leq 2\max_{\mathbf{z}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k} \mathbf{z}^\top\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \mathbf{x}_i^j(1-\alpha)\mathbf{w}_i^{t+1\top} - \frac{1}{rnm}\mathbf{q}_i(1-\alpha)\mathbf{w}_i^{t+1\top}\right)\right)\mathbf{y} \\
&= 2\max_{\mathbf{z}\in\mathcal{N}_d, \mathbf{y}\in\mathcal{N}_k} \sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\langle \mathbf{z}, \mathbf{x}_i^j\rangle\langle(1-\alpha)\mathbf{w}_i^{t+1}, \mathbf{y}\rangle - \frac{1}{rnm}\langle \mathbf{z}, \mathbf{q}_i\rangle\langle(1-\alpha)\mathbf{w}_i^{t+1}, \mathbf{y}\rangle\right) \quad (57)
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{c}{rnm}\|\mathbf{q}_i\|_2\langle(1-\alpha)\mathbf{w}_i, \mathbf{y}\rangle &\leq \frac{c}{rnm}\|\mathbf{q}_i\|_2\|(1-\alpha)\mathbf{w}_i\|_2 \\
&\leq \frac{c'}{rnm}\left((1-\alpha)\sqrt{k}+\alpha C_s\right)^2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) \quad (58) \\
&:= \frac{c'}{rnm}\Delta \quad (59)
\end{aligned}
$$

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

$$
\mathbb{P}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\langle \mathbf{z}, \mathbf{x}_i^j\rangle\langle(1-\alpha)\mathbf{w}_i, \mathbf{y}\rangle - \frac{1}{rnm}\langle \mathbf{z}, \mathbf{q}_i\rangle\langle(1-\alpha)\mathbf{w}_i, \mathbf{y}\rangle\right) \geq s\right) \leq \exp\left(-c_2 rnm \min\left(\frac{s^2}{\Delta^2}, \frac{s}{\Delta}\right)\right). \quad (60)
$$

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

$$
\mathbb{P}\left(\left\|\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top}\right)(1-\alpha)\mathbf{W}^{t+1\top}\right\|_2 \geq 2s\,\middle|\,\mathcal{E}\right) \leq 9^{d+k}\exp\left(-c_2 rnm\min\left(\frac{s^2}{\Delta^2}, \frac{s}{\Delta}\right)\right). \quad (61)
$$

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

$$
\mathbb{P}\left(\left\|\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}\left(\mathbf{Q}^{t\top}\right) - \mathbf{Q}^{t\top}\right)\mathbf{W}^{t+1\top}\right\|_2 \geq c_4\sqrt{\frac{d}{rnm}}\left((1-\alpha)\sqrt{k}+\alpha C_s\right)^2\operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right)\,\middle|\,\mathcal{E}\right) \leq e^{-110d} \quad (62)
$$

Finally, by using $\mathbb{P}(A) \leq \mathbb{P}(A \mid \mathcal{E}) + \mathbb{P}(\mathcal{E})$, where

$$A := \left\{ \left\| \frac{1}{rn} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A} \left(\mathbf{Q}^{t\top} \right) - \mathbf{Q}^{t\top} \right) \mathbf{W}^{t+1\top} \right\|_2 \geq c_4 \sqrt{\frac{d}{rnm}} \left((1-\alpha)\sqrt{k} + \alpha C_s \right)^2 \operatorname{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^* \right) \right\}, \tag{63}$$

we complete the proof.

□

1.3 Main Result

Recall that $\mathbf{Q}^{t\top} = \mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top} + \hat{\mathbf{S}}^{t\top} - \hat{\mathbf{S}}^{*\top}$, plugging this into (7), and without losing generality, we drop the subscripts of $\mathcal{I}^t$ and obtain

$$\begin{aligned} \bar{\mathbf{B}}^{t+1} &= \hat{\mathbf{B}}^t - \frac{\eta}{rnm} \left(\mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^{t\top}) \right)^\top \mathbf{W}^{t+1\top} \\ &= \hat{\mathbf{B}}^t - \frac{\eta}{rn} \mathbf{Q}^t \mathbf{W}^{t+1\top} - \frac{\eta}{rn} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top}. \end{aligned} \tag{64}$$

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

$$\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^{t+1} = \left(\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\mathbf{Q}^t\mathbf{W}^{t+1\top} - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \right) (\mathbf{R}^{t+1})^{-1}. \tag{65}$$

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

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

plugging this into (65) then we reach

$$\begin{aligned} \hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^{t+1} = \Bigg(& \hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t \left(\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top} \right) + \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top} \left(\hat{\mathbf{S}}^* - \hat{\mathbf{S}}^t \right) \mathbf{W}^{t+1\top} \\ & - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top} \right)^\top \mathbf{W}^{t+1\top} \Bigg) (\mathbf{R}^{t+1})^{-1}. \end{aligned} \tag{66}$$

Therefore,

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

$$\leq \left\|\hat{\mathbf{B}}_{\perp}^{*\top}\hat{\mathbf{B}}^t\left(\mathbf{I}_k-\frac{\eta}{rn}(1-\alpha)^2\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right)\right\|_2\left\|(\mathbf{R}^{t+1})^{-1}\right\|_2$$
$$+\frac{\eta}{rn}\left\|\hat{\mathbf{B}}_{\perp}^{*\top}\left(\frac{1}{m}(\mathcal{A}^{\dagger}\mathcal{A}(\mathbf{Q}^{t\top})-\mathbf{Q}^{t\top}\right)^{\top}(1-\alpha)\mathbf{W}^{t+1\top}\right\|_2\left\|(\mathbf{R}^{t+1})^{-1}\right\|_2$$
$$+\frac{\eta}{rn}\left\|\hat{\mathbf{B}}_{\perp}^{*\top}\left(\alpha\hat{\mathbf{S}}^{*}-\alpha\hat{\mathbf{S}}^{t+1}\right)(1-\alpha)\mathbf{W}^{t+1\top}\right\|_2\left\|(\mathbf{R}^{t+1})^{-1}\right\|_2. \tag{67}$$

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

$$\left\|\hat{\mathbf{B}}_{\perp}^{*\top}\hat{\mathbf{B}}^t\left(\mathbf{I}_k-\frac{\eta}{rn}(1-\alpha)^2\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right)\right\|_2 \leq \left\|\hat{\mathbf{B}}_{\perp}^{*\top}\hat{\mathbf{B}}^t\right\|_2\left\|\mathbf{I}_k-\frac{\eta}{rn}\mathbf{W}^{t+1}(1-\alpha)\mathbf{W}^{t+1\top}\right\|_2$$
$$\leq \operatorname{dist}\left(\hat{\mathbf{B}}^t,\hat{\mathbf{B}}^*\right)\left\|\mathbf{I}_k-\frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right\|_2. \tag{68}$$

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 $\sigma_{\min}^{t+1}$. Then as long as $\eta \leq (\sigma_{\min}^{t+1})^2$, we have

$$\left\|\mathbf{I}_k-\frac{\eta}{rn}\mathbf{W}^{t+1}\mathbf{W}^{t+1\top}\right\|_2 = 1-\eta(\sigma_{\min}^{t+1})^2. \tag{69}$$

$$\left\|\alpha\mathbf{S}^{t+1}-\alpha\mathbf{S}^*-(1-\alpha)\mathbf{B}^*\mathbf{W}^*\right\|_2 \tag{70}$$
$$=\left\|(1-\lambda_S)\left(\alpha\mathbf{S}^t-\alpha\mathbf{S}^*-(1-\alpha)\mathbf{B}^*\mathbf{W}^*\right)-\lambda_S\mathbf{B}^t\mathbf{B}^{t\top}\left(\alpha\mathbf{S}^*+(1-\alpha)\mathbf{B}^*\mathbf{W}^*\right)\right\|_2 \tag{71}$$
$$\leq\left\|(1-\lambda_S)\left(\alpha\mathbf{S}^t-\alpha\mathbf{S}^*-(1-\alpha)\mathbf{B}^*\mathbf{W}^*\right)\right\|_2+\lambda_S \tag{72}$$
$$\leq(1-\lambda_S)^{t+1}\left\|\mathbf{S}^0-\mathbf{S}^*-\mathbf{B}^*\mathbf{W}^*\right\|+(t+1)\lambda_S \tag{73}$$

To bound the term of $\frac{\eta}{rn}\left\|\hat{\mathbf{B}}_{\perp}^{*\top}\left(\frac{1}{m}(\mathcal{A}^{\dagger}\mathcal{A}(\mathbf{Q}^{t\top})-\mathbf{Q}^{t\top}\right)^{\top}\mathbf{W}^{t+1\top}\right\|_2$, we have

$$\frac{\eta}{rn}\left\|\hat{\mathbf{B}}_{\perp}^{*\top}\left(\frac{1}{m}(\mathcal{A}^{\dagger}\mathcal{A}(\mathbf{Q}^{t\top})-\mathbf{Q}^{t\top}\right)^{\top}\mathbf{W}^{t+1\top}\right\|_2 \leq \frac{\eta}{rn}\left\|\left(\frac{1}{m}(\mathcal{A}^{\dagger}\mathcal{A}(\mathbf{Q}^{t\top})-\mathbf{Q}^{t\top}\right)^{\top}\mathbf{W}^{t+1\top}\right\|_2$$
$$\leq \eta\left(\delta_k'\operatorname{dist}\left(\hat{\mathbf{B}}^t,\hat{\mathbf{B}}^*\right)+\delta_k''\right). \tag{74}$$

Similarly,

$$\frac{\eta}{rn}\left\|\hat{\mathbf{B}}_{\perp}^{*\top}\left(\hat{\mathbf{S}}^*-\hat{\mathbf{S}}^{t+1}\right)\mathbf{W}^{t+1\top}\right\|_2 \leq \frac{\eta}{\sqrt{rn}}\left\|\hat{\mathbf{S}}^*-\hat{\mathbf{S}}^t\right\|_2\frac{1}{\sqrt{rn}}\left\|\mathbf{W}^{t+1}\right\|_2 \leq \eta 2\sqrt{k}6\sqrt{k}=12\eta k, \tag{75}$$

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

$$\mathbf{R}^{t+1\top}\mathbf{R}^{t+1} = \bar{\mathbf{B}}^{t+1\top}\bar{\mathbf{B}}^{t+1}$$
$$=\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^t-\frac{\eta}{rn}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}+\mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\right)+\frac{\eta^2}{(rn)^2}\mathbf{W}^{t+1}\mathbf{U}^t\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}$$
$$=\mathbf{I}_k-\frac{\eta}{rn}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}+\mathbf{W}^{t+1}\mathbf{U}^t\hat{\mathbf{B}}^t\right)+\frac{\eta^2}{(rn)^2}\mathbf{W}^{t+1}\mathbf{U}^t\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}. \tag{76}$$

Using Weyl's Inequality, we reach

$$\sigma_{\min}^2\left(\mathbf{R}^{t+1}\right) \geq 1 - \frac{\eta}{rn}\lambda_{\max}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\right) + \frac{\eta^2}{(rn)^2}\lambda_{\min}\left(\mathbf{W}^{t+1}\mathbf{U}^{t}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}\right)$$
$$\geq 1 - \frac{\eta}{rn}\lambda_{\max}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\right) \tag{77}$$

where (77) holds since $\mathbf{W}^{t+1}\mathbf{U}^{t}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}$ is positive semi-definite. Further,

$$\begin{aligned}
&\frac{\eta}{rn}\lambda_{\max}\left(\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top} + \mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\right)\\
&= \max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{\eta}{rn}\left(\mathbf{z}^\top\hat{\mathbf{B}}^{t\top}\mathbf{U}^{t\top}\mathbf{W}^{t+1\top}\mathbf{z} + \mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\mathbf{z}\right)\\
&= \max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{U}^{t}\hat{\mathbf{B}}^{t}\mathbf{z}\\
&= \max_{\mathbf{z}:\|\mathbf{z}\|_2=1}\left(\frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^{t}\mathbf{z} + \frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{Q}^{t\top}\hat{\mathbf{B}}^{t}\mathbf{z}\right)
\end{aligned} \tag{78}$$

When considering the first term, we have

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^{t}\mathbf{z} \leq \frac{2\eta}{rn}\left\|\mathbf{W}^{t+1}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^{t\top}) - \mathbf{Q}^{t\top}\right)\right\|_2\left\|\hat{\mathbf{B}}^{t}\right\|_2 \leq 2\eta(\delta' + \delta'') \tag{79}$$

Then we consider the second term in (78),

$$\begin{aligned}
\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{2\eta}{rn}\mathbf{z}^\top\mathbf{W}^{t+1}\mathbf{Q}^{t\top}\hat{\mathbf{B}}^{t}\mathbf{z} \leq &\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{2\eta}{rn}\mathbf{z}^\top\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^{*}\mathbf{W}^{*} - \mathbf{F}\right)\left(\mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top}\right)\hat{\mathbf{B}}^{t}\mathbf{z}\\
&+ \max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{2\eta}{rn}\mathbf{z}^\top\left(\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^{*}\mathbf{W}^{*} - \mathbf{F}\right)\left(\hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top}\right) + \mathbf{H}\mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^{t}\mathbf{z}
\end{aligned} \tag{80}$$

As for the first term in (80), from equation (81) in (Collins et al., 2021) we have

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{2\eta}{rn}\mathbf{z}^\top\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^{*}\mathbf{W}^{*} - \mathbf{F}\right)\left(\mathbf{W}^{t+1\top}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^{*\top}\hat{\mathbf{B}}^{*\top}\right)\hat{\mathbf{B}}^{t}\mathbf{z} \tag{81}$$
$$\leq 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}_{\max,*}^2 + 2(1+\delta)\eta\bar{\sigma}_{\max,*}\left\|\hat{\mathbf{S}}^{*} - \hat{\mathbf{S}}^{t}\right\|_2 + 2(1+\delta)^2\eta\left\|\hat{\mathbf{S}}^{*} - \hat{\mathbf{S}}^{t}\right\|_2^2 \tag{82}$$

As for the second term in (80),

$$\begin{aligned}
&\frac{2\eta}{rn}\left\|\left(\left(\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^{*}\mathbf{W}^{*} - \mathbf{F}\right)\left(\hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top}\right) + \mathbf{H}\mathbf{Q}^{t\top}\right)\hat{\mathbf{B}}^{t}\right\|_2\\
&\leq \frac{2\eta}{rn}\left\|\hat{\mathbf{B}}^{t\top}\hat{\mathbf{B}}^{*}\mathbf{W}^{*} - \mathbf{F}\right\|_2\left\|\hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top}\right\|_2 + \frac{2\eta}{rn}\left\|\mathbf{H}\mathbf{Q}^{t\top}\right\|_2\\
&\leq 4\eta\frac{1}{\sqrt{rn}}\|\mathbf{W}^{*}\|_2\frac{1}{\sqrt{rn}}\left\|\hat{\mathbf{S}}^{t+1\top} - \hat{\mathbf{S}}^{*\top}\right\|_2 + 2\eta\frac{1}{\sqrt{rn}}\|\mathbf{H}\|_2\frac{1}{\sqrt{rn}}\|\mathbf{Q}\|_2
\end{aligned} \tag{83}$$

$$\leq 4\eta\bar{\sigma}_{\max,*} \left\| \hat{\mathbf{S}}^* - \hat{\mathbf{S}}^{t+1} \right\|_2 + 2\eta(1+\delta) \left\| \hat{\mathbf{S}}^* - \hat{\mathbf{S}}^t \right\|_2 \left(2\sqrt{k}\,\mathrm{dist}(\hat{\mathbf{B}}\cdot\hat{\mathbf{B}}^*) + (1+\delta) \left\| \hat{\mathbf{S}}^* - \hat{\mathbf{S}}^t \right\|_2 + \left\| \hat{\mathbf{S}}^* - \hat{\mathbf{S}}^{t+1} \right\|_2 \right) \tag{84}$$

$$\leq 4\eta\bar{\sigma}_{\max,*} 2\sqrt{k} + 2\eta(1+\delta)2\sqrt{k} \times 8\sqrt{k} \tag{85}$$

$$= 8\eta\bar{\sigma}_{\max,*}\sqrt{k} + 32(1+\delta)\eta k \tag{86}$$

Therefore,

$$\sigma^2_{\min}(\mathbf{R}^{t+1}) \geq 1 - 2\eta(\delta' + \delta'') - 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}^2_{\max,*} - 8\eta\bar{\sigma}_{\max,*}\sqrt{k} - 32(1+\delta)\eta k \tag{87}$$

Finally, we have

$$\mathrm{dist}\left(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*\right) \leq \tag{88}$$

$$\frac{(1-\eta\sigma^2_{\min} + \eta\delta'_k)\,\mathrm{dist}\left(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*\right) + \eta\delta'' \left\|\Delta\hat{\mathbf{S}}^t\right\|_2 + \eta(\delta''' + 6\sqrt{k}/\sqrt{rn}) \left\|\Delta\hat{\mathbf{S}}^{t+1}\right\|_2}{\sqrt{1 - 2\eta\delta'_k\,\mathrm{dist} - 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}_{\max,*} - 2\eta(\delta''_k + (1+\delta)\bar{\sigma}_{\max,*})\left\|\Delta\hat{\mathbf{S}}^t\right\|_2 - 4\eta(1+\delta)^2\left\|\Delta\hat{\mathbf{S}}^t\right\|_2^2 - 2\eta(\delta'''_k + \frac{2\bar{\sigma}_{\max,*}}{\sqrt{rn}}\left\|\Delta\hat{\mathbf{S}}^{t+1}\right\|_2) - 4\eta\sqrt{k}(1+\delta)\left\|\Delta\hat{\mathbf{S}}^t\right\|_2\mathrm{dist} - 2\eta(1+\delta)\left\|\Delta\hat{\mathbf{S}}^t\right\|_2\left\|\Delta\hat{\mathbf{S}}^{t+1}\right\|_2}}, \tag{89}$$

where $\delta_k = c\frac{k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}, \delta'_k = c_1 k\frac{\sqrt{d}}{\sqrt{rnm}}, \delta''_k = c_2\frac{\sqrt{kd}}{\sqrt{rnm}}, \delta'''_k = c_3\frac{\sqrt{kd}}{\sqrt{rnm}}, \delta = \frac{\delta_d}{1-\delta_k}, \delta_d = c_4\frac{\sqrt{d\log(rn)}}{\sqrt{m}}$

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